On coordination in non-cooperative game theory : Explaining how and why an equilibrium occurs and prevails

Lauren Larrouy

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
DROIT ET SCIENCES POLITIQUES ÉCONOMIQUES ET DE GESTION

La coordination en théorie des jeux non-coopérative:

A propos de la formation et de la prévalence d'un équilibre

Lauren LARROUY

GREDEG

Présentée en vue de l'obtention du grade de docteur en Sciences Economiques **de** Université Côte d'Azur
Dirigée par : Richard Arena
Soutenue le : 27/05/2021

Devant le jury, composé de :
Richard Arena, Professeur, Université Côte d'Azur
Annie Lou Cot, Professeure Emérite, Paris 1 Panthéon-Sorbonne
John B. Davis, Emeritus Professor, Marquette University
Alan Kirman, Professeur Emérite, EHESS
Dominique Torre, Professeur, Université Côte d'Azur

La Coordination en théorie des jeux non-coopérative :

A propos de la formation et de la prévalence d'un équilibre

Jury :

Président du jury

Annie Lou Cot, Professeur Emérite, Paris 1 Panthéon-Sorbonne

Rapporteurs

John B. Davis, Professeur Emérite, Marquette University
Alan Kirman, Professeur Emérite, Université Aix-Marseille

Examinateurs

Richard Arena, Professeur, Université Côte d'Azur
Dominique Torre, Professeur, Université Côte d'Azur

La coordination en théorie des jeux non-coopérative : à propos de la formation et de la prévalence d'un équilibre

---

Notre thèse propose de changer l'ontologie et la méthodologie de la théorie des jeux, en définissant les jeux comme la compréhension du processus de raisonnement stratégique des joueurs. Notre contribution est basée sur une approche interdisciplinaire pour une réévaluation du type d'intersubjectivité impliquée dans le raisonnement stratégique.
Nous affirmons que l'analyse des jeux doit impliquer l'étude et la détermination du processus de raisonnement qui conduit les joueurs à une solution spécifique. Un jeu ne doit pas être compris, comme dans la théorie des jeux standard, comme une représentation mathématique d'un choix individuel à l'équilibre.
Cela nécessite d'enquêter sur la capacité de coordination des acteurs. Nous affirmons que la compréhension du processus de coordination permet de comprendre le raisonnement stratégique des joueurs. Cela permet d'apporter de nouvelles réponses au problème d'indétermination de la théorie des jeux qui constitue l'une des impasses auxquelles la théorie des jeux est confrontée et qui souligne ses difficultés positives et normatives.
La thèse est fondée sur l'argument selon lequel la compréhension du processus de raisonnement des joueurs dans les jeux nécessite d'abord et avant tout d'expliquer comment les joueurs forment leurs croyances à propos du choix des autres joueurs, leurs perceptions et leurs croyances, ainsi que la façon dont ils raisonnent. L'un des objectifs de la thèse est de montrer qu'une théorie psychologique expliquant la formation des croyances des joueurs est nécessaire pour rendre compte de la coordination, et que la théorie de l'esprit (ToM) offre un cadre psychologique adéquat. Nous suggérons de construire une théorie des jeux alternative basée sur la théorie de la simulation comme théorie de l'esprit. Nous définissons une caractérisation axiomatique des choix rationnels dans les jeux lorsque les joueurs simulent le raisonnement des autres.

Mots clés : théorie des jeux non-coopératifs, coordination

On coordination in non-cooperative game theory : explaining how and why an equilibrium occurs and prevails

---

Our thesis proposes to change the ontology and methodology of game theory, appraising games as the understanding of the players' strategic reasoning process. Our contribution is based on an interdisciplinary approach for a reassessment of the kind of intersubjectivity involved in strategic reasoning.
We claim that the analysis of games should involve the study and the determination of the reasoning process that lead the players to a specific outcome, i.e. to a specific solution. A game should not be understood, like in standard game theory, as a mathematical representation of an individual choice at the equilibrium. This requires investigating the players' capacity of coordination. We assert that understanding the process of coordination allows understanding strategic reasoning and ultimately to provide new answers to the indeterminacy problem of game theory which is one of the stalemates that game theory faces and which underscores its positive and normative difficulties.
The thesis is grounded on the argument that understanding the players' reasoning process in games necessitates first and foremost to explain how the players form their beliefs regarding each other's choices, but also each other's perceptions and beliefs and reasoning processes in a strategic context. One of the purposes of the thesis is to show that a psychological theory explaining the formation of players' beliefs is required to account for coordination, and that the Theory of Mind (ToM) offers such adequate psychological framework. We suggest building an alternative theory of games based on the simulation theory as such theory of mind. We then specify an axiomatic characterization of rational choices in games in the presence of players able to simulate the reasoning of others.

Keywords : non-cooperative game theory, coordination

A Hugo, mon petit frère,

A Aimée et Virgile, mes enfants,

Remerciements

En songeant à ces remerciements j'ai le sentiment que plusieurs vies se sont écoulées depuis le début de cette thèse. J'ai donc bien peur d'omettre de mentionner beaucoup de personnes ici… je m'en excuse.

Avant d'en commencer la liste une adresse toute spéciale me tient particulièrement à cœur. Tout au long de cette thèse j'ai toujours eu en tête mon petit frère… et l'envie de la finir et de la soutenir pour lui surtout et avant tout. J'ai toujours imaginé le jour de la soutenance de le voir dans le fond de la salle, un magnifique sourire aux lèvres et de la fierté dans son regard malicieux. Je sais à quel point il aurait été fier de sa grande sœur !!! Et puis je sais aussi qu'il aurait tout préparé pour le pot de thèse et que nous nous serions régalé… Alors le jour de la soutenance je regarderai dans le fond de la salle et je penserai à toi mon petit frère et j'imaginerai ton si beau sourire rempli de fierté ; et c'est à la fois la plus douce et la plus douloureuse image que j'en ai de cette soutenance ; mais tu seras là !!!

Dans la liste des personnes à remercier et qui ont plus que largement contribuées à l'élaboration de cette thèse il y a évidemment mon directeur de thèse Richard sans qui je n'aurais jamais pensé pouvoir faire une thèse s'il ne m'en avait pas parlé en licence et pour qui j'ai beaucoup d'admiration professionnelle et intellectuelle. Je ne te remercierai jamais assez ; tu m'as fait grandir et prendre confiance en moi… et puis tu ne m'as jamais abandonné même après la mort d'Hugo, sans quoi j'aurai sans doute raccroché…

Et puis évidemment je remercie mon papa qui m'a ramenée de force à Nice pour me remettre au travail après la mort d'Hugo et toute la bande des doctorants d'alors du GREDEG grâce à qui le retour a été bien plus léger et plus festif : Anaïs, Maëlle, Ankinée, Tania, Raph et Ana, Dorian, Guillaume, Jamal, Patrice, et Margo…

Je remercie tout particulièrement les membres du jury, je vous remercie d'avoir accepté de siéger dans ce jury après de si nombreuses années et je vous remercie pour les rôles importants que vous avez tous pu jouer, à votre manière, au cours de cette thèse. Votre présence m'aide à aborder avec sérénité cette soutenance et l'idée de vous revoir tous me réjouit.

Un très grand merci Annie de m'avoir permis d'assister à votre master et de m'avoir intégrée dans le petit cercle très privilégié (je trouve) de vos doctorants. J'y ai découvert un univers tout particulier, un mélange d'amitié et d'émulation qui m'a énormément apporté à beaucoup d'égards ; pas uniquement sur le plan académique.

Merci John pour votre bienveillance, vos encouragements, interventions en séminaires ou en conférence, votre intérêt et votre suivi.

Dominique un grand merci pour votre gentillesse, votre écoute, votre suivi depuis la licence et votre confiance dans le chercheur que je devenais.

Merci Alan pour nos échanges passionnants en conférence sur Schelling et l'émergence, et le multi agent.

Je remercie évidemment tout particulièrement mes co-auteurs Cléo, Guilhem et Cyril, un grand merci pour nos fructueuses collaborations et les avancées qu'elles m'ont permises.

J'adresse un message tout particulier à mes compères HPistes et philosophes Tom et Dorian mes presque co-auteurs (ce n'est pas faute d'en avoir parlé souvent). Un immense merci pour tous nos échanges, y compris tardifs mais là j'en ai beaucoup moins de souvenirs !!! j'ai perdu quelques cellules hépatiques avec vous deux au Pompadou et au Fût et à Mesure … Merci pour ces soirées mémorables… mon foie vous remercie peut être moins ... heureusement que la maternité m'a permis un sevrage radical. Du coup dans ces soirées mémorables la fine équipe de la MSH n'est pas en reste je les remercie tous pour ces soirées joyeuses et endiablées: Alex, Tania, Magalie, Stéph, Elise, Jamal, Saveria, et j'en oublie beaucoup…

Bien sûr il y a la fine équipe du Gredeg : nos soirées, nos restau, et nos voyages: Berlin et Amsterdam. Des moments mémorables, je recommencerai bien ;)!

Et puis il y a l'équipe de la MSE avec encore une fois des soirées mémorables et de belles amitiés ! Une pensée toute spéciale pour Juju (tu me manques), pour Cléo, pour Matthieu, Eric, Niels…

Ce qui me fait penser que je remercie le GREDEG et les organisateurs des nombreuses conférences et écoles d'été auxquelles j'ai eu la chance de participer. J'ai adoré ces moments privilégiés, je trouve, dans la vie de chercheur (mes heures de sommeil en ont pris un sacré coup à chaque fois … mais j'ai découvert bien plus tard qu'on pouvait faire pire … avec la maternité !). Merci aux copains (Dorian, Tom, Cléo, Judith, Niels, Nicolas, sans qui bien évidemment ces conférences auraient été bien différentes, et pour les mêmes raisons, une pensée pour Jean Seb ; pour Agnès et pour Pierre…). Un grand merci aux nombreux chercheurs au cours de ces conférences et écoles d'été qui ont largement contribué à améliorer la qualité de mon travail et pour leur multiples suggestions.

Je remercie également vivement Mme Arfeuil, pour son aide, sa disponibilité et sa réactivité pour finaliser l'inscription administrative et la préparation de la soutenance.

Bien sûr un immense merci à Anaïs pour tous ces moments passés, ces sorties montagnes, spéléo escalades, l'apprentissage de la SLAC, ces journées ou soirées filles, ces apéros (t'as un foie en acier c'est de la triche), ces virées shoping en Italie, ces virées andernosiennes et les sorties bateaux (je ne savais pas jusqu'alors que le mal de terre existait … merci !)… c'était chouette à tes cotés.

Et encore bien sûr un immense merci à ma chouchou d'amour de Maëlle, merci pour nos rigolades, nos papotes, nos soirées, nos sorties (ton foie est moins en béton que celui d'Anaïs ouf !!!!) pour ta complicité, pour nos virées andernosiennes et bordelaises aussi, enfin un grand merci d'être toi quoi !!! et évidemment un grand merci à toi et Grégouille pour votre accueil, enfin, pour vos très nombreux accueils devrais-je dire ; j'avais l'impression d'être à la maison ; c'est dire ! Cette maison me manque d'ailleurs un peu !

Parmi les logeurs de doctorante émigrée je dois remercier Tom et Marie Prune (et presque Rita) qui ont bien voulu nous accueillir Aimée et moi et avec qui nous avons passé une semaine des

plus agréables. Nous leur avons servi de crash test pour savoir quoi sécuriser à la maison !!!! Un très grand merci à vous deux.

Et enfin sans transition aucune, un grand merci aux nombreuses personnes qui ont pris du temps pour relire mes articles et chapitres de thèse et qui ont eu le courage de corriger mon anglais ; dans le désordre, Charlotte, Rudolph, Lindsay Mégraud, Cyril, Dorian, Tom, Nicolas, et Guilhem, Valérie, Robert et Maureen avec vraiment un grand merci pour l'assiduité de Dorian que j'ai beaucoup beaucoup beaucoup sollicité quand même et qui a rarement dit non, et pour Guilhem aussi !

Merci au personnel administratif du GREDEG et de la MSH, aux directeurs du GREDEG et de la MSH ; pour leur travail et pour nous fournir d'agréables conditions de travail ; une pensée toute particulière pour Laurence et sa bienveillance pour les doctorants. Merci aux nombreux chercheurs du GREDEG qui m'ont soutenue au cours de mon parcours de doctorant et même avant (un grand merci tout particulièrement à Muriel évidemment, qui même après 8 ans de thèse me pousse toujours à soutenir et régulièrement !!! qui m'a aidée dans mon grand saut dans le monde des conférences j'étais encore un Master… à Dominique, à Christophe, à Agnès et à tous ceux qui m'ont donné envie lors de mon parcours universitaire de faire de la recherche).

Et enfin un immense merci à mon cher Mari (qui l'eût cru!) A Christophe, qui me soutient depuis que j'ai recommencé mes études et qui m'a toujours porté, et puis qui a traversé toutes ses vies avec moi et qui bien sûr n'a jamais douté que je soutiendrai un jour cette thèse (non sans impatience ces dernières années il faut l'avouer).

C'est avec à la fois crainte et soulagement que j'imagine maintenant cette soutenance, avec une pointe de nostalgie de revenir au GREDEG avec un nouveau regard : celui de maman ; et d'y voir mes enfants le jour de la soutenance : image que je n'aurais jamais imaginée au commencement de cette thèse. Et puis y voir peut être une partie de ce nouveau cercle amical construit dans cette nouvelle vie de maman et d'imaginer tous ces copains qui me poussent aussi à soutenir en me demandant régulièrement quelles sont les avancées de ma thèse et quand est ce que l 'on va fêter ça à Nice : Marguerite, Eloise, Anna, Sabine, Alain (surtout Alain)… et imaginer le GREDEG avec des enfants pour cette soutenance : ce qui est aujourd'hui ma vie !

Contents

Introduction

The aim of this thesis is to examine the conditions under which players may actually implement a 'solution' to a non-cooperative game, i.e. the conditions under which an equilibrium exist, the process of reasoning that leads the players to the solution in question, and how they converge to the identification of the same solution. Indeed, the specific account of equilibrium and solution entailed by the mathematical definition of games in classical game theory, supposes the existence of a solution and focuses on the mathematical conditions of the existence of such solution without any possible explanation of the specific process or the "forces" leading to this solution (Giocoli, 2003). The existence of the equilibrium is assumed though not explained (ibid), even though the purpose of game theory is to "propose" solutions for games (Sugden, 2001), both from a normative and positive point of view in order to define rational play in the game. Game theorists generally show a lack of interest for the conditions, during the agents' interaction, ensuring the existence of the solution. A game is not conceived as a process but as a mere representation of a strategic choice. In that perspective, Sugden (2001, p. 128) mentions the "unwillingness on the part of economic theorists of decision-making to face up to empirical questions. It seems that the most persistent feature of the theory is not any unifying explanatory principle, but commitment to an a priori mode of enquiry."

The thesis therefore proposes to examine the conditions under which a specific solution can emerge. This requires investigating the players' capacity of coordination understood like in Schelling (1960)'s theoretical contribution: as the process of convergence of player's intentions and beliefs, and then actions. The existence of a solution indeed supposes that the player's beliefs on other's choices and behaviors converge, i.e. be consistent each other. Investigating the conditions under which the players' beliefs can converge requires focusing on the player reasoning, i.e. understanding strategic reasoning as an actual reasoning process in which players must adjust each other. It thus necessitates, contrary to what is done in game theory (and even in epistemic game theory, as will be argued in this thesis) to incorporate player's 'mental states' within games. Players' beliefs are generally called mental 'variables' in epistemic game theory (see Perea, 2014), but we will more generally refer to mental 'states' in the thesis; a term borrowed from cognitive sciences. Mental states refer not only to player's beliefs but also to their preferences, their intentions or perceptions. I will argue that two distinct definitions of payoffs and beliefs in games coexist: one as the mere representation of choices in which there is no room left for defining the player's motives and perceptions or beliefs understood in terms of mental variables, and one acknowledging the role of player's mental states and reasoning process. Standard (classical or epistemic) game theory relies on the first account of payoff and thus, does not offer an *explanation* of such choices. This feature derives from the mathematical representation of the solution that prevails in standard game theory.

In Giocoli (2003)'s terms, the aim of the thesis is to ask the question of 'how and why' a specific solution occurs, and investigates the conditions under which it can exist an answer to the question 'how and why' this specific solution can occur. For that purpose, the objective of the thesis is to offer to the reader an ontological viewpoint on coordination.

1. What is game theory?

Game theory is a framework of analysis (Schelling, 1960; Giocoli, 2003; Aumann, 2000; Aumann and Dreze, 2008); it is a mathematical theory that formalizes strategic interactions, i.e. situations of interdependence of individual choices.

Providing a mathematical theory of such situations entails that the individuals involved in the interactions, the players, act within the rules of the games created by the theorists. A game specifies the set of players, their strategies which circumscribe the possible actions available to the players, the outcomes of the interactions, i.e. the payoffs. Acting within the rule of the games thus means that the players know the players with whom they play and their set of strategies, i.e. their possible actions, and the payoff structure.

In its most basic characterization, standard non-cooperative game theory formalizes contexts of strategic interactions without communication. Basically, it means that the outcome of interactions relies on the combination of the interacting agents' individual decisions. Accordingly, players have to take into account other players' possible actions. They have to form beliefs about others' decisions and beliefs. This specific kind of uncertainty entails that game theory requires rigorous principles defining (i) players' decisions (Bacharach, 1976, p. vii) and (ii) the epistemic requirements of players' decisions (Colman, 2003), i.e. the knowledge they have (or beliefs) about the other players.

Common knowledge of rationality is generally assumed which entails that each player is rational, each player knows that each other is rational, that each other knows that each other knows that each other is rational and so on ad infinitum. In this manner the players are able to form beliefs about others' choices and to infer the other players' choice.

The uncertainty that prevails in game theory is of a specific nature as it relies on what the other player may decide and how they can possibly act; although common rationality (being instrumental in case of perfect information or Bayesian in case of imperfect information) and common knowledge of rationality help to circumscribe what the players can rationally anticipate of other players' action, uncertainty is of a particular complexity. As each player's choice rely on each other's choice; players are in symmetrical position. The choice of a player is indeterminate without assessing the others' choice, but every player's being in the same position this can lead to an endless reasoning, and to an indeterminate choice. To break this endless chain of regression common knowledge of rationality is assumed.

As Hargreaves Heap and Varoufakis (2004[1995], p. 6) emphasize, the hypotheses of rationality and of knowledge of the rules of the game are ontological postulates while common knowledge is

an epistemological postulate. While rationality and the knowledge of the rules of the games define what a game is and what is identified as a contribution to game theory, i.e. the mathematical theory of games, common knowledge is a methodological device within this mathematical theory to determine acceptable results. In particular an acceptable result is the definition of a solution for games. The purpose of game theory is to provide determined solutions, and the mathematical requirement for a result of the games to be acceptable is the existence, stability and uniqueness (for modern non cooperative game theory) of the solution of the game (Sugden, 2001).

This emphasis on the solution of the games, applicable for any interactive situation, leads to the search of a solution concept broad enough and rigorous enough to apply to each game.

> "game theory is a sort of umbrella or 'unified field' theory for the rational side of social science … [game theory] does not use different ad hoc constructs… it develops methodologies that apply in principle to all interactive situations" (Aumann, 2000, p. 47)

Such methodologies are first and foremost driven by the search for *a*, and even more, *the* solution concept: "it is up to the solution concept to identify what it is meant by 'good playing', that is, by rational behavior, in a given game." (Giocoli, 2003, p. 212) It is supposed that the principle of rational determinacy prevails in game theory (Sugden, 1991), that is, that there exists only one way to play rationally in games and for every possible interactive strategic situation, i.e. any possible game.

2. The solution concept: two visions the System Of Force vs. System Of Relation view of economics

As explained by Giocoli (2003), the miss-specification of the process leading to a given solution for games is determined by a specific account of equilibrium and a specific conception of economics as a science.

A solution is a function that associates, to each game, a (small) subset of outcomes among the possible outcomes of the game (Giocoli, 2003; Sugden, 2001). From a mathematical point of view, a solution must be defined by an equilibrium point (a fixed point following Nash's contribution). There exist two accounts of equilibrium which are associated, as explained by Giocoli (2003) to two visions of economics and of individual rationality. These two accounts of equilibrium have important methodological and more generally epistemological consequences for game theory as will be detailed below.

Two different accounts of the term "equilibrium" successively existed in so-called neoclassical economics: i) the equilibrium as "an attractor of arbitrary motions of the underlying dynamic process" and ii) the equilibrium as "a state of no motion" (Weintraub, 1991, p. 18). Giocoli (2003, p. 138; referring to Weintraub, 1991, p. 102) adds that the former type of equilibrium "is characterized by the fulfillment of a set of static conditions; there is no mechanism through

which equilibrium is established" while the latter type of equilibrium is "associated with the mechanical image of the achievement of a balance of forces … this requires the existence of an equilibration process, by virtue of which the equilibrium is actually reached"

These two accounts of equilibrium relate to two visions of economics as a science: "two images of economics as a scientific discipline." The first "image" entails a system of relations (SOR) view of economics which is defined by Giocoli (2003, p. 139) as "a condition of mutual consistency between a set of economic relations" and in which the existence of the equilibrium and the properties of the equilibrium are at the center of the analysis. The second "image" of economics, the system of forces (SOF) view, entails according to Giocoli's definition, "equilibrium as the end-point of an economic process that itself constitutes the central topic of investigation." (ibidem).

The changeover from the SOF to the SOR in economics occurred after the WWII. We evolved from an explanation of economic phenomena in terms of markets and market forces, i.e. a dynamic analysis focused on learning processes in which economic analysis investigates how and why a specific equilibrium occurs, to a static analysis of equilibrium in which economic analysis assumes the existence of equilibrium a priori, and never explain it (Giocoli, 2003, p. 202). In the latter case the choice of economic agents are consistent, "in harmony". Thus, no out of equilibrium path exists; there is no investigation of the forces that drive the economic system to the equilibrium: the analysis focuses on the existence of the equilibrium and the properties of such equilibrium (of optimality for instance). It provides a static analysis of the equilibrium: a mere representation of this equilibrium. This is founded on the consistency view of rationality, which states that the plan of every economic agents must be in accordance, i.e. consistent each other. But again "[t]he price to be paid for this solution is the impossibility of explaining or justifying how and why the equilibrium occurs in the first place" (Giocoli, 2003, p. 208).

The account of rationality as the consistency of individual choices dates back to Samuelson (1947) and revealed preferences theory, and culminates with the analysis of the Bayesian foundations of equilibrium concepts. As will be explained in each case, what is offered is the representation of individual choices, i.e. a way to describe choices when they respect the stated axioms, but it does not explain the choice to be made. The difficulty of this approach is that it "has forced neoclassical economics to abandon no less than its major theoretical goal, namely, the explanation of the individual's behavior." (Giocoli, 2003, p.42)

Modern game theorists thus focused on the search of the mathematical conditions insuring the existence, uniqueness and stability of solutions. They have essentially developed and refined the mathematical tools of game theory in order to propose defined solutions for games (e.g. see Schelling, 1960; Bacharach and Hurley, 1991; Bacharach, 1986, 2006; Hausman, 2000; Grüne-Yanoff and Lehtinen, 2012). For instance, epistemic game theory – the contemporary version of non-cooperative game theory with incomplete information (Péréa, 2014) – provides the players' epistemic conditions that are compatible with defined solution concepts. However, as it will be shown, epistemic game theory draws on a very specific definition of players' beliefs, which leads to numerous criticisms, but more importantly, it does not provide the tools to explain where the players' beliefs come from and why they may converge to a solution. Epistemic game theory

merely describes the beliefs of the players and ultimately the choices that the players have made that are compatible with the equilibrium, with the solution concept defined *a priori*.

3. Why focusing on coordination? What is coordination?

Coordination in game theory is identified by specific games in which there exist two or more equilibria. Here follows one of the paradigmatic coordination games: the Stag-Hunt Game.

SH	A_2	B_2
A_1	(3; 3)	(0; 2)
B_1	(2; 0)	(1; 1)

From the perspective of classical game theory, several strategy profiles could be equilibria of the game: both A1A2 and B1B2 are Nash equilibria, A1A2 is the only strong Nash equilibrium, and B1B2 is the only stochastically stable equilibrium (Foster & Young, 1990) for instance. However, from your perspective as a player, classical game theory is of little help: if several strategy profiles can be equilibria of the game (depending on the solution concept), the theory does not give you any criterion to select one of these equilibria. Classical game theory cannot therefore offer an operative theory of rational choice in games.

From an epistemic perspective, your optimal strategy – and then the resulting equilibrium strategy profile – depends on your beliefs about the choice of P2. But the choice of P2 depends on her beliefs about your own choice. Neither of your choices is determined without the other determined but since you are in a symmetrical position, both choices remain undefined as no rational basis can help you to select one equilibrium instead of the other.

Thus, there exists no 'solution' from the point of view of classical game theory: strategic rationality does not provide sufficient reason to choose one equilibrium among the multiple ones.

Because players are in symmetrical position their decision depend on the decision of the other players who are in the same situation so that they cannot determine rationally and independently which equilibrium to choose to coordinate on. This is the case only if the relevant information on which players can count on is included in the description of the game and the rules of the game. Such statement will be extensively explained in the chapter 2, 3 and 4 of the thesis.

It is indeed important to stress that, following von Neumann and Morgenstern's contribution, game theory became "an internally closed procedure which operates according to fixed rules known by all mathematicians [and presently by all game theorists]" (Giocoli, 2003, p. 15; quoting von Neumann 1983 [1931], pp. 61-62). As a result, the premise that mathematics is "the universal

language" (ibid, p. 16) translates into game theory. Following von Neumann and Morgenstern, game theorists assumed – and again, still assume – that every player can do what they are capable of doing as game theorists, i.e. understanding mathematics to draw valid conclusions from a mathematical structure. It implies that all of the relevant information needed for players to make their decision is contained within the mathematical structure of games. The rules of the games suffice to determine from the players' perspective the solution of the game: the rational action to accomplish. This explains why the indeterminacy problem is deeper than a mere methodological problem in this epistemology. What stems from outside the game should not matter. Perceptions of the game and of the strategic situation for instance do not matter, as they cannot be encapsulated in the rules of the game. When it is impossible to derive a solution (i) from the rules of the game, (ii) the agents' knowledge and (iii) the rationality postulate (those things being common knowledge), some determinants of player's reasoning must be found outside the theory of games. As it is impossible with the consistency view of rationality to explain how and why a solution occurs we need to search for such – non mathematical – conditions sustaining the existence of a solution, for such processes leading to the solution. These are the real foundations of the reasoning of the players. Such processes are about coordination.

From a methodological perspective this problem exhibited through the indeterminacy problem led to numerous research programs among which is the refinement program that led to hundreds of contributions. But at the same time, such program revealed to be quite disappointing as no solution concept emerged to provide an answer to the problem of coordination games. This result is even more disturbing if we consider that the problem of the existence, or uniqueness of the equilibrium is the major concern of game theorists and economists. But as the SOR view of economics entail that the equilibrium is an hypothesis of the model and not a result, the analysis of the process that led to this equilibrium being left to other disciplines of social sciences, there exists an indeterminacy that is deeper than the solution problem. Being focalized on the methodological and mathematical conditions justifying the existence of a solution has prevented an ontological thinking on the conditions sustaining and justifying coordination. It has never been at the center of the analysis to determine the players' reasoning process in coordination. The thesis will attempt to take this path, i.e. to examine from an ontological perspective the determinants in the players' strategic reasoning process that ultimately induces coordination. Interesting results in game theory on coordination problems show that this happens when the players reasoning process is analyzed, that the players psychology, their mental states are integrated in games that some answers are delivered.

If behavioral game theory has extended the analysis and formalization of some of the mechanisms leading players to cooperate, and considerably enhanced our understanding of cooperation in strategic context based on empirical data, the study of coordination remains however underdeveloped. Some of the propositions to solve the difficulty to explain and rationalize coordination in game theory, such as the famous focal points, are well known but at the same time, their use and formalization made rather little progresses, even during the later years with the rise of behavioral game theory. As will be detailed in the chapters 2 and 3, a focal point is a behavior, a decision that stands out as a solution from all the other potential solutions. The players perceive that this solution has an attractive power that the others do not have. This accordingly requires accounting from other dimensions in players' reasoning that the mere rules

of the games as supposedly specifying all the information required to determine the equilibrium of the game and therefore the solution.

We argue in this thesis, that coordination is a more general phenomenon than is described in coordination games such as the Hi-Lo or Stag-Hunt games in which two equilibria exist. As will be exposed in the chapter 2, coordination is inherent to each type of game, to each type of strategic interaction even in situations showing a scope for conflict of interest. This is why the coordination issue is of such importance. Coordination concern is broader than the paradigmatic example of the coordination games.

Besides as accurately emphasized by Sugden

> "It might be objected that pure coordination games are highly artificial, and that in the games of real life we almost always find some degree of payoff asymmetry, repetition, or communication. But this, although true, is not an adequate reason for ignoring pure coordination games. These games may be thought of as controlled experiments which, by filtering out other features of real-world games that might induce players to choose one strategy rather than another, isolate the effects of labelling. From Schelling's investigations, and from replications of these, we know that the players of pure coordination games make some systematic use of labels, to their mutual benefit. This raises the possibility that the players of other games may also be influenced by labels, and that by filtering out the effects of labelling, game theory may be neglecting a significant determinant of economic behaviour." (Sugden, 1995, p. 534)

Coordination games when carefully examined exhibit the complexity of the process of reasoning in strategic interactions, the many dimensions that intervene in the convergence of players' perceptions, intentions, beliefs and then behaviors.

Following Schelling's ontology of strategic interactions, we argue that coordination should be understood as the phenomenon of the convergence of players' perceptions – i.e. frames –, intentions, and beliefs. Accordingly, for each type of games, coordination is the main determinant leading to a specific solution. As early pointed out by Schelling (1960) such convergence process remains understudied, so that the psychological and behavioral determinants in players' capacity to come to a solution are undefined. Understanding such process of convergence relies on these psychological and behavioral determinants. As a consequence, we argue that the nature of strategic rationality remains insufficiently investigated, and some of the most famous shortcomings of game theory – e.g. equilibrium selection in coordination games – remain. Yet, again, the purpose of game theory is to propose defined solutions for the games (Sugden, 2001, p. 115). Proposing solutions both from a normative and positive point of view thus requires investigating coordination mechanisms.

4. The impact of the type of players and of their rationality in games

> "[W]hat is missing in game theory is any serious attempt to model the players and that it is this lack which is largely responsible for the difficulties that have arisen in the foundations of the subject." (Binmore, 1990, p. viii)

To understand the problem posed by coordination in non-cooperative game theory, it is important to stress that players in games are portrayed like their rationality. While the picture of the 'player' in game theory has evolved with the change from the SOF to the SOR view of economics, the notion of strategic rationality did not considerably evolve.

If coordination as argued earlier is understood as the convergence of players' perceptions and beliefs so that their actions are ultimately consistent, the psychology of players and therefore the way they are appraised in games is again of particular importance. The role of beliefs in non-cooperative game theory is indeed decisive as will demonstrated in the thesis and in consequence the way players are described, their reasoning, and their mental states become preponderant. For that purpose contrary to the SOR view, players must be human and not be deprived from their psychology (e.g. see Giocoli, 2003, p. 208). Since the consistency view of rationality involved in the SOR is immune to the humanity of the players, it describes *invariably human* agents as if they were *non-human* players.

> "From the characterization of rationality as a mere consistency restriction there emerges a purely formal representation of the decision-maker that fits any kind of agent, be 'it' a human, a group, an institution, or even a computer. In other words, the main notion of rationality in contemporary mainstream theory is at best agnostic with respect to the nature of the agent whose rationality is predicated in the theory and is left with modeling individuals as formal algorithms." (Giocoli, 2003, p. 42)

Nevertheless, rationality in term of consistency of choices is not exclusive and players can be rational by maximizing their expected payoff without both accounts of rationality being conflated. This requires as will be argued in chapter 1 defining the players' payoffs not as von Neumann's Morgenstern expected utility, as is often implicitly assumed. Indeed, it is often neglected that both visions of rationality are nowadays conflated but that they are not mutually exclusive. Leaving aside the consistency view of rationality allows us to escape many difficulties of modern non cooperative game theory (such as the inability to explain coordination or cooperation, while experimental results show that players are generally inclined to cooperate (Sally, 1995) and able to successfully coordinate (Schelling 1980 [1960]; Metha, Starmer and Sugden, 1994a,b).

An important feature of the consistency view of rationality in game theory is that players are deprived from their humanity, from their psychology and therefore from their reasoning capacity. This has again important methodological consequences. If players are deprived from their humanity: (i) they are in a sense naturalized, i.e. their choices and behaviors are treated as natural events, and appraisable by objective probabilities, (ii) if what make their psychology, their subjectivity (i.e. their personal perceptions, beliefs or intentions understood as mental variables)

is discarded from the analysis of games, players cannot be heterogeneous. As Giocoli (2003, p. 208) puts it "the consistency view of rationality does not allow for the modeling of the behavior of heterogeneous players, that is, of the kind of agents that populate game-theoretic environments."

Numerous claims are indeed made with respect to the fact that "the other" is naturalized both in complete and incomplete information games (e.g. see Hargreaves Heap, Varoufakis, 2004[1995], p. 37; Lesourne et al., 2006, p. 69). All the possible dispositions in which "the other" may be are contained within the definition of the game, i.e. within the different states of the world that are described by the game. According to Aumann (1987, p. 1) this specificity is explained by the fact that "probabilities can only be assigned to events not governed by rational decision makers." This means that probabilities can only be assigned to natural events. That is why, in some way, the other is treated like an event of "Nature."

> "[T]he same rationality applies to the actor for individual decision in a risky environment and in game theory, the "strategic uncertainty" about the opponent being in some sense "naturalized" in a "physical uncertainty"." (Lesourne et al., 2006, p. 69)

Therefore, many authors state that game theory stays anchored within individual decision theory (e.g. Harsanyi, Aumann). For instance, Bacharach and Hurley (1991, p. 3) highlight that "a number of questions arise about the relationship between individual rationality and game-theoretic rationality" and ask "whether games may be embedded within supposedly individual decision problems." If this is the case however, the status of strategic rationality can be seriously undermined (e.g. Mariotti 1995), as individual decision theory is unable to properly represent the reasoning leading several players to determine the solution of the game. In other words, the possibility to account for coordination, i.e. the capacity of players to end up with consistent beliefs regarding each other's choices, is seriously undermined.

Game theorists still commonly prescribe to von Neumann and Morgenstern's conception of individual rationality (Bacharach, Hurley, 1993, pp. 3-4; Giocoli, 2003), i.e. "an objective definition of rational behavior that could guide a player's choices in a game independently of his/her psychology and opinion on the others players' psychology" (Giocoli, 2003, p. 13). They "derive a theory of rational play in games from one of rational individual decision-making […] which we may call individualism in game theory" (Bacharach, Hurley, 1993: 3-4). That is why players' rationality in games presupposed – and still presupposes – an internal consistency of choices. Von Neumann and Morgenstern's conception of game theory as a "tool-box of powerful analytical methods" (Giocoli, 2003: 13) and of rationality must be considered alongside with an axiomatic approach which is still reckoned as an untouchable statement of the foundations of standard game theory.

This progressively led to the culmination of the escape from psychology "freeing choice theory of the need to refer to unverifiable mental variables." (Giocoli, 2003, p. 201) The explanation of how players reason toward the equilibrium, i.e. the explanation of how and why they coordinate, was therefore considered to belong to the realm of other social sciences such as sociology, or psychology: game theorists are on the contrary focused on the analysis of the mathematical properties and existence of the equilibrium.

Explaining how players ends up with consistent beliefs regarding each other's choice, i.e. beliefs at the equilibrium necessitates to explain how from possibly different perceptions of the game, different epistemic states, (i.e. states of knowledge and /or of beliefs) they progressively converge to common perceptions, common beliefs and common actions. The problem is that as emphasized earlier the consistency view of rationality prevents us from modeling heterogeneous players. The heterogeneity of players' mental states is precluded from the formalization of games in the theory of games understood as a mathematical theory of games, even though it seems to be the inescapable way to explain coordination, i.e. to explain how players progressively converge from out-of-equilibrium path to the equilibrium.

5. Bayesianism in game theory: on decision theory and game theory

The privileged framework for studying coordination and explaining how and why an equilibrium of beliefs, intentions and behaviors occurs is epistemic game theory, as it requires incorporating the players' reasoning process and beliefs. It is therefore of primary importance to stress the methodological consequences of imposing Bayesian rationality in epistemic game theory, on this reasoning process, and on individual beliefs in games. With respect to coordination, as emphasized by Thomas and De Scioli (2014, p. 658) "The challenge is … epistemological: accurately representing the other actor's state of knowledge. The epistemological problem results from the difficulty of converging on a single solution when more than one is available."

The problem of coordination in game theory is first and foremost an epistemic problem because coordination refers to the state of knowledge that the players hold about each other's state of knowledge, perceptions, intentions and beliefs. Epistemic game theory is the modern account of non-cooperative game theory in which the core of the analysis is the players' epistemic states, their reasoning process and the beliefs they hold about each other's beliefs and choices (Aumann and Drèze, 2008; Lecouteux, 2018b). The epistemic program of game theory intends to answer the question of why an equilibrium is rational, i.e. why the players should play a specific equilibrium and not another.

Epistemic game theory is also often called Bayesian game theory, as there is the introduction of uncertainty, and in this case players are Bayes rational; i.e., they act so as to maximize their expected utility according to their beliefs about other's beliefs and choices in particular. As emphasized by Giocoli, what is at the center of the analysis of epistemic game theory and accordingly of Bayesian game theory is the eductive mechanism; i.e. the reasoning process of the players leading to the equilibrium.

> "The analysis of pure eductive mechanisms is usually conducted with the tools of so-called Bayesian game theory. The link between eduction and Bayesianism follows from the requirement that in any pure eductive mechanism each player must understand the other players' mental processes, or at least hold some definite beliefs about them; this in turn presupposes that some knowledge or ability to handle the information be shared by the players. But these are just the distinctive features of Bayesian games, whose main goal

> is precisely to model strategic situations where it is essential to take into account what a player believes the others would do or think, and whose key assumptions are precisely that some crucial characteristics of the situation are common or mutual knowledge among the players and that the latter share the same a priori probability distributions." (Giocoli, 2003, p. 317)

The introduction of Bayesian decision theory in game theory does not however come without any methodological and ontological difficulties. It is indeed, as will be explained in the first chapter of the thesis, traditionally adapted for decision situations under risk and not uncertainty, i.e. for situations under which the probabilities describing the lack of information are objective probabilities; they concern natural events and not events that are under the control of human mind. "Traditionally, Bayesian decision theory had been perceived as appropriate only to tackle exogenous, and not strategic, uncertainty." (Giocoli, 2003, p. 317) In strategic situations the uncertainty is related to the decision of the other players. This decision is explained by their perceptions, their intentions and beliefs, which are determined by the game, by the strategic decision problem they face; they are endogenous to the game. Uncertainty is therefore of a particular nature. Probabilities are subjective and not objective: they depend on the personal evaluation of the players of the elements in the other's reasoning process that are not known.

One element for the players to form probabilities with respect to other's choice is common knowledge (or belief) of rationality. This allows circumscribing the possibilities that the players believe about others' choices. In particular it induces that if the players believe that the other players are Bayes rational the probabilities attached to the other players playing a dominated strategy (i.e. a strategy that lead to a strictly lower payoff compared to other strategies, independently of the other's action) is zero. But in many situations the subjective evaluation of the players with respect to other's choices cannot be simply determined according to common knowledge or common belief of rationality. This is in particular the case for coordination games. Something from outside the game and the rules of the games and not determined by common knowledge or common belief of rationality must be added. Something within the player's reasoning process that appeal to their gut feelings, or to social knowledge for instance. Social knowledge must be understood in terms of the knowledge of some mode of interactions that generally occur in the situation described by the coordination game, of some way of doing and deciding that allow individuals to coordinate each other: to act in congruence. As will be detailed in chapters 2 and 3, this refers to conventions and institutions in real and social life that organize interactions among individuals of a common society, of a common group, and facilitate coordination.

6. The interest of the inclusion of psychology and players' reasoning process

The thesis is in the vein of the path stating that progress in our understanding of coordination in game theory need the analysis of players' reasoning process in strategic interactions which fall

under the frame of cognitive sciences (i.e. cognitive psychology and the philosophy of mind mainly). Indeed "Most existing research on knowledge about other people's knowledge falls in the area known as theory of mind, intuitive psychology, mind reading, or mentalizing, all terms for the mental representation of other people's mental states." (Thomas, DeScioli, 2014, p. 659) This thesis is explicitly integrated in this trend: chapters 4 will explain how the philosophy of mind and in particular the simulation theory is particularly relevant for game theory, and for understanding the knowledge that the players can have about each other in strategic contexts. Chapter 5 will then show how to integrate into a model of games simulation theory as the basis for players' reasoning.

The progressive rejection of psychology led to the incapacity to explain how and why players converge on equilibrium and thus how and why they are able to coordinate. However the rise of the epistemic program is integrated in a perspective that expresses the acknowledgment that the escape from psychology had achieved a stalemate. Chapter 1 will show that the reasoning process of the players in epistemic game theory is poorly integrated. The foundations of epistemic game theory being Bayesian decision theory, at the end players' beliefs and choices are mere representation of an effective decision that should have respected the axioms of Bayes rationality. It represents the choices and beliefs of the players and do not describe a decision or a process of beliefs formation (Binmore, 1993; Heidl, 2016; Hausman, 2012, 2000). Thus it does not describe or explain a reasoning process but solely describes its result. There is accordingly still an indeterminacy regarding this reasoning process. The way players form their beliefs and progressively accord their beliefs regarding each other's choices is still omitted from the analysis of game and of strategic reasoning; and coordination still unexplained.

Nevertheless "[I]f a game in the formal sense has any coherent interpretation, it has to be understood to include explicit data on the player's reasoning processes. Alternatively, we should add more detail to the description of these reasoning procedures. We are attracted to game theory because it deals with the mind. Incorporating psychological elements which distinguish our minds from machines will make game theory even more exciting and certainly more meaningful." (Rubinstein, 1991, p. 923)

The thesis is grounded on the argument that understanding the players' reasoning process in games necessitates first and foremost to explain how the players form their beliefs regarding each other's choices, but also each other's perceptions and beliefs and reasoning processes in a strategic context. One of the purposes of the thesis is to show that a psychological theory explaining the formation of players' beliefs is required to account for coordination, and that the Theory of Mind (ToM) offers an adequate psychological framework. As will be explained in chapter 1, one of the main problems of game theory and in particular epistemic game theory is a controversial ontology of the players' beliefs. Nothing explain how players form their beliefs regarding each other's choices, beliefs, perceptions or reasoning process, since these beliefs are supposed before the game to be already the result a rational reasoning process, at the end of which players beliefs are mutually consistent. These beliefs therefore are already supposed to translate the idea that the choices of each other's player are rational and are therefore at the equilibrium. The consistency view of rationality states that each player's beliefs are consistent each other's and with the equilibrium, which leaves unexplained the process by which the players converge to those beliefs (and hence, leaves unexplained how players coordinate).

Standard game theory remained anchored to the 'neoclassical' view of economics in which individual preferences are a priori given so that economics is free from any psychological determinant. Our contribution aims at showing that reconciling economics and psychology is the only way to prevent from this stalemate.

7. The organization of the thesis

From a historical and methodological perspective the first chapter portrays the evolutions of standard game theory from its creation, with classical game theory, to its modern version with epistemic game theory. The chapter will analyze the idea of 'solution concept' and how it impacts the way players are portrayed, their rationality and the resolution process of the game.

Classical game theory, following von Neumann and Morgenstern's and then Nash's respective contribution, is a framework of analysis in which strategic rationality is conditioned by the existence of a solution. The mathematical analysis of the solution, its properties (i.e. existence, uniqueness and stability), provides the core of the theory of games. Epistemic game theory, on the contrary, intends to put the players at the center of its analysis, and intends to provide a theory explaining how a player chooses her strategy in a strategic interaction, from her perspective, given her information. Epistemic game theory therefore offers a shift in standard game theory by investigating the players' epistemic states and modes of reasoning and no longer the solution concepts per se, its formal and logical properties (e.g. its uniqueness, stability or optimality). However it will be stressed in chapter 1 that both conceptions go with the SOR view of economics in which it is impossible to explain how and why a specific solution is reached. In other words the resolution process of the games and accordingly the reasoning process of the players towards the equilibrium and the solution are missing from the analysis.

The first part of chapter 1 will detail the construction of game theory as thought by von Neumann and Morgenstern and then Nash and the epistemological and methodological consequences these respective contributions have on the objective and vision of games of the game theorists onwards. In particular, while von Neumann and Morgenstern bring into game theory the view that a game is a mathematical object and that strategic rationality is equivalent to the existence of the equilibrium, Nash on the other hand brings into game theory the consistency view of rationality so that together they induce the application of the solution concept as an hypothesis of the models of games, without explaining why the players should conform to this solution concept. It will be argued that a game is therefore a mere representation of the choices of the players and does not offer an explanation of the choice they may make. When several equilibria exist, the problem is then to define a rational way to play – which is lacking since no explanation of the resolution process of the game is provided. Such theory is thus inoperative to offer recommendations to the players about how to play.

The refinement program that arose after the acknowledgement of the failure of the main objective of the classical game theory (i.e. to propose solution for games), induced the search for new justifications for equilibrium play, the proposition of new solution concepts more operative

than Nash equilibrium. Nevertheless this refinement program as will be explained reveals to be quite disappointing as it does not challenge the way strategic rationality is conceived, it remains focused on the mathematical properties of the different solution concepts proposed. It does not ultimately challenge the way players are portrayed in games as their reasoning process and the way they adjust each other to come to a solution are still missing from the analysis. On the contrary, epistemic game theory puts the players at the center of its analysis, and intends to provide a theory explaining how a player chooses her strategy in a strategic interaction, from her perspective, given her information (Aumann and Drèze, 2008). Epistemic game theory therefore supposedly offers a shift in standard game theory by investigating the players' epistemic states and modes of reasoning and no longer the solution concepts per se, its formal and logical properties (e.g. its unicity, stability or optimality). However, the epistemic program generates several questions and even inconsistencies so that it faces numerous challenges and requires the development of new "game theoretic concepts" (Perea, 2014, p. 20).

The main criticisms of the epistemic program focus on (i) the methodological consequences of imposing Bayesian rationality (Mariotti, 1995, 1996; Levy, 1995; Stalnaker, 1999) and (ii) the existence of common priors, which states that the beliefs of the players, prior to the game, are already equivalent (Kadane and Larkey, 1987; Gul, 1998; Gilboa, 2011; Morris, 2001). The existence of common prior received a great attention in the critical literature and both its philosophical and methodological dimensions have been tackled, however the existence of priors itself has been rarely stressed. We propose to fill this gap in this chapter. Much of the debate in EGT concerning the players' prior beliefs has focused on the nature of these priors, i.e. whether they are common or not, independent or not. Our criticism goes further, by questioning the existence itself of prior beliefs about each other's actions. We question the epistemic program in game theory that consists in justifying individual choices in strategic interactions by imposing rational constraints on players' beliefs prior to the game.

The first chapter thus intends to show that in the end, some questioning remains regarding the existence of a solution and how and why such solution can be reached. The possible answers to this questioning require investigating the coordinating processes in strategic interaction, which implies a different epistemology. It necessitates bypassing the mere logical and mathematical view of games and of strategic rationality.

The second chapter portrays from a historical and methodological perspective, the contribution of T.C Schelling, who is one of the first to open game theory to social sciences like economics, sociology or law. He grasped the challenges of game theory from an interdisciplinary perspective which is inherited from the RAND Corporation, in which the mathematical foundations of modern game theory are laid while being at the same time strongly based on interdisciplinarity and on experimental and psychological methodologies.

From his contribution we keep his ontological thinking on coordination; his view of a game as an interactive process in which the reasoning process of the players explains the solution of the game. Schelling sees games as situations of strong interdependence, which means that the players must accommodate each other, adapt their behavior to each other so that each situation of strategic interaction is inherently about coordination. In his view of games, the solution of the game is primarily determined by the way players react to each other; it is a discovering process.

The solution cannot be determined a priori but explained a posteriori by the way players reason and by the knowledge they have about each other, their intentions, beliefs and reasoning. It thus contradicts the standard view of game theory. Accordingly, Schelling is one of the first to think about answering the question of how and why a specific solution occurs in a game and from the angle of the players.

The interest of Schelling's work for our thesis is to show that, from an epistemic perspective, he proposes some paths of thinking about the way the reasoning process of the player can be integrated to broaden the formalism proposed by epistemic game theorists following Harsanyi's contribution. Schelling insists on the way the player's beliefs can become consistent with each other and on the mechanisms allowing such consistency, to be coordinated on an equilibrium without imposing a priori some prior beliefs consistent with the equilibrium. He shows how such beliefs – that are, prior to the game, free from any formal restriction – become aligned during the reasoning phase of the game, i.e. during its resolution. For that purpose, he integrates many psychological and sociological dimensions that impinge on the players' reasoning process. He also shows, which will be of particular importance in the remainder of the thesis, that it is primarily the capacity of the players to put themselves in the other's shoes, that is, to try the see the problem from the other angle, that is predominant in the resolution process of the game and in strategic reasoning. This is, in the modern account of cognitive sciences what is identified as the simulation theory and which will be developed in the chapter 4 of the thesis and integrated in a theory of games in the chapter 5.

A methodological consequence of his conception of strategic reasoning in games and of coordination is thus a departure from a purely mathematical conception of games and of the solution. The analysis is not focused on the mathematical representation of the solution (confined to the search from its properties and existence) but on the contrary on the conditions that led to a specific solution. The solution is not a priori assumed but searched: this is the purpose of strategic reasoning and the nature of coordination in games. Schelling indeed emphasizes in his work the inadequacy of the mathematical tools of game theory to account for the convergence of players' beliefs, i.e. for their capacity to coordinate. Schelling therefore provides very fruitful conceptual and methodological answers to some of the major difficulties that contemporary game theory faces. His explanation of the elicitation of player's beliefs by their capacity of putting themselves in the other's shoes for instance, provide a methodological ground to overcome some of the main shortcomings of epistemic game theory regarding the explanation of the origin of the players' prior beliefs. Schelling proposes a methodological ground to endogenize the players' beliefs which is not the case in epistemic game theory.

Chapter 3 intends to show that a formal game theorist trained in mathematics and keen to enrich game theory from a formal point of view (M. Bacharach) shared a common social philosophy with Schelling, and succeeded in incorporating many of the conceptual and methodological solutions that Schelling's offers for non-cooperative game theory. Bacharach's work on game theory indeed focuses on coordination. To understand coordination and to explain player's capacity of coordination in games he incorporates in games the player's perception of the games and of the strategic situation they face. He opens up a research program in game theory on framing that has for consequences to translate standard games into framed games, i.e. games in which the perceptions of the players are integrated and which impinge on their reasoning and

therefore on the solution of the game. Like Schelling, Bacharach's grasps the problem of coordination and of the explanation of how and why a specific solution occurs by the integration of player's reasoning process. The main determinant of player's reasoning will therefore be, in his framework the knowledge and beliefs the players have about the other's player perceptions and beliefs. He integrates this concern in a model of games and intends to formalize the process of convergence of players' frames and beliefs towards the equilibrium. He is one of the earliest game theorists to challenge the epistemic foundations of game theory and in particular introduced new formal tools to analyze common knowledge. Common knowledge as it will be explained in chapter 1 is a methodological device to explain how players form beliefs on the other's players' beliefs and choices. But it will be stressed that such assumption at best is in many cases insufficient to come to a conclusion of the choices that the other will make and in the worst case contradictory with some axioms of rationality in game theory.

Bacharach's work on game theory began at a very specific and important moment in the history of game theory: the refinement program. However as will be showed in the first chapter, the refinement program remained mainly confined to propose mathematical answers to the problem of multiplicity of Nash equilibrium while Bacharach was on the contrary already assessing the philosophical and methodological foundations of non-cooperative game theory tackling the problem of the definition of strategic rationality in games. This is in this perspective that Bacharach developed the Variable frame theory (VFT) and then the Team Reasoning theory that will be both detailed in chapter 3.

Bacharach early criticized the assumption of rational expectations which ultimately precludes from understanding the coordinating processes that occurs in games, pointed out some foundational shortcomings of game theory and for that purpose used formal methods from economics, philosophy and logics. He adopted an interdisciplinary approach of game theory allowing him to enrich game theory by pointing out it insufficiencies. He assessed the philosophical foundations of game theory from an interdisciplinary perspective. His interdisciplinary approach of economics, relies on philosophy, psychology, sociology, and evolutionary biology, to propose "new coherent foundations" for game theory through the inclusion of new epistemic fundaments in games; relying on the players' perceptions of the game and of the other players, and on their reasoning process toward the equilibrium and the solution of the games. The interest of Bacharach's work from a methodological point of view is to show that some enrichment proposed by Schelling can be formalized in games to improve game theory and to challenge the mere mathematical view of solution concept. He offers a methodological ground with his formal contribution to the integration of a theory of mind (the simulation theory) in game theory that will be presented and developed in the chapters 4 and 5 of the thesis. The integration of the simulation theory will in particular explain the way players form their prior beliefs regarding each other's choice and the way they revise these prior beliefs to become consistent each other.

Chapter 4 examines a contribution within the social sciences: the philosophy of mind and more specifically the ToM which can provide a methodological ground to explain how the players form their beliefs about others' beliefs, reasoning, perceptions, intentions or behaviors. It can therefore provide the ground to integrate player's mental states and reasoning process in strategic interactions an ultimately to explain coordination, i.e. how and why an equilibrium occurs.

The ToM encompasses different approaches from neurosciences to sociology, cognitive psychology and social psychology in order to investigate the way the individuals attribute to themselves and the others some mental states like preferences, intentions, desires and beliefs. The ToM therefore investigates the way the players may be acquainted with the other's way of reasoning and beliefs, and explains the way the players can have some knowledge or beliefs about the other players' intentions, beliefs, etc. It thus allows an individual either to predict or to explain someone else's choice. Besides introducing a mechanism of third person mindreading, has for consequence to eliminate the controversial requirement of the existence of prior beliefs in epistemic game theory. We will explain the different theories within the ToM, their respective interest for game theory, and their contribution for analyzing coordination in game theory. We will specifically detail the simulation theory and will explain why it is this theory that will be integrated in the theory of games proposed in the chapter 5.

Meanwhile, the chapter 4 will expose a theory called mindshaping which explain prior to the theory of mind how the individuals of a common society are shaped by institutional devices, i.e. some ways of interacting, so that a degree of congruence occur among them and their understanding of situations, their perceptions or beliefs about other's behaviors. In other words, the same social collectives, or institutional devices have shaped the mind of the individuals. This explains how the process of attribution of mental states to someone else may be accurate enough to lead to good predictions or good explanations of the others' behavior. The ToM offers a framework explaining our ability to coordinate with others based on a shared capacity to correctly represent each other's mental states and beliefs, and thus to accurately anticipate the behaviors of others. Different theories of mindreading – the capacity to 'read' each other's minds – have been suggested in the literature, though they systematically start from the principle that the key to human coordination is our ability to form complex epistemic states about each other's mental states, especially beliefs and desires (Zawidski, 2018). This is also the approach implicitly endorsed in the epistemic program in game theory, according to which the analysis of strategic behaviors can be reduced to Bayesian decision theory, while taking hierarchies of beliefs as an input of decision-making. Such representation of complex mental states is in some cases, for instance when the coordination context is familiar unnecessary and in some other cases too complex without any mindshaping devices beforehand. Mindshaping is necessary and both mindreading and mindshaping are complementary dimensions in social cognition, i.e. for human interactions.

Both the ToM and the mindshaping theories will provide the ground for the integration of the players' mental states and reasoning process in games to justify coordination. They will provide the methodological ground to explain the origin of the players' prior beliefs, i.e. their beliefs about others' mental states like their beliefs and then, their behaviors in games.

We thus suggest in the chapter 5 building an alternative theory of games in which the primitive of the theory is Bacharach's Variable Frame Theory and which is grounded on a descriptive theory of framing in games that explains the integration of players' subjective perceptions of the games. We then specify an axiomatic characterization of rational choices in games in the presence of players able to simulate the reasoning of others.

We assume that a proper Bayesian explanation of rational choices in games requires introducing a psychological theory of belief formation. We suggest that an operative theory of rational choice in games – i.e. a theory explaining how a player should rationally choose, given the nature of the game and of the other players – should:

i) consider the problem faced by each player,
ii) define a mechanism consistent with the rationality of the player that gives her a reason to play a specific strategy,
iii) assess the final outcome, from our perspective as theorists.

The first step consists in describing the players' perception of the game, i.e. their beliefs about the types of the other players and about their strategies. The second step explains how the players choose, given their perception of the game. The third and last step consists in defining a solution concept to characterize the resulting strategy profile. The fundamental issue of epistemic game theory [EGT] is that the beliefs that the players were supposed to use at step (i) are defined from the choice observed by the theorist at step (iii). Indeed, once we identified an equilibrium, we can build the ad hoc prior beliefs such that Bayes rational players should play this equilibrium. This however does not give any practical advice to a player. EGT only suggests that, if Bayes rationality is common belief, then the possible sets of prior beliefs the players have at their disposal is reduced (to common priors for instance). Endorsing this approach makes deliberation 'vacuous' (Levi, 1998, p. 181). Indeed, deducing my beliefs from the equilibrium profile of strategies implies that the premise of my deliberation (my prior beliefs) is deduced from its result (my actual choice). We suggest integrating the ST in the model to explain how players form their beliefs in step (i): we will then assume that players maximize their expected pay-off in step (ii), given their beliefs in step (i), and characterize the resulting solution concept as a subjective belief equilibrium.

The contribution of our thesis to the field of game theory is mainly a methodological and ontological change in relation with both the contents of game theory and the concept of game. Game theory is understood by us as a framework of analysis, the core of which is the analysis of strategic interactions between active agents and not a mere mathematical theory describing individual choices at the equilibrium. Games are considered as situations of strategic interdependence which involve a reasoning process towards a solution and not a mere mathematical representation, i.e. a mathematical description of an individual choice. From a methodological point of view this mathematical representation is grounded on the hypothesis of exogenously and a priori given individual. It presupposes the independence of economics from psychology and more generally from the other social sciences.

Our thesis proposes to change the ontology and methodology of game theory, appraising games as the understanding of the players' strategic reasoning process intending to determine why a specific equilibrium of beliefs, intentions and plans of actions, occurs. In that perspective the stalemates that game theory faces both from a positive and a normative point of views can be solved. We claim that the analysis of games should involve the study and the determination of the reasoning process that lead the players to a specific outcome, i.e. to a specific solution. The

solution of games should not therefore be a hypothesis of the game but a result and should thus be explained.

We locate the analysis of coordination at the center of our analysis. We consider coordination as a process: the progressive congruence of players' beliefs about each other action, and accordingly of their perceptions, intentions, or plan of actions. We assert that understanding the process of coordination allows understanding strategic reasoning and ultimately to provide new answers to the indeterminacy problem of game theory which is one of the stalemates that game theory faces and which underscores its positive and normative difficulties.

The interest of the thesis is to gather contributions that stress the intersubjective dimension of games, i.e. the dimension allowing players to reason about others' choices and to gather contributions both coming from within and outside game theory. Such intersubjective dimension in player's reasoning process is underdetermined. Our contribution is based on an interdisciplinary approach which suggests a change of epistemology and not only a change of methodology to overcome this shortcoming.

We finally enlarge the research program implemented with the integration of a framing approach which includes many dimensions not involved immediately in the usual framing literature, in particular the adding of a reassessment of the kind of intersubjectivity involved in strategic reasoning. The thesis shows that the integration of framing can be seen as the beginning of a research program which can be centered around a critical and specific assessment of the understanding and the modeling of strategic reasoning. It is at the core of this understanding and this modeling that intersubjectivity appears in game theory. This intersubjectivity combines the foundations of game theory and a new ontological approach.

Chapter 1

A critical assessment of the evolution of standard game theory

1. Introduction

Game theory can be defined as a mathematical theory of strategic interactions between rational decision-makers, involving sets of 'players', 'strategies' and 'payoffs'. The foundations of game theory on which classical game theory are grounded also concern the analysis of the solution concept, i.e. the rigorous and mathematical definition of the outcome of a strategic interaction within some defined rules of the game, as the existence of such a solution concept entails a defined rational behavior (e.g see von Neumann and Morgenstern contribution in the section 2.1). The object of classical game theory is indeed "to *propose* solutions for games" (Sugden, 2001, p. 115; emphasis in original), and the two usual criteria for distinguishing between acceptable and non-acceptable solution concepts are (i) that the strategy profiles under consideration (the equilibria of the game) should be consistent with the rationality of the players, and (ii) that the set of equilibria should be relatively small, yet non-empty (Sugden, 2001, p. 115).

Classical game theory, its epistemology, i.e. its formalism, methodology, codes and research program, is built on the works of von Neumann and Morgenstern (1944) and even more so on Nash (1950-51). More specifically, classical game theory is identified by Sugden (2001) as the game theory that prevailed in economics in the late 1980s, and which had for ground the search for a *solution concept*. A search for a solution concept entails finding "a rule which applies to all games in some general class and which, for each such game, picks out one or more combinations of strategies as *the* solution or solutions." (Sugden, 2001, p. 115) Thus, one of the main objectives of classical game theory is, according to Sugden (ibidem), to discover the principles of equilibrium selection. The main principle of equilibrium selection in classical game theory entails the combination of instrumental rationality (players attempt to maximize their given individual payoffs according to their knowledge or beliefs of the game and other's choices), and common knowledge of the game (which entails knowledge of the set of players, of a given payoff matrix, and of everything that can be logically deduced from the structure of the game; e.g. see Colman, 2014, p. 36). To go a step further, Bacharach claims that classical game theory is identified by the

adoption of Nash equilibrium as *the* solution concept: "classical game theory = individual choice theory + Nash equilibrium." (Bacharach, 2001, p. 1) Classical game theory raises many difficulties when applied to economic situations, as the condition of uniqueness of equilibrium, a necessary condition of the existence of a solution, is rarely reached. This "state of the art" led to the birth of the refinement program which is an attempt to achieve the conditions of uniqueness of equilibrium of games. The refinement program began with the search for the stability conditions and the attempt to reject the unstable Nash equilibria thus eliminating some of the Nash equilibria in cases of multiplicity (Hargreaves Heap and Varoufakis, 2004 [1995], p. 80).

In the recent history of game theory, the refinement program, which arose in the 1970s, is often acknowledged to be one of the richest periods of development from a theoretical point of view (see Giocoli, 2003). The profound debates it generated constitute contemporary game theory. Besides, it corresponds to the period of the progressive introduction of game theory in economics. This is mainly explained, according to Giocoli (2003), by the introduction of uncertainty in game theory, i.e. the formalization of games in incomplete information: a theoretical framework much more suitable for economic contexts. Many scholars stress the inapplicability of game theory in economics because of the completeness of information: a state of knowledge that cannot describe economic contexts. To a certain degree, the refinement program contributed to the development of enthusiasm towards game theory in economics. It began, as explained by Giocoli (2003, p. 338), with two contributions from Selten in 1965 and 1975, but was fully developed in the 1980s.

Within the refinement program Harsanyi's contribution opens the doors to the introduction of incompleteness of information and later the birth of what is now identified as Bayesian game theory or epistemic game theory (the two labels are interchangeable and refer to the same research program in game theory; see Lecouteux, 2018b). Bayesian decision theory – in the sense of Savage (1954), i.e. as the maximization of subjective expected utility – was first introduced in game theory to discuss issues of incomplete information (Harsanyi, 1967–1968), while keeping the classical approach of studying solution concepts. It was only in the 1980s that game theorists started to analyze the Bayesian rationality of equilibrium play, by assuming that players maximize their expected utility, given some subjective beliefs about the actions of the others, i.e. they attempted to explain how and why a specific equilibrium could occur according to the way players appraise their decision problem, i.e. according to their knowledge of the game and of the other players and according to their beliefs regarding these other players (i.e. their beliefs regarding the other's knowledge, reasoning, beliefs or choices). On the contrary Von Neumann and Morgenstern (1944, p. 19), defined probability "as frequency in the long run" and therefore as objective probability (for a more detailed explanation of the difference between subjective and objective probabilities see the section 5.1 of this chapter).

The stalemate of the relationship between equilibrium and rationality that game theory faces is a hangover of the interwar so-called "neoclassical economics" (Giocoli, 2003, p. 343). However, as already emphasized in the introduction, being embedded in the SOR view of economics game theory tackles the issue of such link between rationality and equilibrium in a very different way than the SOF view of neoclassical economics. Recall that the SOF states that economics attempts to provide an explanation of how and why an equilibrium is reached and how rationality explains such a path to the equilibrium while in the SOR view the equilibrium is postulated and never

explained. In other words as this chapter will shows, the equilibrium of a game is an hypothesis of the theory not the result of a play. Rationality therefore does not explain the equilibrium play. According to Giocoli (2003), such divergence of account between the SOF and SOR view of economics may explain why game theory has failed over many years to pervade economists. Game theory has remained in mathematicians' hands "who governed its development according to the sociology of their own field, for example, rewarding the formal improvements and disdaining the applications." (Giocoli, 2003, p. 343) This chapter will provide an illustration of this view. As will be emphasized, even after its success in economics after the 1960s, and what is called the refinement program, game theory remained largely a mathematical theory, governed by mathematical developments and methods and research agenda. The rise of the SOR view, brings in economics "the triumph of the consistency view of rationality" (Giocoli, 2003, p. 346) concordant with the axiomatic approach.

The epistemic program in game theory, which introduced incomplete information situations in strategic interactions and the acknowledgement of the player's hierarchy of beliefs (their beliefs regarding other's choices, others' beliefs, the impact of others' beliefs on the others' beliefs etc. ad infinitum) into the games structure, was apparently an attempt to overcome the inability to explain how and why an equilibrium exists. The concept of the Players' hierarchy of beliefs opened the door to the introduction of the players' reasoning process in games and therefore to an explanation of how the players could reach a stable state in which every player's beliefs regarding each other's choice was consistent.

However, as will be explained in this chapter, being faithful to the axiomatic approach to epistemic game theory, i.e. to the consistency view of rationality, does not meet this objective.

The 1940s witnessed the beginning of the general transformation of economics into a mathematical discipline (Weintraub, 1992, p. 3) and game theory is a particular example of the triumph of the mathematical approach and its code. The adoption of the axiomatic approach in game theory is the marker of this fact. Whatever the major contributions in the history of game theory: the publication of the TGEB in 1944, the work of Nash and the birth of Nash equilibrium in 1950-1951, the work of Harsanyi in 1975 and the introduction of incomplete information in game theory (Mirowski, 1992, p. 115), and the epistemic program of game theory; these contributions - remain faithful to the axiomatic approach.

The mathematical approach of game theory runs alongside the axiomatic approach resulting in two tendencies or even two schools.

Regarding the axiomatic approach that has been imported into game theory, in particular thanks to von Neumann, Giocoli stresses the difference between the approaches of the Bourbakists and of Hilbert. According to the former, mathematics is seen as an "autonomous subject that needed no input from the real world and that was completely separated from its applications"; they solely attempt to "exploit the axiomatic technique to derive general formal structures devoid of any connection with the physical world." (Giocoli, 2003, p. 373) In contrast, later mathematicians suggested that theories must be connected to the reality: "Hilbert's axiomatic method aimed at developing mathematical theories that captured the empirical substrate from which the axioms had been drawn"(Giocoli, 2003, p. 373).

Modern game theory is strongly influenced by the Bourbakists' account of the axiomatic approach while von Neumann and Morgenstern's view of the axiomatic approach was more in line with the Hilbertian view, considering that the axiomatic method was complementary to the empirical method. When delivering the axiomatic account of a game von Neumann and Morgenstern (hereafter vN/M) attempt to provide "an exact formulation for intuitively-empirically-given ideas" (vN/M, 1953, p. 76).

The different views of axiomatic constitute for Giocoli another reason for the relative neglect of emerging game theory in the 1950s. Indeed, for Giocoli it is only after the Bourbakist's account of General Equilibrium Theory (GET) changes "the image of economics" so that economics became consistent with the SOR view of economics that game theory gained importance in neoclassical economics (Giocoli, 2003, p. 375).

In the consistency view of rationality "[w]hat the theorist has to do is, in short, to pre-determine the solution concept he/she is interested in and then identify the game equilibrium by proving its existence under the assumption that the players are rational, that behave consistently with respect to that particular solution concept. The equilibrium is then taken to embody all the positive and normative content of the strategic situation." (Giocoli, 2003, p. 213)

In that perspective Bicchieri (1993, p. 33) refers to the "centrality of the notion of equilibrium" in game theory. This chapter intends to show how this premise has deep methodological and philosophical impacts in game theory at every stage of its development. In particular it will be stressed that this premise is responsible for the numerous stalemates that game theory faces and which will be identified in this chapter.

The difficulty remains that in the consistency view of rationality, which is equivalent to assuming the existence of a solution without explaining how and why such a solution can be reached, is that even if such a solution exists, as an equilibrium point exists (see Bacharach, 1994), it cannot be certain that the players can reach the solution as no principle of individual decision-making justifies that such solution can be computed by the players (Giocoli, 2003, p. 213).

Strategic rationality should not be understood as a mere consistency state but should "refer to the entire reasoning activity that intervenes between the receipt of a decision stimulus and the ultimate decision, including the manner in which the decision-maker forms the beliefs on which the decision is based" (Bimnore, 1987, p. 181). It will be shown that the consistency view of rationality inherited by the Nash's program precludes the integration of such reasoning processes.

Strangely, the conquest of rationality as a consistency relation is explained by Giocoli (2003, p. 346) as the incorporation of mental variables, such as players' beliefs, i.e. what we more generally refer to as mental states in this thesis. When the constraints imposed on rationality have been extended to account for the players' beliefs, thanks to Savage's subjective expected utility theory, the consistency view of rationality becomes the core of the post WWII neoclassical economics even today. Rationality has been extended to beliefs and expectations so that being rational now means handling rational beliefs and rational expectations (Giocoli, 2003, pp. 346-47).

2. On the foundations of classical game theory

The exegesis of the contributions of the founders of classical game theory respectively von Neumann and Morgenstern (vN/M)on the one hand and Nash on the other, help to stress the belief that never faded that strategic rationality is equivalent to the existence of a solution concept. The solution concept is applied as a norm and strategic rationality means playing the equilibrium, i.e. playing as the rules of the games set by the theorists define a rational play. Players analyze the game like outside observers, like game theorists, and strategic rationality is playing by? the rules of the game. It helps to stress that it is mathematical rigor that drives the creation of game theory and that it is the rules of the mathematical community that will continue to govern the contributions to game theory into the future determining the different amendments that will augment classical game theory until the final stage: the epistemic program in game theory.

2.1. Von Neumann and Morgenstern's contribution

The first publication of von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (TGEB) in 1944 marked the birth of modern game theory though it took decades to penetrate the economic community. The purpose of this book was to propose a theory of rational behavior in strategic context from a normative point of view, i.e. to propose the "handbook for the good player" (Giocoli, 2003, p. 211). To fulfill this objective the authors attempted to define rationality from a normative point of view; from the objective characteristic of what they call a game (which will be described in detail below). Defining a rational behavior requires definition of an appropriate solution concept (Giocoli, 2003, p. 260), free of any *subjective* and *intersubjective* considerations, based only on objective characteristics and more specifically on numerical characteristics, i.e. basically payoffs. The authors claim only a normative power for this solution concept and not a prescriptive or descriptive power. Though they assert that what they call a game and the way they describe a game (which will be detailed below) captures "the essential" of reality (see Giocoli, 2003, p. 246).

The aim is to define rational behavior according to a specific solution concept, solely dependent on the formal, i.e. objective, characteristics of a game free of any mental variable (what we more generally identify in this thesis as mental states which would necessitate an exploration of psychological aspects in their discourse and formalization is in line with the formalist revolution prevailing in mathematics and in particular the new axiomatic approach to which von Neumann is faithful (Giocoli, 2003, p. 216). There is no claim regarding the positive side of this solution concept, i.e. no claim regarding the realism of this solution and its capacity to describe real strategic situations.

2.1.1. An objective characterization of strategic rationality according to the maximin criterion

The authors attempt to offer in the TGEB "mathematically complete principles which define "rational behavior" for the participants in a social economy, and to derive from them the general characteristics of that behavior." (von Neumann and Morgenstern, 1953[1944], p. 31) They supposed that economic interactions and economic behaviors are like in game theory. For them "the typical problems of economic behavior become strictly identical with the mathematical notions of suitable games of strategy." (ibid, p2) They aim to revolutionize the social sciences, asserting that the neoclassical mechanistic view of economics is outdated and that game theory constitutes its replacement.[1]

Von Neumann claims an affiliation with Hilbert's formalism (Léonard, 2010, p. 234). Hilbert sees axiomatization and true rigor as follows: axioms must be complete, independent and consistent so that all the theorems can be derived from these axioms, the removal of one of the axioms would entail proving that one of the theorems following from the axioms would be impossible, and finally the theorems deriving from these axioms cannot be contradictory (Léonard, 1992, p. 41). When von Neumann speaks of Hilbert's axiomatics he stresses the "internally closed procedure which operates according to fixed rules known to all mathematicians and which consists basically in constructing successively certain combinations of primitive symbols, which are considered 'correct,' or 'proved' . . . a combinatorial game played with the primitive symbols" (von Neumann, 1984 [1931], p. 62). Thus, in the TGEB games are defined axiomatically as requiring completeness, to be non-contradictory, and to be independent properties of axioms (see the Chapter 2 of the TGEB).

For von Neumann, the axiomatic method "follows the classical lines of obtaining an exact formulation for intuitively – empirically – given ideas" (von Neumann and Morgenstern, 1953[1944], p. 76). He claims that he succeeded in showing that it "is possible to describe and discuss mathematically human actions in which the main emphasis lies on the psychological side" (ibid, pp. 76–77); without any reference to psychology just with mathematical and objective features, i.e. quantitative features. This is exactly what is proposed in the TGEB.

It is interesting to note that the French vision of game theory is in opposition to this postulate. Borel indeed considers that the mathematics of games cannot prevent an incursion of psychological constraints (Léonard, 1992, pp. 45-46-47) and games of strategy are too complex to be embraced in all of their dimensions by a simple mathematical analysis; a theorem determining

[1] The way vN/M sees modeling and which prevails in the TGEB is that setting a new theory requires an heuristic stage allowing a move from non-analytical and "commonsense considerations" (Giocoli, 2003, p. 248) to formal and mathematical methods. The next step is to provide a general and formal theory which to be rigorous requires conforming to axiomatic methods. Then one should lay down a rigorous and conceptually general formal theory. Such a theory should be applied to the resolution of some elementary problems to test the conformity of the theory and then to more complicated problems to again test its conformity. Finally such a theory must be employed for predictions (von Neumann and Morgenstern, 1953[1944], pp. 7-8).

a good way for a player to play cannot be sufficient (Léonard, 1992, p. 48; see also Léonard, 2010, p. 61).

2.1.2. The solution concept

The maximin solution concept proposed in the TGEB translates the idea that the best strategy for an individual player must secure a minimal outcome whatever the other player's choice is. Such a strategy must guarantee a certain level of payoff when the individual player has no clue of what the other players' choice might be (Péréa, 2013, p. 5). Even in the worst scenario a minimal level of payoff for the other player must be guaranteed.

More specifically, with a two player game where g(x,y) is the payoff induced by the choice x of the player 1 and the choice y of the player 2, player 1 must secure for herself $g(x,y) = \max_x \min_y g(x, y)$, a payoff that is at least equilavent to the best of all of the worst payoffs that she could get according to player 2's choice. The player 2 on the other side, attempts to secure for herself $g(x,y) = \min_y \max_x g(x,y)$ a payoff that is at least equivalent to the worst of all of the maximal payoffs she could get according to player 1's choice (Giocoli, 2003, p. 222). Player 2 attempts to ensure that the player 1 will not earn more than $\min_y \max_x g(x, y)$. Both payoffs, $\max_x \min_y g(x, y)$ and $\min_y \max_x g(x, y)$ are called the security-level that the players can get. Von Neumann declares that the payoff $\max_x \min_y g(x, y)$ of the player 1 is the best outcome that she can obtain if the other player happens to discover, i.e. to correctly anticipate her choice, in the same way as $\min_y \max_x g(x, y)$ is the best payoff that the player 2 can earn if her choice is found out by player 1 (Giocoli, 2003, p. 223). Indeed, for him "it makes no difference which of the two players is the better psychologist, the game is so insensitive that the result is always the same" (von Neumann and Morgenstern, 1953[1944], p. 23). The value of the game depends on this maximin criterion. VN/M consider that if both players play well they can each guarantee themselves the value of the game, i.e. $\max_x \min_y g(x, y)$ and $\min_y \max_x g(x, y)$ (Giocoli, 2003, p. 264). And for vN/M the game is strictly determined when this solution exists, i.e. when $\max_x \min_y g(x,y) = \min_y \max_x g(x,y)$ (Giocoli, 2003, p. 267). For these authors, a solution always exists for the game (at least for a two persons and zero-sum games) when $\max_x \min_y g(x,y) = \min_y \max_x g(x,y)$.[2] It means that, when following the maximin criterion. The choice of both players must be mutually compatible (Giocoli, 2003, p. 223).

The solution is therefore independent of any inferential capacity or psychological disposition to infer other's choices. This neutrality of the solution of the game with respect to the players' inferential capacity guarantees an objective definition of a strategic behavior and of rationality in games "without having recourse to the players' beliefs, expectations or intuitive powers." (Giocoli, 2003, p. 223) Hence, a solution for vN/M is a "list of the mathematical principles defining the rational behavior of an individual playing a specific game." (Giocoli, 2003, p. 253) A

[2] Léonard (2010, p. 64) however claims "[a]lthough, in general, $\max_x \min_y g(x, y) \leq \min_y \max_x g(x, y)$, it is not generally true that the equality holds."

solution concept specifies "a set of rules for each participant which tell him how to behave in every situation which may conceivably arise" (von Neumann and Morgenstern, 1953[1944], p. 31). This is therefore a specific strategy.

The circularity problem of a game situation, i.e. the fact that your outcome depends not only on your decision but on the decision of the other who is in exactly the same position as you, explains why von Neumann (1928) first attempts to define a game in its objective form, getting rid of any subjective reasoning and solely depending on material payoff. In this perspective a strategy is described as a complete plan of action for all of the game (von Neumann and Morgenstern, 1953[1944], p. 79) whatever the actions of the other player. The player must envisage every possible situation in the whole game. It implies that the player "enters the play with a theory worked out in detail" (Von Neumann, 1928, p. 18) about the way the game is going to be played. The information provided by the rules of the game, or by the game structure is sufficient for the player to determine her "theory" in von Neumann's words. No further information such as a prediction of the other's choice, her belief or intention is needed. The player can enter the game with a detailed plan of action for the whole game "in the absolute ignorance of the other players' choices (ibid, p 19)." Contrary to Nash's solution concept, as we will see in section 2.2, the player's choice of strategy does not depend on her knowledge of the other players' choice. The definitive choice of strategy defined for the whole game is therefore made in complete unawareness. This is for Giocoli (2003, p. 260) the hallmark of von Neumann's attempt to describe a game as objectively as possible. By this means von Neumann frees the game of any reference to psychology, and to player's mental variables such as their beliefs or their expectations.

A game in the TGEB therefore remains in a "black box", free of any consideration of the way the game is played, instead solely described by a set of strategies which are transformed into a set of payoffs. Such an objective description of the game is consistent with the formalist method and its rigor (Giocoli, 2003, p. 279). In von Neumann's words, "[n]othing is left of a 'game of chance'", and "everything takes place as if each of the players has his eye on the expected value only" (von Neumann, 1928a, p. 21), i.e. on her payoffs understood in terms of expected utility, only.

However, as von Neumann (1928b, p. 26) declares, chance remains "an intrinsic part of the game itself" (in Léonard, 1992, p. 43). This chance parameter in games persists in the mixed aspect of players' strategy, i.e. in the probabilities attached to the other's choice even though he has integrated "the effect of the chance moves on the player's payoff, f(.), by replacing the latter with his expected result g(.)." (Léonard, 2010, p. 63) Thus while the probabilistic dimension of a game has been discarded by being introduced in the expected value of players' payoffs, the dimension of chance remains with the probability distribution that sets the players' strategy (Léonard, 2010, p. 65). The stochastic dimension of their strategy prevents the players from being 'discovered by the other; it is a form of protection (Giocoli, 2003, p. 224).

It is very important to note that the probability distribution employed in the definition of a strategy is an objective probability distribution, i.e. long run frequencies, because von Neumann would have denied that a player can "theorize" about the other's belief or choice as this would

require the introduction of mental variables in games. Probabilities can by no means be subjective as in Bayesian game theory.

2.1.3. Strategic rationality

While vN/M tend to characterize rationality according to an "appropriate solution concept" Giocoli points out that rationality is, in their work, "the output, and not an input, of their analysis." (Giocoli, 2003, p. 211) A rational behavior implies following the minimax criterion. Thus, rationality means achieving the best possible outcome without any knowledge regarding the others' choice and without any precise expectation regarding the other's choice, i.e. in a context in which every choice is admissible, there is no difference between a rational or a non-rational one. Rationality is defined as a mode of behavior that gives the player full control of the issue of the game, they have only to refer to the structure of the game and do not have to infer others' choice. As vN/M declare "we formulate the principles and justifications for rational behavior, every possible conduct of the other players must be taken into account. In other words, if we aim at demonstrating the superiority of rational behavior over other kind of conduct, its characterization must necessarily embody the rules of conduct required to face every possible situation, including those where the other players act irrationally." (von Neumann and Morgenstern, 1953[1944], p. 32) Rationality is therefore defined objectively, i.e. independently of players' psychology, without any reference to the players' mental states, i.e. to the players' beliefs expectations, preferences, intentions, etc. "No player (or players) tries to predict what other players will do by putting him/herself in their shoes and thinking about what they might be thinking, and so on." (Brandenburger, 2010, p. 2)

In the TGEB, no belief of any sort with respect to other's beliefs, intentions, or choices enter the players' reasoning, even the belief that the other is rational in the sense defined by the TGEB. Players do not believe that the other can be shaped to follow the principle of a good play established by the maximin criterion. This frees the analysis of strategic behavior in games from any notion of psychology or any concern for players' mental states or mental variables. They escape from the problem of perfect foresight that is, on the contrary, central for Nash equilibrium.

vN sees human interactions as deprived of any "moral or psychological connotation and completely a-temporal." (Giocoli, 2003, p. 223) Human interactions must be "reduced to abstract forms" (ibidem) i.e. to the quantitative and mathematical structure of the game and the definition of players' strategies which specify for the whole game. vN/M did not attempt to describe the actual behavior of players in games; there was not any positive content in the players' strategies. On the contrary they assert that all of the strategical dimensions of a game can be summarized by the payoffs, i.e. by "a set of numbers" (Giocoli, 2003, p. 225).

Nevertheless, according to Giocoli (2003, p. 277) vN/M's account of strategic rationality provides an explanation of 'how and why' players adopt the maximin strategy. The explanation of how and why the players must adopt the minimax strategy is that it secures to them at least a

minimal payoff, the best payoff in the worst scenario. This does not correspond to the contemporary consistency view of rationality inherited from Nash's program and unambiguously adopted going forward. VN/M are therefore still in line with the neoclassical account of rationality in terms of a SOF before WWII and not of a SOR (as the consistency view) (Giocoli, 2003, p. 207).

VN/M's explanation of how and why the players plays the minimax/maximin strategy, is for Giocoli reinforced by the technical proofs they give to support their solution concept. To prove the minimax theorem vN/M rely on both what are called the direct and indirect proof methods. The direct proof method proposes to explain how and why a player should follow the minimax strategy; which for Giocoli (2003, p. 259) means that they still belong to the SOF view of economics. It necessitates a calculation in order to prove the existence of the mathematical object under study: "the existence of the object is proved by the fact that we know how to 'calculate' it." (Giocoli, 2003, p. 270)

The indirect proof method, on the contrary, is related to the formalist approach and entails a SOR view of economics as favored in the consistency view of rationality. This method entails "supposing that the object whose existence is asserted does not exist and then showing that this assumption leads to a contradiction. This logically proves that the desired object does exist, but no technique is provided by which it can actually be determined." (Giocoli, 2003, p. 270)[3] This is not a positive validation of the existence of the object under study and consequently, this is not a positive validation of the existence of a solution.

One important dimension of the formalist approach and the consistency view of rationality is its tight link with the fixed-point theorem, which relies on the indirect proof method. It is interesting to note that the minimax theorem initially does not rely on the fixed-point theorem of Brouwer (Léonard, 2010, p. 65).[4] This confirms that the minimax theorem still partly conforms to a SOF view of economics, as argued by Giocoli (2003).

This explanation of strategic rationality and the tendency towards a positive account of strategic rationality in games is however threatened by the extension of the maximin strategy from two-person zero sum games to n-players games and non zero-sum games which is highly problematic in both cases.

Non zero-sum games cannot be reduced to zero-sum games (Mirowski, 1992, p. 139) because the existence of a saddle point guarantees that the maximin and minimax strategies of the players coincide in 2-players zero-sum games; this is no longer the case in n-players non zero-sum games (Mirowski, 1992, p. 140). To apply the minimax solution concept to n-players games, vN/M attempt to divide games in two phases: (i) a first stage, a "fictitious play", setting the coalitions and then, (ii) a second stage which is the game between the coalitions in order to reduce the n-players game into a two-player game (von Neumann and Morgenstern, 1953[1944], pp. 505-506).

[3] For more details on the indirect proof method see Giocoli (2003, pp. 263-76)

[4] The misapprehension of the proof of the minimax theorem of von Neumann in 1928 is, according to Léonard (2010, pp. 66-67), explained by von Neumann himself because in his equilibrium growth model of 1937 he makes the formal link between the fixed-point argument of the growth model and the minimax theorem.

In the first stage each coalition is defined by its "unique value" and then the coalition plays against the remainder, the players that did not enter the coalition (Mirowski, 1992, p. 140; referring to von Neumann and Morgenstern, 1953[1944], p. 220).

Nevertheless, no unique solution exists in n-players games so the rationality of playing the maximin strategy is endangered. The multiplicity of solutions in n-players games is for vN explained by the fact that a solution is interpreted as a "standard of behavior", that numerous "stable social structures" exist in social reality and that "many differing conventions can endure" (Kuhn and Tucker, 1958, p. 103; in Mirowski, 1992, pp. 140-41). VN/M declare " [w]e shall in most cases observe a multiplicity of solutions. Considering what we have said about interpreting solutions as stable standards of behavior this has a simple and not unreasonable meaning, namely that given the same physical background different established orders of society or accepted standards of behavior can be built." (von Neumann and Morgenstern, 1953[1944], p. 42) Thus, this is not a surprising result for them.

In n-players games the solutions of games become a stable set of imputations, with imputations understood is terms of a vector of individual payoffs (ibid, p. 34). Therefore while vN/M see the existence of a solution in games as an important matter for the theory of games, the unicity of the solution on the contrary is not. This explains why vN in particular rejects the Nash equilibrium as a solution concept; it is too much attached to the problem of unicity. The multiplicity of solutions is on the contrary an advantage for vN. As already seen, an imputation is interpreted as a standard of behavior or an "order of society"; the existence of a set of imputations thus means that from the same initial conditions different accepted standards of behavior can arise (ibid, pp. 41-42). As Shubik (1992, p. 155) emphasizes, relying on personal conversations with von Neumann, "He felt it was premature to consider solutions which picked out a single point and he did not like non-cooperative equilibrium solutions … he did not particularly like the Nash solution and that a cooperative theory made more social sense."

2.1.4. Quid of Morgenstern's view of strategic rationality?

As Péréa (2013, p. 2) emphasizes, Morgenstern was one of the first to ask for the integration of individuals' reasoning processes and in particular of individuals' beliefs regarding other's choice and reasoning in economic models. He claimed that to understand individuals' choices and therefore to set the conditions under which an equilibrium could be reached required investigating the players knowledge about others' choice and the way they form their expectations of the others' choice (Giocoli, 2003, p. 163). And according to him, establishing a theory of equilibrium in economics cannot be based on an assumption of perfect foresight as this would suppose models peopled by "not ordinary men, but rather, at least to one another, exactly equal demi-gods" (von Neumann and Morgenstern, 1953[1944], p. 173).

Besides, the perfect foresight hypothesis as explained in the Holmes Moriarty paradox leads to an infinite regression. The economic sphere is determined by "an endless chain of reciprocally conjectural reactions and counter-reactions. This chain can never be broken by an act of

knowledge but always only through an arbitrary act" (Morgenstern 1935 (11-12), Box 5 OMP) As Morgenstern puts:

> "as Sherlock Holmes, chased by his enemy Moriarty, leaves from London to Dover with a train which stops at an intermediate station, he gets off the train instead of going on to Dover. He saw Moriarty at the train station and, considering him very intelligent, expected that Moriarty would take the faster train, to await him in Dover. This anticipation of Holmes turned out to be right. But what if Moriarty had been even more intelligent and had considered Holmes' capacities even greater, and therefore had predicted Holmes' action? Then he would obviously only have gone to the intermediate station. Again Holmes would have calculated that, and therefore would have chosen Dover. Thus, Moriarty would have acted differently. And from so much thinking, there would have come no action, or the less intelligent one would have handed himself to the other at Victoria Station because all the fleeing would have been unnecessary. Examples of that kind could be taken from everywhere. Chess, strategy, etc., but there one needs to have special knowledge, which simply makes the examples more difficult" (1928, pp. 98–99).

Therefore, Morgenstern (1935, p. 75) claims that perfect foresight and economic equilibrium are "irreconcilable with one another."

The mixed strategy account developed in the TGEB is a way to bypass this indeterminacy problem, i.e. the chain of infinite regression of expectations that perfect foresight entails. In economics when agents have to form expectations on the way things are going to turn out, Morgenstern declares that some expectations must regard economic variables in the sense of dead variables (and determined by "nature") and some others regard live variables, i.e. the others' behavior and decision (see Schotter, 1992, p. 97; and see Morgenstern, 1928, 1935).[5] The types of probabilities, i.e. objective probabilities, integrated in the TGEB are about nature, i.e. dead variable and do not refer to other's actions or decisions.

The static character of the theory of game developed in the TGEB is contradictory to Morgenstern's earlier works (Rellstab, 1992, p. 78); recall that vN's vision of rationality and strategic rationality is atemporal. As Morgenstern (in Schotter, 1976, pp. 180-81) declares "it is clear that a theory of economic equilibrium which 'explains' only a static situation, which is given as unalterable and which, because of this basis assumption, is completely unable to say anything about the economy when a variation occurs, is utterly unimportant from a scientific point of view. It would hardly deserve the names of theory and science."

---

[5] This is linked to the maxims of behaviors. As explained by Giocoli (2003, p. 235), people must decide to follow a maxim of behavior or not: "There are two types of maxims: the unrestricted maxims, which can be followed regardless of the actions of the other agents, and the restricted maxims, which are followed on the basis of whether the others do the same or not… In the case of unrestricted maxims the choice poses no problems, because the agent's evaluation is not disturbed by the behavior of other individuals. In general, however, the decision depends upon the agent's forecasting ability which, in the case of restricted maxims, must also embrace the other agents' behavior."

2.1.5. And what after the publication of the TGEB?

The post TGEB era has been focused on the mathematical properties of the minimax theorem especially in 2-players zero-sum games (Giocoli, 2003, p. 217). Many critics have raised on the minimax theorem as a solution concept for n-players zero-sum games. The TGEB has been attacked by many scholars with respect to the problem of multiplicity of solutions; just to mention some of them, by Simon, Savage or Harsanyi (see Léonard, 2010, pp. 260-61). The latter argued that if the TGEB did not generate great interest among economists after its publication it is because of the lack of determinate solution (e.g. see Harsanyi, 1976); while Simon declares "I have further difficulty in pinning down the operational meaning of the term "solution". It is clear that only one imputation can follow from a single play. If a single imputation were observed, it could only be concluded that the solution which held for the players of the game was one of all those solutions of which //the particular observed imputation was a member. There might be an infinity of such solutions. If the society were 'stable' over a period of time then successive observations of imputations might successively narrow the set of possible solutions. Since an assumption of stability would be necessary to determine the solution, the stability of the solution could not be empirically tested" (Simon to O. Morgenstern, August 20, 1945, OMDU, Container 32, File 90; in Léonard, 2010, pp. 160-61) On the other hand Marschak (1946, p. 104) criticizes the TGEB with respect to its static analysis of strategic interactions asserting, like Morgenstern, "All is not well with static economics."

The expected utility theory on the contrary has been considered to be of great interest for economics by some economists and statisticians like Harsanyi, Friedman and Savage (Léonard, 2010, p. 264). It became a central element of game theory and later of epistemic game theory as the payoffs in games are standardly described, and it is still the case nowadays, in terms of vN/M utilities.

The work of vN/M after its publication therefore remained confined to a military use instead of immediately inspiring the formal revolution in the social sciences expected by VN/M (Léonard, 2010, p. 264; Giocoli, 2003). Game theory remained in the field of operations research during the WWII and the post WWII era so that the scope of application of game theory was reduced to military concerns only (Léonard, 2010, p. 265) – which also reinforced the predominance of 2PZSG (2-players zero-sum games). Nevertheless, although game theory was intended, in this context of operations research and of the RAND **Cor**poration, to solve tactical and military problems during the WWII and the Cold War, it seemed that game theory was revealed to be disappointing and of little help for practical military problems (Léonard, 2010, p. 298).

One important dimension, however, that arose in the context of the RAND corporation is the role of experiments in game theory and it appears that the concepts of "strategic interaction", "threats", or "credibility" became of particular importance though they were initially dismissed from the analysis by vN/M, being considered as being psychological or morals aspects that do not matter for a quantitative view of games and strategies. We will see in the chapter 2 that these new experimental considerations in game theory and some of the determinants of strategic rationality such as the threats or credibility of some player's strategic moves is of great

importance for Schelling. It contributes to the introduction of players' reasoning in games and of some determinants that belong to psychology.

Consequently, an experimental methodology developed at RAND and many experiments of gaming have been conducted from an interdisciplinary point of view (see Léonard, 2010). Experiments on game theory began in 1949, informally at first, with Flood who considered that "game theory was mathematically rigorous and of great value, but of questionable validity insofar as it had not been shown to stand up to experimental test." (Léonard, 2010, pp. 319-320). He declares that even in 2PZSG the TGEB does not give the player a clear-cut recommendation of what to play grounded on randomization (Léonard, 2010, p. 320).

From that perspective, there was a great dissatisfaction with the solution concept of vN/M, and then of Nash (which will be discussed in the next section). Thus game theory was accused of being unable to embrace the complexities of human decision making in strategic context: "the basic information provided to the mathematician was inadequate." (Léonard, 2010, p. 330) Experiments undertaken on game theory at RAND under the supervision of Kennedy after Flood were based on this statement. A widespread opinion at RAND was therefore that the value of game theory was more its framework of analysis than its solution understood as a numerical value. This vision of the value of game theory for social science will be characteristic of Schelling's contribution as will be shown in the next chapter. The interest in game theory was about its qualitative content, i.e. about the "reciprocation", "opposing intentions", or the "credibility of threats" (Léonard, 2010, p. 315). This is what produces the experimental turn in game theory at RAND in particular with the experiments headed by Flood and Kennedy.

Apart from the Cowles commission and Chicago, the TGEB was largely ignored in economic departments but was on the contrary very popular among mathematicians in mathematical departments (e.g. in Princeton and Michigan) and in military inspired organizations (e.g. in Rand Corporation and the Office for Naval Research (ONR)). Consequently, in Giocoli's (2003, p. 348) words "the mathematicians set the research guidelines in the field following their own priorities and standards of evaluation: rigor, elegance and generality became the key requirements that any new contribution should satisfy, while the possible applications to concrete socio-economic problems received little if any attention." This situation has pervaded game theory since its creation and as we will see in this chapter is still true today.

The lack of interest can be explained by the fact that most of the economists was not trained in mathematics; but also as Shubik stressed, that the TGEB was founded upon perfect information; a dimension that reveals to be quite inadequate for economic contexts. This would imply that cooperation occurs among perfectly informed players. However, cooperative games with perfect information applied to oligopoly situations would entail that the firms collude in order to gain the highest possible payoff (Shubik, 1952, p.146). However, in reality, interactions occur in the context of incomplete information, and economic competition results in non-cooperative games in which the players only know their own payoff (Shubik, 1952, p.149). This is why Shublik is more inclined to adhere to Nash's approach.

2.2. Nash's program

Nash attempts to solve the problems of the minimax strategy as a solution concept and in particular the problem of its uniqueness for any type of games. He indeed declares contrary to vN/M

> "We proceed by investigating the question: what would be a rational prediction of the behavior to be expected of rational playing of the game in question? By using the principles that a rational prediction should be unique, that the players should be able to deduce and make use of it, and that such knowledge on the part of each player of what to expect the others to do should not lead him to act out of conformity with the prediction, one is led to the concept of a solution defined before [i.e the Nash equilibirum]." (Nash, 1950, 23)

From that perspective, Nash proved that there exists a Nash equilibrium for any finite games being a zero sum game or not. It means that he proposed a solution which operates for any type of interactive context, i.e. any type of game contrary to von Neumann's minimax theorem.

According to Brandenburger (2010, p. 2) Nash's program relies on three premises: (i) there is only one "correct" way to analyze a game, (ii) players can analyze the game in such a way, and (iii) each player plays her best reply i.e. chooses her best available strategy against the others' best strategy choice. Point (ii) means that a player can analyze the game like a theorist, i.e. as an outside observer. This view contradicts the analysis of games that will be proposed by the two authors studied in the part 2 of the thesis: Schelling and Bacharach. This also contradicts the position adopted in the last chapter in the model of games that will be proposed. In brief, assuming that the point of view of the players and of the theorists differ is one way to escape the indeterminacy problem and therefore to explain and justify players' capacity of coordination. Regarding Nash the points (i), (ii) and (iii) together imply that there is only one way to play which is the Nash equilibrium (Brandenburger, 2010, p. 3).

One major achievement of Nash that is worth noticing here, apart from the definition of the Nash equilibrium as a solution concept, is to his reduction of cooperative games to their non-cooperative parts (Brandenburger, 2010; Giocoli, 2003). He was the first to propose a definition of cooperative and non-cooperative games that is still valued today. A cooperative game entails that the players "are supposed to be able to discuss the situation and agree on a rational joint plan of action, an agreement that should be assumed to be enforceable" (Nash, 1953, p. 128) while on cooperative games "it is impossible for the players to communicate or collaborate in any way" (ibid, p. 129). The reduction of cooperative to non-cooperative game is explained by Nash as follows: "one makes the players' steps of negotiation in the cooperative game become moves in the non-cooperative model. Of course, one cannot represent all the possible bargaining devices as moves in the non-cooperative game. The negotiation process must be formalized and restricted, but in such a way that each participant is still able to utilize all the essential strengths of his position (Nash, 1953, p. 129).

Hence Nash provides an axiomatic characterization of bargaining situations and conditions the solution of the game as a search for an equilibrium point. The consequence of that is the removal of the negotiation process from the analysis. Nash declares "we abstract from the situation to form a mathematical model in term of which to develop the theory." (Nash, 1950, p. 156) The axiomatization of Nash thus results in the outcome of the bargaining game depending solely on the initial demands of the bargainers which depend on the maximization of their utility function. As in vN/M, the analysis of the bargaining game is therefore static; the negotiation process is removed from the game analysis. Nash defined "a set of reasonable conditions" that a solution should fulfill and axiomatize these conditions so that according to Giocoli (2003, pp. 302-03) his approach corroborates the Hilbertian one.

More specifically, according to Nash the solution of a bargaining game is understood in terms of rational expectations, i.e. self-fulfilling expectations, regarding each player's expected payoffs Nash, 1950, p. 158). Nash did not however provide an explanation of how and why the players should conform to this agreement (Giocoli, 2003, p. 303).

Nash equilibrium more specifically is formally defined as a profile of strategy: $\hat{s} = (\widehat{s_1}, \widehat{s_2}, \dots, \widehat{s_n})$ such that for every $s_i \in S_i, u_i(\hat{s}) \geq u_i(\widehat{s_1}, \dots, \widehat{s_{i-1}}, \widehat{s_i}, \widehat{s_{i+1}}, \dots, \widehat{s_n})$, with u_i the utility function of player i and S_i the set of her strategies (see Giocoli, 2003, 296).

It means that each player plays her best reply against the other players' best reply.[6] A best reply reasoning means that a player i plays the strategy, say c_i, leading to the highest payoff given the strategy c_j played by the other player j ($j \neq i$). A Nash equilibrium is therefore a profile of strategies for which each player has played her best reply to each other best reply. No player can improve her outcome by deviating from the strategy she chooses (e.g. see Nash and Shapley, 1950, p. 106). A Nash equilibrium in mixed strategies corresponds to a specific distribution of probabilities, i.e. a specific probability is assigned to each pure strategy and no player has an incentive to adopt another strategy than the mixed strategy. Nash equilibrium in mixed strategies can be interpreted as a Nash equilibrium in pure strategy but in uncertainty, i.e. when the players have to form beliefs regarding the other player's choice.

Nash stated that each game possesses a Nash equilibrium either in pure or in mixed strategy. To prove the existence of the equilibrium point Nash uses the fixed theorem of Brouwer (Giocoli, 2003, p. 304)

While in 1950a Nash proves the existence of an equilibrium he does not provide an explanation of its interpretation and its link with strategic rationality (Giocoli, 2003, p. 302) This approach differs from vN/M's one which explains strategic rationality as a means to free herself from other's mental states and behavior, i.e. from any intersubjective consideration (Giocoli, 2003, p. 283). Nash on the contrary ties together the players' choice of strategy: the choice of a player is a function of the choice of the others (Giocoli, 2003, p. 283). Recall that the solution of the

[6] Such interpretation in terms of best reply reasoning is the most common one but Nash actually presents the Nash equilibrium in terms of countering (see Giocoli, 2003, p. 301 for more details)

bargaining game relies on the players' self-fulfilling expectations; they have to correctly anticipate each other's choice.

Considering that each player plays her best reply strategy to each other' s best reply strategy this can lead to an infinite regression that can only be broken if there exists a unique best reply strategy for each player, in other words, if only one Nash equilibrium exists. This is ensured by the fixed-point theorem at the heart of the proof of existence of the Nash equilibrium. The problem remains that the fixed-point argument supposed that the rationality of the agent, which is playing a best reply reasoning, is already assumed and not explained:

> "the equilibrium is simply imposed on the model without any reference to the process by which it is reached, while rational behavior is treated as something already achieved by the players that only needs be given a formal characterization… players are simply taken to be rational, that is, consistent, in their behavior, while nothing is said on what motivates them or on how they arrive at playing their equilibrium strategies." (Giocoli, 2003, p. 315)

If the fixed-point argument justifies the existence of the equilibrium, it does not justify its occurrence. How and why the players play the Nash equilibrium remains unexplained. For that purpose we need "a theory explaining why the players either agree on that equilibrium or follow that convention, or why their expectations should focus precisely on something whose salience is what should be proved in the first place." (Giocoli, 2003, p. 316) Justifying that the players play the Nash equilibrium would require a theory that explains how from a 'disequilibrium situation', i.e. from initial beliefs free of any rationality considerations – beliefs in which every strategic choice is admissible, they would progressively refine their anticipations of each other's choice so as to end up with correct anticipation of each other's choice.

Recall that questioning of the learning process involved when agents in disequilibrium situations progressively learn to ultimately reach the equilibrium situation is of major concern for interwar neoclassical economics. For Giocoli this provides one explanation for the lack of interest manifested by the economists in Nash's two major publications in 1950 and 1951, setting out his program and his equilibrium notion as a solution concept (Giocoli, 2003, p. 297). Nash equilibrium corresponds to a consistency view of rationality that conveys a SOR view of economics that was still at that time inconsistent with the neoclassical SOF view of economics. One major concern for neoclassical economists was to justify how and why the agents could, through a learning process, reach the multi-agent equilibrium point, i.e. how they could progressively refine their expectations about each other's choice so as to end up with consistent expectations (Giocoli, 2003, p. 314).

Nash, though, initially proposed two interpretations of his notion of equilibrium which could have explained how and why the players play a Nash equilibrium, but they have disappeared. In his own words, he attempted "to show how equilibrium points and solutions could be connected with observable phenomena" (Nash, 1996, p. 32). The first explanation was "mass-action" the second was "rationalistic." "Mass-action" means that with the repetition of interactions the players can progressively learn the equilibrium play. In modern terms this corresponds to repeated games in which players progressively acquire information regarding the correct and rational way to play. Players are not perfectly rational and are not assumed to possess perfect

knowledge of the game and of the rules of the game (Giocoli, 2003, p. 310). Nash (1996, p33) declares that "the "mass-action" interpretation leads to the conclusion that the mixed strategies representing the average behavior in each of the populations form an equilibrium point." The "rationalistic" interpretation refers to rational expectations. Nash argues that "a rational prediction should be unique, that the players should be able to deduce and make use of it, and that such knowledge on the part of each player of what to expect the others to do should not lead him to act out of conformity with the prediction, one is led to the concept of a solution defined before." (ibidem) This view requires that the players have perfect knowledge of the structure of the game and of the other players' knowledge and rationality (which is supposedly symmetrical) which is why for Nash such an interpretation is an "idealizing" one (ibidem).

As explained in the introduction of the thesis, evolutionary game theory is not discussed in this thesis as it relies on a very specific vision of the players not as reasoning humans, but solely as adaptive entities which still maximize their expected utilities which is contradictory with the mere definition of non-rational players (see Sugden, 2001; and Giocoli, 2003). Recall that we are interested, in this thesis, in the influence that the players' reasoning processes and the integration of their mental variables such as beliefs or intentions, have for the analysis of coordination in non-cooperative game theory.

Nevertheless, and contrary to Nash's claim, Nash equilibrium is not necessarily unique in that it does not provide, with respect to Nash's own program, a satisfactory solution concept. In some games Nash equilibrium leads to sub-optimal solutions, in others there are multiple Nash equilibria and in others no Nash equilibrium exists in pure strategy. In the latter case, it thus requires a search for a Nash equilibrium in mixed strategies. However, if playing a Nash equilibrium in mixed strategy helps to bypass the problem of indeterminacy in some games, for instance when there exist two Nash equilibria and one Nash equilibrium Pareto dominates the other, when no one Pareto dominates the other there is no rational belief justifying one equilibrium rather than the other. Thus, playing a Nash equilibrium in mixed strategy can be irrelevant (Hargreaves Heap and Varoufakis, 2004[1995], p. 71). Nash equilibrium in mixed strategies can help to solve the indeterminacy problem only in eliminating unattractive Nash equilibria. As this is not always the case and as no rational prediction in case of multiplicity can justify that one equilibrium profile of strategy will be selected at the expense of the others, this solution concept is unsatisfactory. This explains why Luce and Raiffa declared in their 1957 handbook, that Nash equilibrium couldn't be the solution concept of a "general theory of non-cooperative games". Besides, as declares Bicchieri (1993, p. 65), "Nash equilibrium is just a consistent prediction of how the game will be played;" the implicit message is therefore that some other predictions, with respect to the way the game is or should be played, exist.

Nash equilibrium is nevertheless the most popular and important solution concept within game theory. However, the fact that some games do not have a solution with Nash equilibrium as the solution concept casts doubts on its supremacy within the field. Something more is required. As Hargreaves Heap and Varoufakis (2004 [1995], p. 78) put it "something must be supplied by extra data on the social context in which the game is played, or the history of the people who play it … or/and a different conception of rationality. However game theory strives for a solution that draws exclusively on the assumptions that agents are rational and operate under CKR [common knowledge rationality]."

Because the indeterminacy problem is not unconditionally resolved with Nash equilibrium in mixed strategy, the refinement program is of particular importance. The refinements proposed of Nash equilibrium within the program provide new 'sophistications' for strategy profiles to be an equilibrium (Hargreaves Heap and Varoufakis, 2004[1995], pp. 78-79). As we will see in the next section, these social or historical dimensions that Hargreaves Heap and Varoufakis appeal for are however not one of the types of propositions made within the refinement program to discriminate between the multiple Nash equilibria when they exist.

To conclude while the EUT presented in the TGEB still remains a pillar of game theory for both classical and epistemic game theory, the program of vN/M did not have a lot of success compared to Nash's program. One the one hand, Nash equilibrium became the central solution concept of game theory and in particular of classical game theory. Recall that classical game theory basically refers to the account of game theory supposing the exclusive emphasis on equilibrium analysis and the hypotheses of common knowledge and joint rationality (Raiffa, 1992, p. 174). We will see that the refinement program and then the epistemic program induce the proposition of new solutions concepts, so that different solution concepts progressively became coexistent with Nash equilibrium. On the other hand, the non-cooperative dimension of Nash's program was considered as being much more appropriate for economic theory contrary than vN/M's cooperative games.

There are however common points between vN/M's and Nash's respective contributions. They both attempt to define strategic rationality and to provide a normative account of rational behavior in games. Besides, according to Giocoli (2003, p. 327) "[f]or both, strategic rationality was the output and not the input (that is, an assumption) of the analysis." They both provide a static analysis of game deprived of any learning process, of any reasoning and adjusting process (ibidem).

3. The refinement program

Raiffa's reminiscences of the troubles in which he found himself in the 1950s provide a good explanation of the challenges that game theory faced at the birth of the refinement program.

The author was investigating many configurations of 2-players games when he progressively discovered that there were many "anomalies" as he puts it. There were many problems related to the multiplicity of equilibria. Nevertheless as he noted "I thought we could get consensus about what we might mean by a game having a "solution" for the case in which there was a unique equilibrium pair that jointly dominated all other criteria pairs." (Raiffa, 1992, p. 171) But then he wondered "Solution in what sense?" (Raiffa, 1992, p. 171) He gave the following example to clarify his worries concerning the notion of solution in game theory

	C1	C2
R1	(0,-1000)	**(10,8)**
R2	**(12,10)**	(0,0)

This game presents two equilibria (R2,C1) and (R1,C2). (R2,C1) Pareto dominates (R1,C2), it should therefore be a good candidate for being the solution of the game; but Raiffa declares "I wasn't happy" with such a solution. To explain the problem he develops the following reasoning "Player 2 could think to herself "I'm in deep trouble if I take c1 and he takes r1. But he's no dope; he might think I'd be afraid to take c1 and therefore he might be inclined to take r1. But now that I can give some rationalization for him to take r1, I certainly should stay away from c1." The argument becomes cyclic. The stability of the equilibrium (r2,c1) is destroyed once one gives some probability of a deviation from that equilibrium.

"Wise advice, jointly given potentially to both players, would suggest r2 for 1 and c1 for 2. But is c1 a wise advice to player 2 in isolation?" (Raiffa, 1992, p. 171) He began "to worry about the question "solution in what sense?"" (ibidem) so that he finally tackled both the descriptive and predictive power of game theory: he became concerned with "the efficacy of the equilibria theory as either a descriptive or a predictive theory." (ibidem) He goes further declaring that "equilibria theory" may lead to a stalemate (Raiffa, 1992, p. 173). For him "There were too many inadequacies in a theory based on just equilibria analysis. As far as I was concerned the theorizing I did lacked descriptive or prescriptive applicability" (Raiffa, 1992, p. 173).

More generally, the main aim of this program was to refine the requirement to attain a unique Nash equilibrium in games. The refinement program offers different possible ways in which "the definition of an equilibrium … can be sharpened by invoking additional criteria derived from decision theory." (Govindan and Wilson, 2008, p. 1) Govindan and Wilson identify two technical criteria for sharpening Nash equilibrium : (i) invoking sequential rationality which basically entails invoking dynamic games, and (ii) modifying games to allow perturbed games. Among the large amount of refinement propositions, three types of theories can be identified according to Bacharach (2006): (i) the 'respecification theories', (ii) the 'bounded rationality theories' and (iii) the 'revisionary theories'. The respecification theories attempted to translate a pure coordination game into a game ensuring that the solution corresponds to the Pareto optimal equilibrium. The 'bounded rationality theories' relaxed the rationality assumption – even the perfect rationality. Evolutionary Game Theory belongs to this category of theories. And finally, the 'revisionary theories' imply a revision of the notion of rationality; they use different concepts of rationality.

The refinement program attempted to provide answers to two main questions: (i) why playing a Nash equilibrium, which means to what extent rationality in games would entail playing a Nash equilibrium, and (ii) which Nash equilibrium should the players play. While the refinement program fails to answer the second question, evolutionary game theory answers both questions.

That is why some scholars claim that the rise of evolutionary game theory is explained by the failure of the refinement program (Giocoli, 2003, p. 322; see also Sugden, 2001). In that respect, it is interesting to note that Hillas and Kohlberg (2002, Section 8.2) make a distinction between what they call a selection program and the refinement program. The latter attempts to provide sufficient conditions for selecting the right equilibrium not necessary conditions (see Brandenburger 2010, p. 3). The former on the contrary entails the identification of necessary conditions. In a comparative way Myerson (1941, pp. 241-42) distinguishes the true refinements on the one hand, which entail the identification of a more precise characterization of a rational behavior in games, from the false refinements that on the other hand provide solely a selection criteria.

The refinement program could have therefore been an important theoretical project ensuring the explanation of how and why an equilibrium could be reached and thereby explain how and why players can coordinate. As will be shown in this section, this has not been the case.

3.1 On the introduction of dynamics in game theory

A dynamic game is a game in which players act repeatedly, and are therefore able to infer the others' preferences, strategies, information, etc. There can be perfect or imperfect information and new solution concepts are proposed in each case. But before turning to the solution concepts proposed in dynamic games per se, let us explore the concept of elimination of dominant and strictly dominant strategies. The refinements necessitating the elimination of dominant and strictly dominant strategies do not necessarily involve dynamic games but they are of particular importance for dynamic games as the order of deletion can involve substantial differences in the ultimate solution. In the successive elimination of strictly dominated strategies or iterated dominance there is indeed path dependence. The selection of the different Nash equilibria can be affected by the order of deletion.

A strategy is dominant if it leads to a higher payoff whatever choice the other player makes: strictly dominant if it leads to a strictly higher payoff and weakly dominant if it leads to higher or equal payoff and more specifically a higher payoff for at least one of the other player's choices. The elimination of strictly or weakly dominated strategies can lead to a solution for some games. However if a Nash equilibrium is stable when induced by the elimination of a strictly dominant strategy this is not the case for weakly dominant strategy; in the latter case the Nash equilibrium is unstable (Hargreaves Heap and Varoufakis, 2004[1995], p. 51). In the successive elimination of dominated strategies, CKR is a necessary but not sufficient condition to eventually lead to a Nash equilibrium (Hargreaves Heap and Varoufakis, 2004[1995], pp. 55-56). This is illustrated by the concept of rationalizable strategies defined by Bernheim (1984) and Pearce (1984). Rationalizable strategies correspond to strategies that survive the successive elimination of strictly dominated strategies. In order for rationalizable strategies to be operative players must have common knowledge, i.e. to know that the other eliminated those strategies, to know that the other knows that they eliminated those strategies, and so on…. (ibidem). The Nash equilibrium profile of strategy is rationalizable however the converse is not true. If there is no unique rationalisable

strategy per player in this case, the multiple rationalizable strategies may not lead to a Nash equilibrium (Hargreaves Heap and Varoufakis, 2004[1995], p. 56). As the concept of rationalizable strategies is not so restrictive, in some contexts, it can leave the games' solution undefined. In some games there are indeed several plausible beliefs rationalizing more than one strategy as a best reply strategy to another best reply strategy.

Now, when the game is repeated as suggested earlier, information can be perfect or imperfect. When there is perfect information all the data of the game (e.g. the set of players and their strategies) is available and players can observe every action at each node, i.e. they know all the history of decisions during the game. It means that players do not act simultaneously so that they can observe the decision made by the others. In the case of perfect information each decision initiates a subgame. A subgame exists when the information set of the players is a singleton (Hargreaves Heap and Varoufakis, 2004 [1995], p. 93). This is the case if she knows exactly at which node she is, i.e., what was the strategy played by the player before her turn. Accordingly, an information set is a singleton if it contains one and only one node of the tree, the player knows at which node of the tree she is and therefore knows the previous moves in the game. On the contrary the information set is not a singleton when the player does not know at which node she is.

When games are of perfect information they are decomposed into subgames which entails that strategies that are dominated are progressively deleted. Players can unambiguously choose a priori an optimal strategy (when it exists) for the whole game. This optimal strategy can be determined thanks to backward induction, i.e. from the final node to the initial node. Backward induction runs as follow: the player that has to choose first imagines what the player who has to choose second will choose with respect to her own choice and then chooses accordingly; she reasons backward from the end of the game to the beginning.

The first equilibrium proposed with this informational structure linking the Nash equilibrium concept and backward induction is the subgame perfect Nash equilibrium (SPNE) proposed by Selten (1965). In particular a SPNE entails that a Nash equilibrium exists for each subgame.

Together with Harsanyi's introduction of imperfect information in game theory the work of Selten and the introduction of dynamics in game theory both contributed to an upsurge of interest for game theory in economics in the 1970s.

The SPNE entails the iterative deletion of dominated strategies when the players have a strictly dominant strategy (Hargreaves Heap and Varoufakis, 2004[1995], pp. 91-92). SPNE generally requires backward induction and "Nash's assumption" (Hargreaves Heap and Varoufakis, 2004[1995], p. 93), i.e. that each player plays her best reply strategy at each node of the game. This is what is called Nash backward induction. The difference between backward induction and Nash backward induction relies on the need for common knowledge rationality (CKR). While backward induction does not entail CKR because a strictly dominant strategy exists, Nash backward induction does because there is no strictly dominant strategy (Hargreaves Heap and Varoufakis, 2004[1995], p. 93).

The SPNE implies that for each subgame each player eliminates those of her strategies that are not best replies to the other players' best reply strategy at each subgame. In other words,

strategies are in SPNE in a game in extensive form if they are equivalent to a Nash equilibrium in each of the subgames of the game. It is a strategy profile for which each player has played her best reply according to the strategies chosen by the other players for each possible history whether or not they occur (Hargreaves Heap and Varoufakis, 2004[1995], p. 95). This last statement is controversial as CKR prevents off equilibrium paths, i.e. players cannot admit that the other players play strategies that are not optimal, i.e. that are not best reply, so that every possible history of the game is not conceivable. The difficulty in Nash backward induction results from the fact that the players never question the occurrence of out of equilibrium paths. They follow the profile of strategy dictated by the SPNE (Hargreaves Heap and Varoufakis, 2004[1995], p. 112) No rationality principle can tell the players how to reach the equilibrium path as they cannot find themselves out of this equilibrium path in first instance. This is the same problem as in classical game theory: when a solution concept is applied as a norm, rationality is assumed to equate to equilibrium play but nothing explains how and why the players should play this equilibrium profile of strategies.

One other possibility is forward induction and several propositions for this principle exist. Kohlberg and Mertens (1986) require that the set of refined equilibria must be an optimal response to any equilibrium; in other words, each strategy that is not optimal to any equilibrium of the set must be deleted. The set of refined equilibria must contain a subset that survives deletion of strategies that are not optimal responses at any equilibrium in the set.

Van Damme (1989, 1991) postulates that if a player rejects an option even if the option a priori seems for the other player better than some other equilibria in the subsequent subgame, it means that he anticipates better outcomes and that one equilibrium among the set of possible equilibria is better in the resulting subgame (Govindan and Wilson, 2008, p. 2).

The idea developed in forward induction is that the player commits themselves to some strategies, making her opponent drawing inferences on how the game is going to be played (Hargreaves Heap and Varoufakis, 2004[1995], p. 108); "forward induction … permits players to send meaningful messages … concerning their future intentions." (Hargreaves Heap and Varoufakis, 2004[1995], p. 109) Forward induction provides a means to signal her intention of play for future nodes. This could be related to a form of bluff or of signaling; a way to send messages through the strategies that we choose as players. For Hargreaves Heap and Varoufakis (2004 [1995], p. 108) this is very interesting from an analytical point of view as it shows that CKR can be dissociated from backward induction contrary to what is generally supposed. It is indeed generally admitted that CKR is equivalent to backward induction in dynamic games (see Hargreaves Heap and Varoufakis, 2004 [1995], p. 111).

On the contrary when information is imperfect players act simultaneously so that they cannot observe others' choice, there is no subgame and each player has private information (e.g. regarding her preferences, or expected payoffs, etc.). When information is imperfect the player may not know which node she at; she may not know the path that leads her to the decision she has to make and in consequence may not be able to anticipate the choice of the other players. Players therefore form beliefs about the others' choice; there is a probability distribution for the set of possible contingencies that the players can face, i.e. for each possible path. There exists a

conditional probability for each history of the game, i.e. for each node to which the player has arrived. Beliefs are updated according to Bayes rules.

In such dynamic games with imperfect information the refinement of Nash equilibrium entails that each strategy that is not optimal for any possible contingency, i.e. for any possible node of the game, must be deleted. The strategies that remain must be optimal. It means that a strategy that remains at a certain node must be optimal for all nodes in the remainder of the game whatever the future paths that will be followed. This is what is called sequential rationality (Govindan and Wilson, 2008, p. 4). Within this sequential rationality account there exist three types of refinements: the perfect-Bayes equilibrium, the sequential equilibrium, and the lexicographic equilibrium (Govindan and Wilson, 2008, p. 4).

The weakest possible refinement in dynamic games with imperfect information is the perfect-Bayes equilibrium (Fudenberg and Tirole, 1991) which states that (i) each player's strategy is optimal according to her belief on the others' choice, and (ii) Bayes' rule is satisfied. A stronger refinement is the sequential equilibrium (Kreps and Wilson, 1982) which requires that the players' beliefs must be consistent with the structure of the game.[7] Then, the sequential equilibrium proposed by Kreps and Wilson (1982a) is an extension of the subgame perfect equilibrium to cases of imperfect information in which there is uncertainty regarding the other players past move or types. But the "basic idea" is the same (Hargreaves Heap and Varoufakis, 2004[1995], p. 98). Both require backward induction and that strategies played by the players are best reply strategies. The only difference lies in the nature of the information set. When the information set is not a singleton, the players' strategy must be a best reply to the others' strategy for any information set. Accordingly the sequential equilibrium involves a strategy profile and a system of beliefs which stipulates for each information set and for each player the belief held by the player regarding the others' choice before her turn, i.e. the belief regarding the node at which she is.

Another refinement is related to the quasi-perfect equilibrium (van Damme, 1984), which requires that at each node the players would not want to deviate from the strategy they intend to play.

There exist other refinements that add restrictive criteria, in particular regarding the set of beliefs that the players can hold with respect to unexpected contingencies, i.e. unexpected paths. For instance, we can refer to Cho and Kreps (1987) who give the example of the probability assigned by a player to the possibility that the other player is of a type that deviates without any gain from deviation. It means that deviation must necessarily be a credible signal of the players' type. And obviously each player's strategy is optimal given her belief regarding others' strategy.

A last refinement is the lexicographic equilibrium (Blume, Brandenburger and Dekel, 1991). In this case each player forms different possible theories on the other's type and choice that are classified according to the credibility – or perhaps the probability – she attributes to each of these theories. In other words, each theory is ranked from the most plausible to the least plausible. The

[7] More specifically, from a formal point of view, consistency entails that the set of players' beliefs is "the limit of the conditional probabilities induced by players' strategies in some perturbed game." (Govindan and Wilson, 2008, p. 6)

player assumes in first instance that the theory that is ranked highest regarding the other is true and she plays her optimal strategy with respect to this theory. If the other's choice contradicts her theory, she assumes that the second theory prevails and so on.

3.2 On the building of perturbed games

The proponents of the perturbed games approach consider that the same results that in dynamic games could have been derived from static games. To reinforce such position they mention the fact that a sequential equilibrium, for instance, can select inadmissible strategies and that the set of possible equilibria is not invariant under the addition or deletion of redundant strategies, so that dynamic games do not provide rigorous solution concepts (Govindan and Wilson, 2008, p. 8)

The basic idea of perturbed games is that the players can deviate from their initial optimal strategies; they can 'tremble', for exogenous reasons that are not initially specified in the description of the games. A game played with 'trembles' is a game in which there is a positive (even if very small) probability that the players do not choose their optimal strategy and therefore there is a positive probability that every strategy can be played. Players can act differently than anticipated, i.e. than rationality would have led them to choose. Trembles are considered as mistakes or lapses of rationality or deviations, etc (Hargreaves Heap and Varoufakis, 2004 [1995], p. 81). Such trembles can be explained by distinguishing two phases: (i) a reasoning phase where the players form intentions to play a particular strategy and (ii) an implementation phase during which the tremble can intervene (ibid, p. 83). This implies that the optimal choice for each player in these perturbed games must take into account these possibilities of deviation. Perturbations can intervene at the level of strategies or payoffs (Govindan and Wilson, 2008, pp. 8-9). The refined equilibria in this account are therefore the equilibria that are slightly perturbed by all the possible deviations and perturbations.

When games are perturbed in this a way, the criteria involved to identify refined equilibria are perfection or stability like the perfect and proper equilibria in which the equilibria are slightly perturbed by deviations of players' strategy. The truly perfect equilibria entail a set of equilibria that are very little perturbed under any deviations of strategy. A more restrictive refinement requires stability. The set of equilibria must be stable under multilateral deviations of strategy, under the non-optimality of players' choice, under variation of payoffs, and under any deletion or addition of strategies that are redundant (Govindan and Wilson, 2008, p.5)

The trembling hand perfect equilibrium proposed by Selten (1975) is a Nash equilibrium that survives in a perturbed version of a game in which the probability of trembles tends to zero (see Hargreaves Heap and Varoufakis, 2004[1995], p. 82).

The trembling hand perfect equilibrium is a refinement of Nash equilibrium that helps to rule out Nash equilibria that are supported by weakly dominant strategies. However such refinement does not rule out "implausible" Nash equilibria in some cases (Hargreaves Heap and Varoufakis,

2004[1995], p. 84). More generally, it does not rule out all of the multiple Nash equilibria, in particular when there are no weakly dominant strategies. In this case, a problem of elimination remains (Hargreaves Heap and Varoufakis, 2004 [1995], p. 85)

The perfect equilibrium (Selten, 1975) illustrates for Govindan and Wilson (2008, p. 9) the "basic method" which necessitates two steps: (i) there exists a perfect equilibrium for each probability ε according to which every players' strategies has some positive probability to be selected but any strategy that is not an optimal response has a probability to be played that is inferior or at best equal to ε. – i.e. every strategy has a positive probability to be played even if it is suboptimal; and (ii) a perfect equilibrium is obtained by "the limit of a convergent subsequence of ε-perfect equilibria." (Govindan and Wilson, 2008, p. 9) A perfect equilibrium is equivalent to a sequential equilibrium in the dynamic version of the same game (Govindan and Wilson, 2008, p. 9), for Hargreaves Heap and Varoufakis (2004 [1995], p. 83) it is "the least restrictive concept involving trembles" used to battle the indeterminacy problem as it does not necessitate "to assume anything specific about the nature of the trembles."

The proper equilibrium (Myerson, 1978) involves the same basic method as the perfect equilibrium but it is a stronger refinement as it supposes that the lower the payoff of a strategy the less it is likely that its corresponding strategy be chosen. A proper equilibrium is related to the sequential equilibrium concept as "[a] proper equilibrium induces a sequential equilibrium in every one of the equivalent descriptions of the dynamic game." (Govindan and Wilson, 2008, pp. 9-10)

Subgame perfect Nash equilibrium and sequential equilibrium help to cut down certain Nash equilibria but not others. The difficulty related to sequential equilibria is linked for Hargreaves Heap and Varoufakis (2004 [1995], p. 107) to the fact that the origin of the players' beliefs regarding the other's choice is undefined. To determine in an extensive game which of the two possible Nash equilibria will occur when none of the SPNE and sequential equilibria break the indeterminacy problem, there is the need to determine which of the Nash equilibria is more reasonable or more likely to be the outcome of the game. This calls for investigating out-of equilibrium beliefs. This perspective is offered by perturbed games. For some of the Nash equilibrium there is a higher probability that the players tremble. We have to determine why and what could be the determinant of such probabilities of trembling, otherwise there is a lapse of rationality. When considering these out of equilibrium beliefs there is however the need to impose some restrictions (Hargreaves Heap and Varoufakis, 2004[1995], p. 107). The strict dominance of a strategy is one of such possible restrictions; it influences the beliefs that the player has on which strategy is more likely to be played. It is indeed less likely that when a player trembles she plays a strictly dominated strategy if another one is dominant. If trembles are considered as a kind of "experiment", for instance, there is no interest in experimenting with a strictly dominated strategy (Hargreaves Heap and Varoufakis, 2004[1995], p. 107). Myerson (1978) asserts that the likelihood of a tremble should be determined by the cost of such a tremble, which implies that the probability of a tremble is low when the cost is high and the higher the cost the less the probability is important. If the cost of trembling from one Nash equilibrium is higher than another such Nash equilibrium it is less likely to be the solution. This is the suggestion of Myerson to discriminate between Nash equilibria. Accordingly, Hargreaves

Heap and Varoufakis (2004[1995], p. 84) state that the origins of the 'trembles' should be explained or investigated. This is for them "a potential weakness with all refinements based on "un-theorised' trembles: they need a plausible theory of trembles to go with them and one that players share." (Hargreaves Heap and Varoufakis, 2004 [1995], p. 85)

The perfect and proper equilibria thus depends on the kind of perturbations supposed, so that according to the restriction existing on ε they can be "essential" or "truly-perfect" which states that the games can be perturbed by less than ε, i.e. by any σ lower than ε (Govindan and Wilson, 2008, p. 10). For instance, we can mention the work of Kohlberg and Mertens (1986) on stability and hyperstability which each time requires stronger criteria of existence. This will not be discussed here as it is a mainly technical detail. Hyperstability however tends to induce a larger set of equilibria than stability as hyperstability requires robustness against more types of perturbations (Govindan and Wilson, 2008, p. 11).

3.3 Other propositions

The payoff dominance principle has been proposed together with the risk dominance principle by Harsanyi and Selten (1988) within the well-known refinement program. It provides a principle of equilibrium selection that prevails for specific games, i.e. "common interest games" (Aumann and Sorin, 1989), i.e. those games in which one Nash equilibrium, among the multiple Nash equilibria, is Pareto dominant, and in which both players are in a symmetrical position: they possess the same set of strategies and receive the same payoffs in each outcome.

Coordination games such as the Stag-Hunt Game and the matching games (Bacharach, 1993; Bacharach and Bernasconi, 1997) belong to this class of games (see Colman and Bacharach, 1997).

Stag Hunt Game

	A	**B**
A	5,5	0,4
B	4,0	2,2

Hi-Lo Game

	H	**L**
H	2,2	0,0
L	0,0	1,1

Assurance Game

	C	**D**
C	4,4	1,2
D	2,1	3,3

These games all possess two Nash equilibria but in one of them Pareto dominates the other, in the sense that it gives to both players a higher payoff. In the Stag Hunt Game (A, A) Pareto dominates (B,B), in the Hi-Lo Game (H,H) Pareto dominates (L,L) and in the Assurance Game (C,C) Pareto Dominates (D,D). According to Harsanyi and Selten (1988)the payoff dominance

principle was supposed to provide a principle of equilibrium selection ruling out the Pareto dominated Nash equilibrium.

The principle entails that if one equilibrium offers the players a higher payoff they should play their part in this equilibrium: they should choose the corresponding profile of strategy. This principle therefore means that the players should choose strategy A in the stag hunt, H in the Hi-Lo and C in the assurance game. Harsanyi and Selten (1988) even considered that this principle should be a standard principle of rationality and therefore that it should be of common knowledge in games; in the same way as it is normally supposed that rationality is of common knowledge or at least of common belief in order for players to have expectations regarding the others players' choice of strategy (see Colman, 2004, p. 295). The Pareto optimal Nash equilibrium would then be a mutual best reply.

It is indeed tempting to say that if one of the Nash equilibrium payoff dominates it should be rational for the players to play the strategy that will lead them to such an outcome, since each player will be better off. The intuitive appeal of the payoff dominance principle is moreover well recognized (see Gauthier 1975; Lewis 1969; Farrell, 1988; Crawford and Haller, 1990; Bacharach, 1993; Sugden, 1995; Colman and Bacharach, 1997; Colman, 1997, 2004). However such an equilibrium cannot be justified rationally from a standard perspective. That is why according to Colman (2004, p. 295) Harsanyi and Selten (1988) introduced "the payoff-dominance principle as an axiom." A player indeed cannot play the strategy leading to the payoff dominant Nash equilibrium unless she knows or at least has strong reason to believe that the other players will do the same, i.e. that everybody will play their part in the Pareto optimal equilibrium profile of strategy. But since the game is symmetrical the other players are in the same situation, saying to themselves that they will play the strategy leading to the payoff dominant Nash equilibrium only if the others do the same, etc. This leads to infinite regressions and no rational basis is given to such equilibrium play, unless, as emphasized by Colman, the payoff dominance principle is *artificially made* a rational principle, i.e. an axiom. Even assuming common knowledge of the structure of the game and of instrumental rationality cannot provide a basis to rationalize the payoff dominance principle.

Besides, Harsanyi and Selten (1988) defined another competing principle: the risk dominance principle, which can prevent the players from playing the Pareto optimal profile of strategy. This is particularly true for the Assurance Game. A risk dominant Nash equilibrium is an equilibrium corresponding to a strategy choice less risky than the other. When players are uncertain of the strategy choice of the other players they choose to play the strategy leading to the risk dominant equilibrium. Let's take the example of the Assurance game: if player 1 (the line player) chooses to play strategy A she can get either 5 or 0; thus if she is not sure that player 2 (the row player) will play A she risks having 0 instead of 5 whereas if she plays B she can get either 4 or 2. When being uncertain of the choice of the other player, playing B is thus a less risky choice. Subsequently, if each player makes the same reasoning, the equilibrium is (B,B). Nonetheless, there exist empirical data showing that players are generally inclined to play the payoff dominant strategy profile (see e.g., Colman, 2004, p. 296; Cooper, DeJong, Forsythe, and Ross, 1990; Mehta, Starmer, and Sugden, 1994a,b)

As Colman and Bacharach (1997, p. 3) assume, a "plausible attempt" to rationalize the payoff dominance principle requires more than respecification theories; rather, we claim, it requires "revisionary theories". They list three lines of research: bounded rationality theories, such as the level-k theory (Stahl and Wilson, 1994; Bacharach and Stahl, 2000), team reasoning (Hurley, 1991; Sugden, 1993; Bacharach, 1999, 2006) and the third they propose in their 1997 paper: the Stackelberg heuristic. All these lines of research bypass the mere respecification of coordination games. The main attempts to rationalize the payoff dominance principle in standard game theory were however mainly mere respecification theories (see Colman and Bacharach, 1997, p. 3). As examples, Colman and Bacharach (ibidem) mention the attempts of Aumann and Sorin (1989) who introduce repetition and of Anderlini (1999) and Farrell (1988) who introduced a cheap talk stage with costless talks, before the game per se. Strangely while they mention that Bacharach (1993) used the concept of payoff dominance to "provide determinate solution" (Colman and Bacharach, 1997, p. 2) they do not assert that Bacharach's VFT could have been a way to rationalize the payoff dominance principle with a respecification theory.

Such a proposition within the refinement program is mentioned by especially the scholars working on coordination problems but not so much in contributions on the refinement program that concentrate on technical and mathematical issues like Godivan and Wilson, for instance. But this concept of payoff dominance is particularly relevant for Bacharach's VFT which will be presented in more detail in chapter 3 of this thesis as a theoretical representation of Schelling's focal point (which will also be presented in detail in the next chapter) to justify players' capacity to coordinate.

3.4. What is the headway of the refinement program?

The many refinements of Nash equilibrium proposed within the so called refinement program entail "[t]he development of increasingly stronger refinements by imposing ad hoc criteria incrementally … Eventually, one wants to identify decision-theoretic criteria that suffice as axioms to characterize refinements." (Govindan and Wilson, 2008, p. 12) The emphasis was therefore on the identification of mathematical representations of axioms sufficient to support the mathematical requirements imposed by Nash (1950, 1951). The building of dynamic games and of perturbed games approaches the same objective differently. While perturbed games as proposed first by Selten (1975), and then Myerson (1978), and Kohlberg and Mertens (1986), attempted to satisfy the criteria of stability, they remained very dependent on the kind of perturbations supposed, which makes them very *ad-hoc*. "Perturbations are mathematical artifacts used to identify refinements with desirable properties, but they are not intrinsic to a fundamental theory of rational decision making in multi-person situations." (Govindan and Wilson, 2008, p. 12) That is why the second class of refinement attempts to refer to decision theoretic criteria such as admissibility, iterative elimination of dominated strategies, backward or forward induction, etc. But again "the ultimate aim is to characterize refinements axiomatically." (Govindan and Wilson, 2008, p. 12)

The refinements program did not contradict the view of a game theory that remained a mathematical theory concerned with the laws and rules of the mathematical community: "this program's mathematical models were often inspired by purely mathematical concerns rather than the economic phenomena they were intended to model." (De Bruin, 2009, p. 290)

The refinement program raises many methodological and philosophical difficulties in particular regarding the role of instrumental rationality in games. As Hargreaves Heap and Varoufakis (2004 [1995], p. 117) put it, taking the example of the subperfect Nash equilibrium (SPNE)

"The crux of the problem facing game theory is that it has to introduce the possibility of some lapse of rationality to explain what rationality demands. This is because what rationality demands is often determined in dynamic games by a consideration of what would happen if rational players actually end up in what turn out to be out-of-equilibrium decision nodes. But why should one assume that players behave rationally when they find themselves at an out-of-equilibrium decision node? Surely if the analysis of SPNE is correct, then rational players should not reach these out-of equilibrium nodes."

In order to justify out of equilibrium paths and the rational response that the players can make to these situations, perturbed games have been linked to dynamic games. But the difficulty remains, once trembles or perturbations have been accepted; how to justify why players do not a priori integrate into their reasoning the possibility of trembling. Trembling could become a specific strategy, players could intentionally and purposefully decide to tremble; "it may even lead rational players to decide to tremble." (ibidem) This would imply that the rules of the games change and that players acknowledge this before playing. Otherwise, trembles must remain purely random events. In other words, depending on the way a tremble is interpreted the methodological and philosophical consequences are very different.

4. From Harsanyi (1967-68)'s contribution and the introduction of player's hierarchy of beliefs to the birth of the epistemic program in game theory

4.1 Harsanyi's introduction of uncertainty in game theory

Contrary to many of the propositions of equilibria within the refinement program that revealed quite disappointing results and that did not meet the requirements for being integrated in standard game theory as they remained mathematically centered, Harsanyi's (Harsanyi, 1967-68) introduction of uncertainty into game theory met the requirement of economics (Giocoli, 2003). Situations of imperfect information, as already mentioned, are more usual in economics than situations of perfect information as originally supposed in game theory. Thanks to Harsanyi's work, game theory became progressively integrated into the economic tool box becoming one of its core tools.

In addition, Harsanyi's work opened the door to the epistemic program that was very promising in the 90s as it, supposedly, put the analysis of players' reasoning at the center and linked the result of a game to the knowledge attributed to the players with respect to the other players' information, belief or reasoning. This dimension should have been essential, from its early days, in the understanding of strategic rationality.

Harsanyi (1967-68) introduced uncertainty in game theory, but with the adoption of a specific formalism, i.e. with the encoding of the players' hierarchy of beliefs in the definition of their type. He found a way to transform games of imperfect information into games of perfect information using the definition of probabilistic beliefs regarding each other's type. The type of players in Harsanyi's (1967-68) original formulation specifically defined their utility function and beliefs hierarchy. The players' hierarchy of beliefs integrates what the players believe about the game payoff function, what they believe that the others believe about such payoff function and so on ad infinitum.[8] More specifically, player i's hierarchy of beliefs is of the form: i believes that p, i believes that j believes that p, i believes that j believes that i believes that p, and so on ad infinitum (e.g. see Péréa, 2013, p. 6). Harsanyi introduces the concept of "types" as an artefact to encode these hierarchies of beliefs.[9]

Harsanyi first introduced this idea of the player's infinite hierarchy of beliefs in 1962 in the context of bargaining games (Péréa, 2013, pp. 6-7). He described the infinite hierarchy of beliefs in bargaining contexts as follows:

> "In bargaining, and more generally in all non-trivial game situations, the behavior of a rational individual will depend on what he expects the other party will do. Party 1 will ask for the best terms he expects party 2 to accept. But party 1 will know that the terms party 2 will accept in turn depend on what terms party 2 expects party 1 to accept. Thus, party 1's behavior will depend on what may be called his second-order expectations, i.e., on party 1's expectations concerning party 2's expectations about party 1 's behavior. These again will depend on party 1's third-order expectations, i.e., on his expectations concerning party 2's second-order expectations, etc." (Harsanyi, 1962, p. 29)

The beliefs within these hierarchies of beliefs are however non-probabilistic beliefs. Harsanyi (1967-68) thus extends the use of this formalism for incomplete information games and for probabilistic beliefs (see Péréa, 2013).

Harsanyi originally assumed uncertainty only with respect to players' payoff. He indeed argued that any kind of uncertainty can be reduced to uncertainty regarding the players' payoff (may it be uncertainty regarding the outcome of the game, or a players' utility function, or a players' set of strategies). In this case players may not know some of the other players' payoffs. Thus if player 1

---

[8] The formal justification for the encoding of players' hierarchy of beliefs according to the definition of the players' type is later provided by among others Armbruster and Böge (1979), Böge and Eisele (1979), or Mertens and Zamir (1985), (see Brandenburger, 2010). Mertens and Zamir (1985) in particular proved that the type of the player allow encoding of the player's infinite hierarchy of beliefs.

[9] For an epistemogical reflection on the concept of "types" in Harsanyi's work, see Hargreaves Heap and Varoufakis (2004[1995])

is uncertain about player 2's payoff, i.e. about the type of player she is, player 1 will form a probability regarding the event that player 2 be of one type or another. In order to determine her best choice player 1 will maximize her expected utility, i.e. will adopt her best choice by taking into account (i) the different possible payoffs associated with each type of the other player and (ii) her beliefs (probability assessment) regarding the possibility that the other is of one type or another.

According to Harsanyi the fact that situations of imperfect information give rise to an infinite hierarchy of beliefs for each player explains why the study of these situations had made so little progress until his contribution (Harsanyi, 1967). With his definition of type, he manages to encode in a compact way the infinite hierarchies of beliefs that considerably simplifies the analysis, as there is no need to write down the whole hierarchy of belief for each player. More specifically, the type of a player i is defined by a vector $c_i = (a_i; b_i)$ with a_i the utility function and b_i the infinite hierarchy of beliefs (c_i is also called the information vector or attribute vector of player i [Péréa, 2013, p. 8; Myerson, 2004, p. 1819]). b_i thus describes the player i's infinite hierarchy of beliefs, i.e. her first order beliefs: the beliefs about her opponents' utility function, her second order beliefs: the beliefs about the opponents' beliefs about the other's utility function (i.e. the opponent's first order beliefs), her third order beliefs: the beliefs about the opponents' second order beliefs, and so on ad infinitum. Each player's type ultimately defines what the player knows at the beginning of the game, i.e. her private information (Myerson, 2004).

A game is therefore described by a set of types for each player, a probability distribution about the players' type which corresponds to the belief that the other players hold about the other players' true type within the set of the potential types, and a set of perfect information games for each combination of players' types. In this context, Nash equilibrium becomes a Bayesian Nash equilibrium which is a profile of strategy that corresponds to each player's best reply according to her probabilistic beliefs about the other players' type. A Bayesian Nash equilibrium is therefore a Nash equilibrium of the corresponding Bayesian game.

In 1973, Harsanyi offered a new interpretation of Bayesian Nash equilibrium in mixed strategy games from the one corresponding to a randomization behavior as in 1967-68. He indeed recognized that it is very unlikely that players randomize in a strategic context but more likely form subjective beliefs about the other players' choices and in turn act according to these beliefs. Then in 1975, Harsanyi extended his framework to uncertainty over the other players' choices. He thus finally considered the issue of the origin of the players' beliefs, i.e. of where the players initial beliefs, prior to the game – what is called their prior beliefs – come from (see Harsanyi and Selten, 1988).

> "[T]he resolution of this problem will be a very essential part of our whole theory of solutions for n-person noncooperative games, because in general, the tracing procedure to be described will select different equilibrium points s* as solutions for any given game Γ, depending on the n-tuple p of prior probability distributions used as a starting point (Harsanyi, 1975, p.63)

Assuming subjective beliefs is a step into the epistemic program of game theory. The reduction of incomplete information games into complete information games depends, according to

Harsanyi (2004), on a specific distribution of beliefs that will later be called the common prior assumption. Such new thinking regarding the nature and more specifically the consistency of players' beliefs is indeed typical of the epistemic program in game theory. A Bayesian Nash equilibrium indeed depends on the hypothesis of common aligned beliefs (CAB) which means that there is only one probability distribution for each player's belief: there can be only one set of beliefs (Hargreaves Heap and Varoufakis, 2004[1995], p. 89). CAB entails that the players have common knowledge of the probabilities regarding each other's type and it is necessary to the existence of a Bayesian Nash equilibrium (Hargreaves Heap and Varoufakis, 2004[1995], p. 279). The common knowledge rationality hypothesis (CKR) ensures that players are able to form expectations with respect to other players' choices, which are described by the set of hierarchies of beliefs.

However Harsanyi finds a formal artifact to avoid such epistemic intrusion into his model of games of incomplete information as he declares that assuming CAB avoids resorting to hierarchies of subjective beliefs.

> "if we use the Bayesian approach, then the sequential-expectations model for any given I-game [Incomplete information] G will have to be analyzed in terms of infinite sequences of higher and higher-order subjective probability distributions, i.e., subjective probability distributions over subjective probability distributions. In contrast, under our own model, it will be possible to analyze any given I-game G in terms of one unique probability distribution R* (as well as certain conditional probability distributions derived from R*)." (Harsanyi, 2004, p. 1807)

As will be explained in the next section, assuming common prior entails that the subjective beliefs held by the players are derived, according to Bayes's rule, from a common probability distribution (Myerson, 2004, p. 1821). Before the game, a lottery defines which of the different possible types a player is (Harsanyi, 1995). As summed up by Myerson (2004, p. 1808) in Harsanyi's methodology

> "Instead of assuming that certain important attributes of the players are determined by some hypothetical random events at the beginning of the game, we may rather assume that the players themselves are drawn at random from certain hypothetical populations containing a mixture of individuals of different "types," characterized by different attribute vectors (i.e., by different combinations of the relevant attributes)."

Then the player will try as an informed observer to "estimate the probabilities of the lottery" and in particular "each player will act on the assumption that the other player will estimate these probabilities … much in the same way as he does." (Harsanyi, 1995, p. 298) Even if the players do not have perfect knowledge of the outcome of the lottery, i.e. they do not know exactly of what type are each other players, they know the "basic probability distribution" (Myerson, 2004, p. 1804). Each player will derive her hierarchy of beliefs (i.e. her probability distribution) from the same "basic probability distribution" of the lottery and it is in that manner that the players' probability distributions are consistent each other (Myerson, 2004, p. 1804).

With such formalism, players' beliefs regarding the others' type are exogenous; i.e. they are independent of the specific structure of the games under scrutiny. "Harsanyi 'tied down' his

Bayesian-Nash equilibrium on a solid foundation of beliefs that were imported from outside the game … They reflect the distribution of characters or types in the broader population." (Hargreaves Heap and Varoufakis, 2004[1995], pp. 279-280)

The difficulty that arises from Harsanyi's methodology as will be explained below is that the priors that the players hold are built in such a way that they are derived (by Bayes updating) from priors that are at the equilibrium: they correspond to the solution of the game (Lecouteux, 2018b). Harsanyi and Selten (1988, p.140) state that simply assuming that the players maximize their expected utility according to their subjective beliefs regarding the others' choice is insufficient: "Unfortunately this simple theory will not work because this best reply combination … will generally not be an equilibrium point of the game, and therefore it cannot be the outcome chosen by a rational outcome-selection theory". This explains why they require CAB, i.e. subjective beliefs supposed to be rational and already at the equilibrium.

There is accordingly no explanation of the rationality of the strategy chosen as it is merely supposed as in classical game theory that the equilibrium profile of strategy is rational. There is no explanation of why the players should choose the equilibrium strategy. The rationality of a specific strategy remains defined by the fact that it corresponds to an equilibrium profile (Lecouteux, 2018b, p. 8). On the contrary, the epistemic program in game theory attempts to solve this question and determine why a player should choose her equilibrium strategy.

Many of the contributions that flourished immediately after the rise of what is called the type-based approach developed by Harsanyi concentrate on the analysis of games of incomplete information (e.g. Zamir 1971; Mertens 1971; Mertens and Zamir 1971; Harsanyi and Selten 1972; Ponssard and Zamir 1973; Kohlberg and Zamir 1974; Ponssard 1975a,b; Kohlberg 1975a,b; Sorin 1979, etc.). They however strictly adopt Harsanyi's original formulation in which the uncertainty is about the payoff function of the game and not the choices of the players. They do not assess the Bayesian foundations of games of incomplete information like the epistemic game theory. The objective of the latter is to challenge the study of solution concepts in game theory and to assess whether the players' knowledge regarding the other players' reasoning impacts on such solution concepts (Péréa, 2012, p.2).

4.2 The birth of the epistemic program in game theory

The pioneering work of Harsanyi (1967-68) introduces incomplete information into the realm of game theory and as a consequence puts an emphasis on the importance of the players' hierarchy of beliefs in strategic contexts. When incomplete information is added into games, the primitive of the models of games become the players' hierarchies of beliefs (Bonanno and Nehring, 1999) as they are uncertain about some of the characteristics of the structure of the game. Such an emphasis on the players' hierarchy of beliefs came very late in game theory development despite Morgenstern' (1935) concerns in the very early stage of modern game theory. Harsanyi (1967-68)'s and later Aumann (1987)'s works gave birth to epistemic game theory that progressively became the core of standard non-cooperative game theory, so that finally the players' reasoning,

in particular about others' eventual choices and beliefs, became the center of the analysis of strategic contexts (Péréa, 2013, p. 2).[10] The epistemic program really originated from a moment of uncertainty that had been introduced regarding the strategies and choices that players can make; this was not the case in Harsanyi's contribution (see Brandenburger, 2010).

Besides, while classical game theorists were interested in the properties of a solution concept (e.g. its existence, uniqueness or multiplicity, stability, and optimality), without investigating how rational players could reach such a solution, epistemic game theorists instead study strategic interactions from the perspective of the decision-makers (Aumann and Drèze, 2008), in order to determine the conditions on which the knowledge and beliefs of the players depend so that their choices correspond to a specific solution concept. Taking the players' hierarchy of beliefs as the primitive of the analysis of games and investigating the conditions of knowledge and beliefs (that the player holds regarding others' choices, strategies, payoffs or beliefs) compatible with an equilibrium play are at the core of the epistemic program in game theory. Thus it becomes possible to provide epistemic conditions for solution concepts (Brandenburger, 2010, p. 6).

Many works that flourished in the 1980s attempted to identify the epistemic requirements compatible with equilibrium plays (see for instance Pearce, 1984; Bernheim, 1984, 1986; Aumann, 1987; Brandenburger and Dekel, 1987b or Tan and Werlang, 1988). This change of perspective from the study of solution concepts to the study of the epistemic conditions underlying individual rational choice is, in Aumann (2010, p. 29)'s words, a "revolution".

The main challenge faced by non-cooperative game theory that epistemic game theory had to explain was why playing an equilibrium profile of strategies, and in particular why playing the Nash equilibrum profile of strategy, is rational. Recall (as described in section 2 of this chapter) that Nash equilibrium entails a consistency view of rationality which supposes a SOR view of economics in which no dynamic process exists, where no equilibrating forces towards the equilibrium exist. Thus, no justification of the equilibrium play has been provided by classical game theory (e.g see Giocoli, 2003, p. 314; Lecouteux, 2018b, p. 1421). However, making the epistemic states of the players an input of the model can provide this justification, by deriving solution concepts according to the specific knowledge that the players hold and their beliefs regarding others' choices, strategies or payoffs. In standard game theory of complete information, the structure of the game, i.e. players' utils over each combination of the pure strategies, and players' rationality are common knowledge. In this context the incontestably dominant solution concept is Nash equilibrium. Nash equilibrium entails that each player plays her best reply to the best reply of each other player. Each player is rational if she plays her part in this equilibrium knowing that each other player does the same. In other words, it recommends the players to play a specific strategy when knowing that the same strategy is implemented by each other rational player. The solution concept is applied as a norm and should be followed by any rational decision-maker. Every player knows her rational choice and others' rational choice before making

[10] See Péréa (2012, 2013) and Brandenburger (2010) for an historical perspective on the transition from classical to epistemic game theory. See Brandenburger (1992, 2007), Geanakopolos (1992), Dekel and Gul (1997), Battigali and Bonnano (1999), Dekel and Siniscalchi (2015) for surveys on the epistemic program on game theory, its main assumptions and results, from a formal perspective.

a choice. However, according to Aumann (1987) the answer to the question "why the players should use their equilibrium strategies" remains unspecified. Aumann and Drèze (2008) emphasize that, while classical game theory suggests strategy profiles to the players, a theory of rational choice in games should instead explain why it is individually rational to play a specific strategy.

Epistemic game theory provides various concepts to limit the set of admissible beliefs in games. As Péréa (2013, p. 11) claims "all concepts in epistemic game theory can be viewed as a collection of conditions on belief hierarchies." For Brandenburger (2010, p. 6) epistemic game theory in particular arises from the search for formal proofs establishing that common belief in rationality entails the elimination of strongly dominated strategies. Besides, as Péréa (2013, p. 11) stresses, most of the solutions concepts (which will be presented in this section) rely on common beliefs of rationality and propose some variations of the constraint imposed on the hierarchy of players' beliefs.

4.3 The standard hypotheses of epistemic game theory

As Aumann and Drèze (2008, p.73) suggest, the epistemic approach is to take the standpoint of a player (P1), and consider a strategic game G defined by a set of player, their strategy sets and payoff functions. It is assumed that P1 has a belief about the strategies played by the other players, and also about their beliefs (about what they think P1 plays, but also what they think P1 thinks they play, and so on). The "context" of the game is described by this belief hierarchy and they argue that "the fundamental object of game theory should be the game situation" (ibidem. 82), i.e. the belief hierarchy of P1 in G and the strategic game G, rather than its underlying game G, contrary to standard game theory. The knowledge of this underlying game G is for them generally insufficient to determine what a rational player should do. "The essential element" of an epistemic game is therefore, for Aumann and Drèze (2008, p. 73), the "context" of the game. Such context and "the restrictions" placed on the players' hierarchies of beliefs according to the rationality principle, are for them "implied by the formal definition of the game itself, whatever the specific situation may be." (Aumann and Drèze, 2008, p. 73)

In the epistemic program a game is therefore described in addition to the standard characteristics, i.e. the set of players, their set of strategies and payoffs, by their types in what is called the type-based approach developed by Harsanyi, or by the players' information partitions or by the set of possible worlds, in the 'Aumann-Kripke' approach. The types of the players or equally the information partitions or sets of states of the world are of particular importance in the description of the game as the outcome of the game is determined by it (Brandenburger, 2010, pp. 6-7).

In the state-based approach developed by Kripke (1963) and Aumann (1976), a state of the world describes everything that is relevant in the game, what the players know and believe about themselves and about the other players (including the other players' beliefs), and about the possible states of affairs they "deem possible" (Kaneko, 2013, p. 2; Lecouteux, 2018b). A state of

the world therefore describes the states of affairs that a player deems possible, the states of affairs that the other players deem possible, what each player deems possible about what the others deem possible, etc. However, the player does not know exactly what is the real state of the world (Péréa, 2013, p. 9).

More specifically in Aumann's words "[t]he term "state of the world" implies a definite specification of all parameters that may be the object of uncertainty on the part of any player of G." (Aumann, 1987, p. 6) A given state of the world provides a complete description of the world but the players have access to only a partial description of such world (Kaneko, 2013, p. 2). Within the set of states of affairs that the players deem possible at a given state of the world there is the true state of the world. Aumman (1987, p. 6) adds that a given state of the world ω provides a "specification of which action is chosen by each player of [the game G] at that state ω." In this way the state-based approach can be equivalent to the type-based approach (Péréa, 2013, p. 10). Although the Aumann –Kripke state-based approach is originally non probabilistic, by defining a probability distribution describing the states of affairs that the players deem possible, a hierarchy of beliefs can be derived for every player and such approach can be equivalent to the type-based approach (see Péréa, 2013, p. 10; Tsakas, 2012; Aumann 2010; Brandenburger and Dekel, 1993; Tan and Werlang, 1992).[11] That is why Gul (1998, p. 930) declares "an information model is simply a notational device for representing the n-tuple of infinite hierarchies of beliefs"

In such an approach a difficulty arises regarding the information needed to be of common knowledge among the players. In an epistemic model, as Bacharach (1994, p. 178) emphazises, "a condition p is c.k. [common knowledge] if p holds in all states of the model." It means that in each state of the world the players know that p, know that the other players know that p, the players know that the other know that they know that p, etc. ad infinitum. This is for Kaneko a "difficulty" as it implies that the players' information that is supposedly private, i.e. subjective must be commonly known, i.e. objective (Kaneko, 2013, p. 2).

As Aumann (1987, p. 9) declares

> "A question that often crops up when models of differential information are discussed is whether there can be uncertainty on the part of one player about the information partitions P' of other players. The answer is "no". While Player 1 may well be ignorant of what Player 2 knows – i.e., of the element of p2 that contains the "true" state ω of the world – 1 cannot be ignorant of the partition p2 itself. Indeed, p2 is part of the description of the model, and does not enter the description of any particular ω; it therefore cannot be the object of uncertainty, it must be common knowledge. This is not an assumption, but a "theorem", a tautology; it is implicit in the model itself."

In epistemic game theory, players are Bayes rational; for instance, they maximize their expected utility given their beliefs regarding the other players' strategies, payoffs and beliefs. It means that they play their best reply or their optimal choice given their information and given their beliefs

[11] The formal proof of the equivalence is given by Brandenburger and Dekel (1993).

regarding others' strategy choice (Hart, 2006, p. 205). But this hypothesis of rationality goes further by imposing restrictions on the players' beliefs and in particular by imposing the consistency of players' respective beliefs and the correctness of such beliefs. The hypothesis of rationality and the methodological consequences this hypothesis has on the players' beliefs and in particular the players' prior beliefs will be discussed extensively below in the section 5.1.

Consistency implies that if player i has a belief, say b_i about the player j, the player j must believe that the player i believes b_i. The beliefs player i has about a player j are among the set of beliefs j has about i (Bicchieri, 1993, p. 51; see also Tan and Werlang, 1986). "The knowledge of the two players must be consistent; one player cannot know something that the other knows to be false." (Gul, 1998, p. 930) Correctness states that "correct beliefs are the result of a rational procedure such as, for example, Bayesian updating from the so-called priors." (Bicchieri, 1993, p. 33)

This belief consistency requirement is, for Bicchieri (1993, p. 55), the epistemic rationality which ultimately implies common knowledge of players' beliefs (Bichhieri, 1993, p. 81). Mutual consistency requires that each player knows each other's beliefs. It is assumed that every player knows the beliefs of each other player (e.g. see Aumann, 1987; Bicchieri 1993, p. 81) and how everybody updates her beliefs (i.e. according to Bayes rule). This is established by the so called Aumann-Harsanyi doctrine (Aumann, 1976, 1987) which also implies that the priors are common knowledge. As Aumann (1987, p. 9) declares "[o]nce one accepts the Bayesian viewpoint, that each player has a prior on Q, it follows that there cannot be any uncertainty on the part of one player about other players' priors. Each player's prior must be common knowledge among all players."

The prior stage from which players' prior beliefs are deduced, specifies all states of nature that players consider as possible before playing the game, i.e. all of the possible combinations of individual choices. It is indeed supposed that there exists a prior stage, an interim stage and a posterior stage. At the posterior stage the equilibrium profile of strategies is revealed so that each player's beliefs at the equilibrium are commonly known. At the interim stage players have private information, a private signal is revealed to them (such as their type or their information partition) while such private information is not revealed yet at the prior stage (Aumann and Drèze, 2008, p. 80).

In all of these possible states of the world every player is known to be Bayes rational – or at least every player believes that every player is Bayes rational. And when strategies are included in the priors everybody knows the others' strategy profile. This is what Levi (1995, p. 175) calls the "Joint Action Space Prior Assumption." In this formal framework, in order for a player to have second order beliefs, for instance for player j's second order beliefs to be determined, player i must have beliefs on his own action. From a realistic point of view those requirements are heroic.

This state of the art is explained by the central hypothesis of common belief in rationality. Each player is Bayes rational and each of them know or at least believe that the other is Bayes rational, each believes that each believes that the other is Bayes rational and so on ad infinitum. According to Péréa (2013) the epistemic program in game theory began precisely when papers such as the ones of Milgrom (1981), Brandenburger and Dekel (1987a), Bernheim (1984), Pearce (1984), Monderer and Samet (1989), or Rubinstein (1989) investigated the methodological consequences

of imposing a common belief of rationality (hereafter CBR). CBR is "the central idea in epistemic game theory" (Péréa, 2013, p. 13). Each solution concept is derived from the CBR hypothesis providing more or less restrictions (ibidem).

CBR entails that there already exists a restriction regarding the beliefs that the players can hold at the prior stage. At this prior stage each player believes in the notion that "each believes that the other is Bayes rational" (Péréa, 2013, p.11). CBR and Bayes rationality therefore restrict the set of 'rational' prior beliefs that the players can hold (Lecouteux, 2018b, pp.19-21)

CBR was originally proposed by Friedell (1967, 1969) and Lewis (1969) who defined the concept of common belief and common knowledge, and Aumann (1976) who later offered a formal definition of it (see Péréa, 2013, pp. 11-12 for a presentation of each proposition).[12] Aumann more specifically defines common knowledge as follow:

> "When we say that an event is "common knowledge," we mean more than just that both 1 and 2 know it; we require also that 1 knows that 2 knows it, 2 knows that 1 knows it, 1 knows that 2 knows that 1 knows it, and so on. For example, if 1 and 2 are both present when the event happens and see each other there, then the event becomes common knowledge." (Aumann, 1976, p. 1236)

The first papers to provide a formal definition of CBR were however Böge and Eisele (1979) and Armbruster and Böge (1979) and later Bernheim (1984) and Pearce (1984) who stated and proved that CBR is equivalent to the solution concept of rationalizability (see Péréa, 2013, pp. 14-15).

We will again see the methodological and philosophical consequences of this in section 5.1.

4.4 The main solution concepts of epistemic game theory

Bernheim (1986) proposed to categorize the main hypotheses of epistemic game theory on which all the principle equilibrium concepts are based with four axioms. These four axioms are the following:

(i) "optimization": each player adopts the strategy that maximizes her expected payoff according to her belief regarding the other players' choices (Bernheim, 1986, p. 473). This is equivalent to Bayes rationality which states that each player will select her optimal strategy with respect to her beliefs (e.g. see Hart, 2006, p. 205)

[12] Friedell (1969, p. 31) more precisely defines the concept of "common opinion" and states the real conditions in which common opinion can rise. Like Lewis, and contrary to Aumann (1976) Friedell is "interested in real life situations" (Péréa, 2013, p. 12) Aumann only states that "at a given state ω there is common knowledge of an event E among two persons A and B, if at ω both A and B only deem possible states in E, if at ω both A and B only deem possible states at which A and B only deem possible states in E, and so on, ad infinitum." (Péréa, 2013, p. 13)

(ii) "consistency" which states that the players only assign positive probability for other players' choices that are rational (Bernheim, 1986, p. 474)
(iii) "independence" which means that the player's choice are independent from each other (Bernheim, 1986, p. 474)
(iv) "common priors" which mean that the players' beliefs are derived from the same probability distribution and which implies that the players' beliefs are mutually consistent (Bernheim, 1986, p. 474).

From the less demanding equilibrium concept to the most demanding equilibrium concept we have: 1) iterated dominance that requires the axioms (i) and (ii), 2) rationalizability that requires the axioms (i) to (iii), 3) the correlated equilibrium that requires the axioms (i), (ii) and (iv), and 4) the Nash equilibrium that requires axioms (i) to (iv).

According to these four equilibrium concepts, there is a specific structure of prior that is reiterated in the following table:

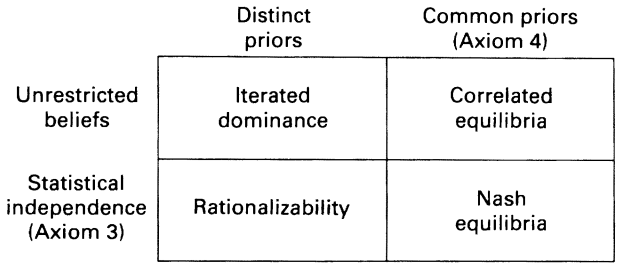


Fig. 1. Relationships between alternative solution concepts.

For Böge and Eisele (1979), CBR implies a "recursive elimination procedure" i.e. the recursive elimination of dominated strategies. It requires common knowledge of rationality as CKR implies the iterative deletion of dominated strategies.

Brandenburger and Dekel (1987) show that CBR implies correlated rationalizable choices and Tan and Werlang (1988) show that CBR plus independence of beliefs leads to uncorrelated rationalizable choices. Correlation means that players observe the same random event so that their beliefs are derived from the same probability distribution. Thus independence means that the players' observations are independent, they are not based on the same observation of the same random variables (ibidem).

Bernheim (1984) and Pearce (1984) also show that rationalizability is linked to this concept of subjective correlated equilibrium. More precisely, a rationalizable outcome belongs to a set of subjective correlated equilibria (see also Tan and Werlang, 1984, and Brandenburger and Dekel, 1985). However, rationalizability generally does not imply that the players believe in the correlation of their strategies

According to Aumann (1987, p. 2) if players have a subjective distribution of probability over the set of all the possible states of the worlds, if each player is Bayes rational in the sense that she maximizes her expected utility according to her beliefs and if there is common knowledge of

rationality, the equilibrium profile of strategies is a correlated equilibrium. This however requires the hypothesis of common priors which means that the players' beliefs are correlated; they are derived from the observation of the same random event (Aumann 1987, p. 3). In the absence of common priors Aumann (1987, p. 14) suggests that while the correlated equilibrium is no longer reachable, there is instead a "subjective correlated equilibrium distribution." However, for him, this does not provide a rigorous solution concept, as there exist very few restrictions on the eventual outcomes that can arise in this case. He argues that "the subjective correlated equilibrium is a relatively "weak" concept, giving little information." (Aumann, 1987, p. 15) Aumann goes further, emphasizing that, as Harsanyi (1967) states it, when players have distinct priors this is an "inconsistent case".

The underlying premise of the correlated equilibrium is the following "If two people have the same priors, and their posteriors for a given event A are common knowledge, then these posteriors must be equal. This is so even though they may base their posteriors on quite different information. In brief, people with the same priors cannot agree to disagree." (Aumann, 1976, p. 1236)

Recall that the Harsanyi doctrine (1967-68) states that when facing the same information rational players must draw the same inference, i.e. come to the same conclusion. This leaves open the possibility that players may come to different conclusions when having different initial information. However Aumann (1976) strengthens the Harsanyi doctrine by adding that rational players must consider as relevant the same information, so that a rational player must come with the same information before playing. They must consider exactly the same information as relevant for playing. The Harsanyi-Aumann doctrine therefore implies that 'players cannot agree to disagree'. The methodological consequence is that when facing the same information players must necessarily hold the same beliefs regarding what all the rational players will play. In other words players must hold the same prior beliefs. It appears that when rational players under CKR learn that they hold inconsistent beliefs they revise such beliefs in order to come to consistent beliefs.

According to Aumann's (1987) and then Brandenburger and Dekel's (1987) formalism a correlated equilibrium is sustained by an epistemic model in which for every state of the world the player's choices are optimal given their belief regarding the other players' choice. This asserts a stronger condition than the CBR as it requires that the criteria of optimality must be satisfied in each state of the world of the model (Péréa, 2013, p. 18). Although Aumann (1987) and Brandenburger and Dekel (1987) define this situation as simply CBR it is defined by others as a condition of "universal rationality" (Péréa, 2013, p. 18).

Finally, the difference between the correlated equilibrium and the Nash equilibrium is that in the former players receive a public signal, i.e. receive the same signal (or signals are correlated), while in the later those signals are stochastically independent (see Hart, 2006, p. 202).

Nash equilibrium implies consistent beliefs and correct beliefs regarding each other's choices as there is common knowledge of each other's beliefs (see Tan and Werlang, 1988; Brandenburger and Dekel, 1989; Aumann and Brandenburger, 1995; Polak, 1999; Perea, 2007; or Bach and Tsakas, 2012; related in Péréa, 2013). As Tan and Werlang (1988) show a NE is an equilibrium of

subjective beliefs rather than of strategies as is the case in classical game theory. Nash equilibrium requires consistent aligned beliefs (CAB), i.e. when each players' anticipation or beliefs regarding the other's choice is correct. No player will change either her belief or her strategy in reaching a Nash equilibrium. In order for players' beliefs to be consistently aligned there must exist for each player only one way to rationally play the game (Hargreaves Heap and Varoufakis, 2004 [1995], p. 60).

The definition of the Nash equilibrium is as follow: if the player i expects the player j to choose her strategy xj the player i has no incentive to choose another strategy than xi and the same prevails for player j; if the player j expects the player i to choose her strategy xi the player j has no incentive to choose another strategy than xj.

Nash equilibrium relies on the Harsanyi-Aumann doctrine and on the CAB hypothesis which is sustained by common priors (Hargreaves Heap and Varoufakis, 2004[1995], p. 78). A common knowledge assumption is always required for justifying the existence of a NE. This explains Gioccoli's (2003, p. 319) criticism claiming that common knowledge is required at the level of individual beliefs, which for him is "debatable" as individual beliefs are supposedly subjective. In support, he refers to Binmore (1987, pp. 209-12), who argues that a Nash equilibrium in an epistemic model relies on a fixed-point argument to state that, as in classical game theory, the explanation of the equilibrium is still static. This is linked to the consistency view of rationality which implies that the equilibrium is still postulated, as a fact, and not explained contrary to what was the declared objective of the epistemic program in game theory.

The defense of Nash equilibrium in mixed strategy, as interpreted by Aumann, i.e. as a play of pure strategy with subjective beliefs, and which necessitates CAB, begs the question of how players come to know the beliefs that the other players hold about the way they are going to play? This relies on the common prior assumption which ultimately entails that prior to the game the players make the same assessment about the way the game is going to be played. The belief that player i holds regarding the way the player j is going to play must be the same as the belief that the player i herself holds about the way she is to play. This explains why Sugden (1991, p. 78) argues "by pure deductive analysis, using no psychological premises whatever, we have come up with a conclusion about what rational players must believe about the properties of a psychological mechanism."

5. Adressing a methodological assessment of the epistemic program of game theory

5.1. On the prior assumptions and the nature of probabilities it implies: the methodological consequences on players' beliefs

Recall that at a prior stage, the players hold beliefs about all of the possible states of world that may occur and in which "all states of nature describe contingencies that at a prior stage were considered as possible resolutions of the uncertainty." The priors thus supposedly resolve all uncertainty. To meet this objective, very restrictive conditions are imposed on the prior beliefs that will be exposed in this section. This has furthermore, as will be explained below, important methodological and philosophical consequences regarding the nature of the probabilities defining the players' beliefs and their epistemological status.

The status of the players' prior beliefs raises many questioning and controversies that mainly focus on the assumption of common prior which is supposed to be a consequence of Harsanyi's methodology in order to introduce incomplete information in game theory.

Although the hypothesis of common priors is particularly dubious, this section will more generally assess the nature of priors and their meaning, being common or not, and the impact they have on the nature of the probabilities defining the players' beliefs.

The main restrictions placed on the players' prior beliefs are the following: they must be consistent with the concept of rationality and with the concept of equilibrium defined by the theorists. The problem is that rationality as defined by Bayes rationality can justify that players update their beliefs according to Bayes rules or that they maximize their expected utility according to their beliefs but it cannot define the prior beliefs they handle (Morris, 1995, p. 235). A belief can be said to be rational only if it concerns objective events and therefore accounts for long run frequencies (ibid, p. 238); but when it concerns subjective beliefs, i.e. subjective feelings, it is very dubious to postulate rational prior (even before Bayes updating when new information is revealed so that the beliefs can be revised until becoming consistent).

In order to provide a determined solution for games, only one profile of beliefs or priors is allowed: "player must have a unique (and correct) subjective prior distribution on the behavior of the other players." (Roth and Schoumaker, 1983, p. 1337) When beliefs become defined by strictly subjective probability the danger, for some game theorists (e.g. Harsanyi, 1967-68, 1982; Aumann, 1987, 1998), is that any kind of belief can be admissible, any kind of behaviors can be anticipated even completely irrational ones. A range of prior distributions is however according to Roth and Shoumaker (1983) consistent with the data of games so that the players can handle different priors regarding each other's strategy choice. This is also compatible with the rationality assumption: different prior distributions in which the players believe or know that the other players are rational can exist (ibidem). However, due to concerns regarding tractability the multiplicity of prior distribution has been largely discarded (see Lecouteux, 2018b). This confirms

the power of mathematical concerns and mathematical rigor that are still imposed on game theory which remains like classical game theory, dominated by the codes of the mathematical community.

This is also why common knowledge of rationality is of such importance for game theorists as it allow them to circumscribe the set of beliefs that can be plausible, and further that can be rationalized. And the priors must be of common knowledge (Aumann, 1998, p. 937)

The difficulty of agreeing with the common prior assumption (CPA) is even greater when it is stressed that at the prior stage players have already the same beliefs regarding the way the game is to be played.

Harsanyi (1967-68) states that players fed with the same information should hold the same beliefs, i.e. the same subjective probabilities, which for Aumann (1976, pp. 1237-38) means that the players have the same priors. The CPA is defined by Aumann (1987, p. 7) as follow "All the priors p1 are the same; that is, there is a probability measure p on ω such that pl = p2= ...= pn = p."

The underlying premise of the common prior assumption (CPA) is that "[i]f one sets forth all relevant information in sufficient detail, then in principle, there should be no room for differing probabilities." (Aumann, 1998, p. 932) For Aumann (1987 pp. 13-14) "there is no rational basis for people who have always been fed precisely the same information" to hold different priors. Besides, Harsanyi (1982a, p. 120) declares "all competent Bayesian statisticians will recognize that in some situations there is only *one* rational prior distribution… those involving random devices with suitable physical symmetries … *all* reasonable people will use the *same* … probability distribution." This means that the players must appraise the game in the same manner. Recall that in addition, it is supposed that the only relevant information is the definition of the game, the structure of the games. This reinforces the relevance of common priors. However, the hypothesis of rationality cannot justify the existence of the CPA. The CPA cannot be seen as a consequence of rationality: "there is … a vast conceptual distinction between the traditional … notion of 'rationality as consistency' – which … implies expected utility maximization with respect to some probability distribution – and the common prior assumption." (Morris, 1995, p. 235) Besides if there was a real prior stage in which the players' beliefs, i.e. their priors, would be equivalent, they would not want to revise these beliefs, yet it is supposed that they update their beliefs between the prior and the interim stage.

Hence for Dekel and Gul (1997, p. 115), the hypothesis of common priors seems purely artificial; they view the existence of a prior stage as of a purely "artificial nature". The consistency of players beliefs that is required with the CPA is a purely mathematical property so that "we do not know what it is that we would be accepting if we were to accept the common prior assumption" (Gul, 1998, p. 926; see also Bonanno and Nehring, 1999, p. 410). If such consistency of beliefs is assumed prior to the game, we do not see how the players could update their beliefs to hold subjective beliefs, ultimately at the equilibrium. This is, for Gul (1998, p. 924), "antithetical to the Savage-established foundations of statistics (i.e., the "Bayesian view"), since it amounts to asserting that at some moment in time everyone must have identical beliefs" so that everyone agrees on the future actions of everyone, i.e. everyone hold the same beliefs regarding each

other's action. Gul adds "Neither the theorem nor the analysis provides any new argument for why one would expect actions to be generated [by the CPA] nor why beliefs for the players (or an outside observer) should be as if behavior were generated in this manner." (Gul 1998, p. 927)

However, for Aumann (1976, 1985, 1987), the prior stage and the common prior are actually results rather than hypotheses of the epistemic model (see also Gilboa, 2011, p. 301). The reply of Aumann (1998, p. 929) to Gul's attacks is simply that the CPA "is essential for the derivation of correlated equilibrium". It means that the CPA is a result of the existence of a correlated equilibrium.

Harsanyi's view regarding the nature of players' beliefs and the CPA, is expressed as follow "The logical justification for replacing the original game containing *subjective* probability distributions with a probabilistic game model involving only *objective* probability distributions is the well-known fact that any Bayesian decision maker (or player) will always act exactly the same way, regardless of whether he interprets the numerical probabilities he assigns to various events as *objective* probabilities corresponding to long-run frequencies or as *subjective* probabilities expressing merely his own personal beliefs." (Harsanyi, 1982a, p. 123)

The subjective nature of players' beliefs imposed in epistemic game theory begs the question as to whether the restriction imposed on priors beliefs endangers their subjective nature. Harsanyi confirms that a purely subjective probability "permits the decision-maker to choose his subjective probabilities in any arbitrary way" while "[i]n contrast, the *necessitarian* version uniquely specifies the subjective probabilities which a rational decision-maker can use in a given situation." (Harsanyi, 1982a, p. 120).

There exist three possible interpretations of probabilities according to Morris (1995): (i) the "frequentist probabilities" which rely on "the observation of some repetition of the event" (Savage, 1954, p. 3), i.e. "long run frequencies of repeated events" (Morris, 1995, p. 231); (ii) the "personalistic" or "subjectivist Bayesian" probabilities which rely on the "individuals' willingness to bet" (Morris, 1995, p. 231), i.e. on "the confidence that a particular individual has in the truth of a particular proposition" (Savage, 1954, p. 3); and (iii) the "necessary", "logical" or "objective Bayesian" probabilities; in this case the probability "measures the extent to which one set of propositions, out of logical necessity and apart from human opinion, confirms the truth of another." (Morris, 1995, p. 231) The frequentists and the logical or necessitarian interpretations, in Harsanyi's terminology, are "independent of any individual," hence "they provide a possible origin for a common prior" (ibidem); they "exist logically prior to the individuals' decision problem, in which case we can think of them as exogenous to the decision problem." (ibid, p. 233) They are not "a function of individual choices" contrary to the personalistic view of probabilities (ibid, p. 231).

Laying the foundations of an epistemic game on probabilities that prove to be by nature exogenous to the decision problem faced by the players undermines the subjective nature of their beliefs. Indeed according to Morris (ibid, p. 233) assuming at the same time that the players' beliefs are subjective and that there exist common priors which means that such beliefs are exogenous probabilities independent of the way the players perceive their decision problem is contradictory. Player's beliefs are supposed to be endogenously determined in the game by the

understanding they have of the game in order for these beliefs to be subjective, i.e. to reveal the player's willingness to bet.

That is why Morris (ibid, p. 227) is led to ask "Why has the economic community been unwilling, in practice, to accept and actually use the idea of truly *personal* probabilities in much the same way that it did accept the idea of personal utility functions?" Aumann (1987, p.13) gives the following answer

> "Perhaps the most basic reason is that utilities directly express tastes, which are inherently personal. It would be silly to talk about "impersonal tastes", tastes that are "objective" or "unbiased." But it is not at all silly to talk about unbiased probability estimates, and even to strive to achieve them. On the contrary, people are often criticized for wishful thinking-for letting their preferences color their judgment. One cannot sensibly ask for expert advice on what one's tastes should be; but one may well ask for expert advice on probabilities."

Even if Morris raises the problem of the inconsistency of postulating subjective priors and assuming common priors, we could apply the same critique for the existence of the other form of priors. In any case, priors are supposed to be a consequence of Bayes rationality and common knowledge or at least common beliefs of rationality. As Bicchieri points out, the hypothesis of players' common knowledge of mutual beliefs entails that the probabilities defining the players' beliefs "derive from public relative frequencies" (Bichhieri, 1993, p. 81) which confirm that they cannot be subjective. They remain "an exogenous construct" (Heifetz, 2018, p. 3) as it is required that the players "know the structure of the state space, which is an abstract construct of the modeler, and not necessarily the form in which the individuals grasp the uncertain environment in which they act." (Heifetz, 2018, p. 3) The existence of priors is exogenous to the game as they preexist the game per se; this is therefore contradictory for Morris, as players' beliefs are about events endogenous to the game. This situation is mainly explained by the fact that the most important requirement in game theory is the existence of a solution. Ascertaining the existence of this solution, i.e. "the perceived necessity of deriving determinate solution concepts", leads to the introduction of "special assumptions" by Harsanyi (1967-68) that threaten the subjective view of probability and instead imply a necessitarian view (Kadane and Larkey, 1982, p. 115)

In fact, as presented early on by Harsanyi, priors are mere "mathematical artefacts", which cannot be given a positive interpretation for the priors: "the prior distribution is merely a mathematical artifact introduced into the model so that the *subjective* probability distributions used by the players in deciding their strategies can be replaced by *objective* (conditional) probability distributions in order to obtain a game model admitting of analysis by the usual analytic methods of game theory." (Harsanyi, 1982, pp. 122-23) Indeed, "the CPA is a mathematical property whose conceptual content is not clear." (Bonanno and Nehring, 1999, p. 410) Such criticism is in particular led by Lipman (1995), Dekel and Gul (1997) and Gul (1998).

In Harsanyi's view, priors are just a convenient formal way to avoid assuming subjective and 'free' beliefs. The reference to specific priors simply allows it to be stated that the posteriors are consistent with the equilibrium defined by the theorist. Priors are relevant only insofar as they concern the posteriors (Morris, 1995, p. 237)

> "The priors are artifacts of a notational device to represent the infinite hierarchies of beliefs … of the players, i.e., their "posteriors" at the true state of nature …The fallacy of viewing the "priors" … as beliefs at some hypothetical prior stage becomes clear, if we remember that the information model above is nothing more than an equivalent representation of the infinite hierarchies of belief … there are no data in either of these hierarchies regarding any "prior" stage. Nor are there any data in the infinite hierarchies regarding what the agents would have believed had their information been "less" or "more" than what it in fact is. Hence, no such data can be present in the equivalent representation of these hierarchies of beliefs." (Gul, 1998, pp. 925-926)

Aumann answers to the criticisms by arguing that the common priors are a formal consequence of the axioms of the models and nothing else:

> "The interpretation of a common prior derives naturally from the axioms: It represents the "correct" or "appropriate" probabilities – what people would or should believe – if there were no private information. Admittedly, the existence of such "neutral" probabilities is not obvious. But that is what the axioms yield; and in any case, the meaning or interpretation of this concept is clear." (Aumann, 1998, p. 936)

The leaders of the epistemic program in game theory contest the descriptive interpretation of player's priors, they contradict any possible positive interpretation of these priors. For instance Aumann and Brandenburger (1995, p. 1174) argue that "[a]n interactive belief system is not a prescriptive model; it does not suggest actions to the players. Rather, it is a formal framework – a language – for talking about actions, payoffs, and beliefs." Any other interpretation than the purely formal is vacuous. What matters for game theorists is the consistency of players' beliefs and it is not relevant from a positive perspective (Bicchieri, 1993, p. 86). If priors were indeed interpreted as real beliefs, as Morris (1995, p. 238) emphasizes, players' beliefs would possibly not converge as they are about endogenous events to the game; in this individual and subjective learning process there is no guarantee of convergence. The endogenous status of players' beliefs, if they were real beliefs, thus also contradicts the hypothesis of consistency of posteriors. Players' posteriors may not converge.

In Bacharach and Hurley's (1991, p. 26) words, the priors, interpreted in terms of real beliefs, are thus 'the central unknowns of the theory':

"What brings me to have the prior probabilities that I do for your deciding on one option and another is a question not answered (and rarely asked) by the Bayesian theory of games. The absence of an independent account of what is in the players' priors is a grave lacuna. There are many games for which, once the priors are given, the identities of the rational acts follow trivially, and then game theory itself is trivialized if it is merely assumed that the prior are such and such. To avoid this trivialization by 'Bayesianization' we must take the content of the priors in such cases to be the central unknowns of the theory, endogenous to it."

5.2. What kind of players peopled the epistemic games

If players have different priors it is argued it is because they are filled with different information, which means different personal history and experiences of life (see Aumann, 1998, 1987; Harsanyi, 1967-68). Recall that for Harsanyi, having different priors reveals an inconsistent case, i.e. a situation in which the players' beliefs are inconsistent and which threatens the existence of an equilibrium to the game. Negating such differences means that players, when facing a game, must be symmetrical in their analysis of the objective game as defined by the theorists; they must pay attention exactly to the same information as the other players and furthermore to exactly the same information as the theorist would do. They act as impartial and wise observers of the game as if they were theorists. This premise implies a negation of player's subjectivity and accordingly there is a misrepresentation of the supposed bottom up methodology of epistemic game theory. As argued by Aumann and Drèze (2008), the game is not formalized from the perspective of the players, their epistemic states and mode of reasoning, it is still formalized from an external perspective when a solution concept is applied normatively and imposed on the players. This raises the first philosophical contradiction we would like to emphasize in this thesis. Imposing priors means that players are free from any subjective content, even from their own identity. Players could be anything, computers, countries, etc. But how can such an empty entity in that case be the bearer of mental states like beliefs, and how can theorists in that case argue that they tend to analyze the game form the perspective of the players if they can void a player of her humanity. Further more, when nature reveals to the players their type, which entails the transition from the prior to the interim stage in which the players have finally subjective information, it begs the question of how nature can give to the player the knowledge of their own identity? It means that the players' identity, formalized by their type ""pops up miraculously" in one's mind" (Gilboa, 2011, p. 307)

> "When they come to play the game they know not only their identity but also their utilities and priors. However, before the game begins they are only empty shells a and b, neither of which knows its identity. That is, a complete description of their knowledge does not use the term "I"." (Gilboa, 2011, p. 306)

Aumann was inclined to defend the common prior assumption thanks to the "empty shell argument" as exposed by Gilboa (2011), but he dismissed such a statement in his final version of his 1987 paper. The empty shell argument is presented by Gilboa as follows:

> "players may have different beliefs (priors) due to different information they acquired during their lives. Theoretically, one may try to model all this learning as simple Bayes' update of a prior one has at birth… At the moment of birth (or conception, or even much earlier, depending on the reader's faith and social policy preferences), this intelligent entity – the empty shell – learns the genes it got, updates its prior and becomes a "regular" player with a utility function and beliefs that are now the posterior. However, the "empty shell" argument concludes, before learning the genes, there is no reason to distinguish between these empty shells. They are all identical, since any distinction among

them is assumed to be learned later on. In particular, they all have the same prior." (Gilboa, 2011, pp. 301-02)

In order to evacuate the "empty shell" statement when imposing the common prior assumption, as already mentioned earlier, Aumann (1985, 1987) argues that it is not a hypothesis of the model but a *result* of common knowledge of information partitions and priors and of rationality.

As a result of the axioms of the model, priors are at the equilibrium as they already encompass consistency of beliefs and CBR. As the priors and the way they are defined are the result of the equilibrium defined by the theorist, they cannot be considered as the way players analyze the game from their perspective, and they cannot be interpreted as a mental variable, i.e. as a way to finally formalize the players epistemic states and mode of reasoning, as was the purpose of the epistemic program in game theory. Again this shows that the model remains a mathematical formulation and the player is a mere "logico-mathematical entity" (Gilboa, 2011, p. 303)

To void players of any identity is however philosophically impossible for Gilboa (2011, p. 305). An empty shell cannot have knowledge but the fact is that players are attributed some knowledge: of their strategy, of the set of player's belief to ensure the consistency of the hierarchy of beliefs as in Aumman's work for instance, of other's rationality, etc. Besides, as players are Bayes rational they maximize their utility given their beliefs about others' strategies or choices. It means that they must know their utility which is supposed to be an inner characteristic of their identity as it translates what they desire, what they want to achieve. Again this provides a sense of identity.

In the same way it is contradictory to assume that a player knows her identity, i.e. her type, which determines her strategy set and therefore gives her the knowledge of her strategy set and at the same time gives her beliefs towards this strategy set (e.g see Levi, 1995). Spohn (1977, p. 115) indeed raises the absurdity of this postulate: "anyone will find it absurd to assume that someone has subjective probabilities for things which are under his control". Thus while nature gives to the player the knowledge of her identity she must have at the same time beliefs regarding her own identity in order to finally reach the required common and consistent beliefs at the equilibrium.

This strategy tends to deprive players of their own identity, i.e. not to formalize the players as real individual is for Giocoli an epistemological consequence of the modern account of rationality since WWII, i.e. rationality as the consistency of choices: "a neoclassical agent is modeled as someone who always consistently reacts to incentives— where the requirement of perfect consistency is what characterizes his/her rationality as far as beliefs, expectations, plans and choices are concerned … the consistency view seems to entail the disappearance of no less than the main subject of neoclassical analysis, that is, the individual as a human being." (Giocoli, 2003, pp. 399-400) Recall that in epistemic game theory, consistency not only refers to the players' choice but also to their beliefs, i.e. their epistemic states or to put it differently, their mental variables. This is perfectly acceptable for Aumann as in his own words, "Game theory is a sort of umbrella or 'unified field' theory for the rational side of social science, where social' is interpreted broadly, to include human as well as non-human players (computers, animals, plants)." (Aumann, 1987b, p. 460)

From another front, Hargreaves, Heap and Varoufakis (2004[1995], p. 27) attack the non-human character of the players when they declare

> "So long as you treat other people as 'things', parameters like the weather, you can plausibly collect information on how they behave and update your beliefs using statistical inference, like Bayes's rule (or plain observation). But the moment you have to take account of other people as like-minded agents concerned with being fashionable too (a kind of common knowledge, like CKR), the difficulties multiply."

The nature of probabilities behind player's beliefs as defined in epistemic game theory, that is, as objective probabilities, confirm Hargreaves, Heap and Varoufakis' claim. The other players are in this manner like things for which observations and long runs frequencies can be established.

To sum up, many methodological and formal aspects of the epistemic models in game theory tend to show that the players that people the game are not real entities, which makes dubious the claim that their epistemic states, i.e. their mental variables are actually taken into account in game theory in contrast to classical game theory.

5.3. Rationality and reasoning: are they compatible?

In a game context "When I deliberate, I have to consider not only what the causal effects would be of alternative choices that I might make, but also what the other agents might believe about the potential effects of my choices and how their alternative possible actions might affect my beliefs." (Stalnaker, 1999, p. 3) Such counterfactual reasoning concerns players' thinking about others' choices. This could be related to the learning process that Morris (1995) is asking for, i.e. the transition from the prior to the interim and the posterior stages which finally, from various and free initial subjective beliefs, ends with consistent beliefs at the equilibrium. Counterfactual reasoning concerns the "off equilibrium path" however the existence of an off equilibrium path is precluded in the reasoning of players (Stalnaker, 1999; Morris, 1995; Levi, 1995). While it could be supposed that there is a form of learning between the prior and the interim stages there is no such thing (Binmore, 1993, 2009). Priors are only an analytical construct "that keeps track of the library of … posteriors" (Binmore, 1993, p. 234). The resulting problem is that the missing part of the epistemic model is the justification of why a player chooses to play a particular equilibrium or at least "a description of the process that leads players' beliefs to become mutually consistent." (Bicchieri, 1993, p. 85)

When players' strategies are included in the description, this raises a logical inconsistency. Players have priors for their own actions. This assumption is problematic (Gul, 1998, p. 927; Levi, 1995). How at the same time can a player know her strategies space and have beliefs on her strategies space? Besides, the joint prior space assumption conflicts with the common knowledge of rationality because it renders it impossible for a player to compare the desirability of all the states of the world. It makes counterfactual reasoning impossible, i.e. reasoning on the off equilibrium paths. Bayes rationality and common knowledge of rationality a priori excludes all of the non-

admissible strategies. Counterfactual reasoning is therefore impossible (see Stalnaker, 1999, for a comparative claim) and players' reasoning seems accordingly vacuous. Indeed assuming at the same time priors and common knowledge of Bayes rationality implies that prior to making a decision a player not only predicts what the others' will do but predicts her own actions and the strategy profile at the equilibrium (Levi, 1995, p. 175). But the fact is that according to the hypothesis of Bayes rationality a player should be able to evaluate all of the options she has prior to making a decision in order to compute her expected payoff for every states of the world.

Bayes reasoning entails that when a player has identified a strategy she considers admissible (i.e. which leads to the higher expected payoff) she knows she will implement this strategy and in Levi's (1995, p. 176) words "regards the matter as settled". Accordingly, the player updates her beliefs on the joint action space over the combination of the others' actions according to her choice and computes her expected payoff in such a situation. This is for Levi (1995, p.176) the "Bayes posterior expectation". The player forms this expectation by conditionalizing on the information that she has when having chosen and implemented a given strategy. The Bayes posterior expectations, once the choice is made, are equal to the Bayes expectations before making the choice (Levi, 1995, p. 176). Indeed, Bayes rationality states that the conditional probabilities over others' choices once a player has made a choice is equal to the players' conditional probability distribution prior to the decision. This condition in a strategic context is particularly doubtful since it would mean that others' actions are probabilistically independent of the player's own action (Levi, 1995, p. 176).

When resorting to a prior stage, the player is committed to the strategy she identifies before making her choice even if she should compare the expected utility reached with such an option with all of the other options (Levi, 1995, p. 180). As a consequence, "if an agent is certain prior to choice that he will not implement a certain policy, then from that agent's point of view at that time the policy is not optional even if we outside observers think the agent should not have been so certain. What we think the agent ought to believe is not relevant to our assessing the rationality of his judgments, as long as his beliefs meet the standards of weak rationality that theories of expected utility require. If an agent is certain that he will not perform a certain course of action, the agent is obliged to rule out choosing that course of action as available for him as an option." (Levi, 1995, p. 180)

For instance, in Aumann's framework retrospectively, in order for a strategy to be rational and therefore implemented it must be ratifiable, i.e. its expected value should be higher than the other strategies provided all of the potential choices of the other players when such a strategy is chosen. And prospectively there must exist a prior that is a correlated equilibrium (Levi, 1995, p. 180). Deliberation should identify a set of admissible options among all of the strategies available for players, however in such a framework, deliberation is vacuous since the priors already contain a correlated equilibrium. There is therefore no need to resort to ratifiability since everything is already settled prior to making a decision, a player does not need to compare an option with her other alternatives. "There is only one option available" (Levi, 1995, p. 180): the strategy which is already a correlated equilibrium.

To avoid this issue one of the assumptions, priors or CBR, must be dropped. Again the hypothesis of CBR should be a result of the reasoning process. It should result from a learning

phase, from a real prior to an interim stage in which players effectively update their initially free prior beliefs.

> "[CBR] imposes restrictions on the belief hierarchy of a player in a game. However, it does not necessarily tell us how a player reasons his way towards such a belief hierarchy. In that sense, the concept only imposes restrictions on the output of the reasoning procedure by a player, not necessarily on the reasoning process itself." (Péréa, 2013, p. 13)

In the same way, the prior which induces the "Joint Action Space Prior Assumption" entails that the players attribute unconditional probabilities to their own actions prior to the game (Levi, 1995, p. 181). Unconditional probabilities prior to players' decisions also leads to vacuous recommendations like common knowledge of rationality because the player must be sure prior to deciding that she will act rationally. Levi (1995, p. 181) suggests that the principle of rational decision cannot be postulated prior to choice if we want to model deliberating agents . If the object of game theory is to model the decision of rational players in a strategic context, game theorists should not attribute to players a priori unconditional probability distributions encompassing their own strategies. A player cannot at the same be free to choose an option according to her deliberation or her reasoning and consider her choice as a state variable (Levi, 1995, p. 182). Game theorists should therefore avoid considering that an agent is predicting her own choice a priori and when reasoning about the way the other players reason about herself, she should not assume that these other players consider that she has predictions about her own choice (Levi, 1995, p. 182)

But in Aumann's words:

> "The reader may still be puzzled by the fact that in deriving his posterior about other players' choices, each player conditions on his own information, which includes his own choice. How can a player's own choice help him to guess what other players will do? But again, this paradox is illusory. The player is not really conditioning on his choice, but on the substantive information that leads him to make this choice. This substantive information leads to a posterior on the other players' choices, which in turn leads to an optimal choice, or to a set of such choices. Intuitively, the fact that a player's choice is part of his information is used not so much in deriving his own posterior about what others will do, but rather in estimating the others' posteriors about what he will do. They cannot simply make arbitrary "guesses" about this, but must take into account that he is maximizing given his own information" (Aumann, 1987, p. 9)

It means that the inclusion of the player's own strategies in the description of the world is not descriptive but a purely notational and formal device to ensure the existence of each other's hierarchy of beliefs. It is at this condition that the other players can form beliefs on the choice of each other player, and therefore ensure the consistency of every player's hierarchy of beliefs.

However no answer to the question of why players should play the action dictated by the equilibrium concept is provided. "Neither the theorem nor the analysis provides any new argument for why one would expect actions to be generated in this way nor why beliefs for the players (or an outside observer) should be as if behavior were generated in this manner." (Gul,

1998, p. 926) – who claims "[the outside observer] has no private information, his probability must be a common prior."

The problem is that "games studied by an outside observers are different from the point of view of a player. Thus, an outside observer will have a probability distribution on the priors of the player of the game." (Kadane and Larkey, 1982, p. 116) The independence among each other's actions imposed in the description of the state of the world and the building of the player's hierarchy of beliefs are the consequences of this external perspective condition.

The reasoning process of players cannot be integrated in games if the norm is to apply a solution concept as in classical game theory. There is no room for a deliberating process from the perspective of the players as if they were in the shoes of real players involved in a strategic interaction. Therefore, neither from a normative perspective (stating what the players should do), nor from a positive perspective (stating how the players effectively reason), does an epistemic model of games provide an acceptable account of a strategic interaction.

To the claim that there is no freedom of choice in an epistemic game in game theory because player's own strategies are included in the description of the world and accordingly settled before the game, Aumann (1987, p. 8) answers that ""freedom of choice" is not an issue" as each player analyzes the game as an external observer; which in consequence, entails that the players treat each player's private information as specific descriptions of the states of world:

> "The chief innovation in our model is that it does away with the dichotomy usually perceived between uncertainty about acts of nature and of personal players. Of course, a player always knows which decision he himself takes; here, this information is not treated differently from private information he may have about other aspects of the state of the world. This point is a little subtle and is worth some discussion. In traditional Bayesian decision theory, each decision maker is permitted to make whatever decision he wishes, after getting whatever information he gets. In our model this appears not to be the case, since the decision taken by each decision maker is part of the description of the state of the world. This sounds like a restriction on the decision maker's freedom of action; at a given state w, it is as if the model forces him to take the decision dictated by w. But closer examination reveals that the model describes the point of view of an outside observer. Such an observer has no a priori knowledge of what the players will choose; for him, the choices of the players are part of the description of the states of the world. This does not mean that the players cannot choose whatever they want, but only that the observer will not in general know what they want." (Aumann, 1987, p. 8)

The outside observer argument allowing theorists to justify controversial assumptions, such as the existence of CP, is for some scholars not sustainable because it would require that the analysis of this outside observer would be incorporated in the model. Her belief should therefore be formalized like those of the players (Gul, 1998, pp. 926-27). For Gul (ibidem) however this does not guarantee that the outside observer's priors are a correlated equilibrium. The common prior assumption is not sufficient to lead to this result. Aumann answers to this criticism that if the outside observer's perspective is taken into account in the game so that her beliefs are integrated

they would also be a correlated equilibrium. Thus, for him, neither of the results or axioms stated in his model would be different:

> "when the term "outside" has its straightforward, usual meaning, as in [Aumann, 1987] – that the observer has no private information. If, as Gul suggests, we take him as an ordinary person with his own beliefs, then our axioms (i.e., the CPA) apply to him as well as the players; so since he has no private information, his probability must be a common prior of the players. According to Aumann, [1987], this is a correlated equilibrium. In any case, outside observers are not central to our view; in Aumann, 1987 they appear only once, and then not in connection with the CPA." (Aumann,1998, p. 937)

Therefore, the following statement that Aumann makes is hardly tenable if we take as granted his previous arguments:

> "While a part of each player's information is undoubtedly generated by overt signals from the outside world, another part is obtained by reasoning about how other players reason. … Some readers might prefer to call that kind of information "endogenous." In actuality, much information is a mixture of both kinds, which it is not easy to untangle." (Aumann, 1987, p. 11)

That is why again we are tied to the claim that a game in epistemic game theory remains a mathematical abstract representation that ultimately does not integrate player's strategic reasoning (see Lecouteux, 2018b, pp. 28-29). As Kadane and Larkey (1982) emphasized early on, an empirical literature on the elicitation of individuals' beliefs must be integrated into game theory. There is no such concern about the elicitation of beliefs in game theory. To their knowledge, the only attempt made was by Savage but it was for one person decision-making problems which does not fit with the strategic dimension of game theory (see also Binmore, 1993). EGT is therefore grounded on the assumption that rational behaviors in strategic interactions are an extension of standard decision theory in which the realization of the state of the world depends on the strategy of other rational players, i.e. on the result of the others' reasoning process. The strategic dimension of games, i.e. players' ability to form expectation about the others' choice, should be captured by the assumption of CKR, or, by a weaker requirement: common belief in Bayes rationality. However, since each state of the world captures the result of the rational deliberation of each player, i.e. the choice of the players and that players prior must already be consistent with rationality and the equilibrium, there is no need to guess what a rational player should do in each circumstance (see also Lecouteux, 2018b). The result is already known. Such an assumption entails that players never have to represent what the others are thinking or wanting. This explains why some philosophers and game theorists attack the epistemic program of game theory for the vacuousness of players' reasoning processes.

5.4. Mentalism vs. behaviorism

Economists fear both the EUT and the SEUT as "a return to a 'metaphysical' characterization of rationality, imbued with unverifiable mental variables and a naïve psychologism" (Giocoli, 2003, p. 388) but on the contrary, both provide "a general characterization of rational behavior totally devoid of any psychological assumption." (ibidem) Savage's formalism of SEUT, on which epistemic game theory is grounded, offers an axiomatic belief in which, if the players' beliefs respect the axioms he decreed, they are consistent, in a Bayes sense. Savage therefore extended the consistency criterion from choices to epistemic states: the players' beliefs.

If the agents' choices are consistent as specified by the axioms it is as if they have maximized their subjective expected utility, i.e. their utility according to the "subjective" probability distribution regarding the set of possible states of the world. The subjective dimension of the probability distribution regarding the possible states of the world and which represent the players' beliefs "arises as a consequence of his axioms" (Luce and Raiffa, 1957, p. 304).

According to Savage's methodology, the choice of the players not only reveals their preferences but their beliefs (regarding the probability of occurrence of the events under scrutiny). That is why the term "revealed subjective probabilities" is given by Giocoli (2003) with respect to Savage's SEUT. Subjective probabilities are revealed according to a gamble set up. A finite set of states of the worlds or events can occur and the gambling procedure lies in the bet the individual takes regarding the likelihood of occurrence of the possible states of the world.[13] The probabilities that an individual attaches to the occurrence of the different possible events or states of the worlds is revealed by her choice, however the "result derives from the particular set of axioms" (Giocoli, 2003, p. 392). This combination of the axiomatic and behavioristic approaches provides for Giocoli the culmination of the escape from psychology. He indeed declares that "the final step(s) in the process" are vN/M's EUT and Savage's SEUT. Recall that in vN/M's EUT players' utility is not representing "a particular psychic state" but is "an analytically useful implication of the set of axioms." (ibid, p. 379) EUT is simply "a formal deductive consequence of a set of axioms." (ibid, p. 380) Probabilities in this case are objective probabilities, i.e. they represent mere long run frequencies (von Neumann and Morgenstern, 1953[1944], p. 19). They are derived from the observation of regular events.

This is why according to Levi (1995, p. 180), as in Aumann formalism, Bayes rationality is "retrospective … it applies to the evaluation of options by agents once the choice has been made". However, such rationality should be and is supposed to be prospective too. In a prospective account the strategy that a player should rationally choose and implement, is recommended given the information she has before she effectively chooses (ibidem). A central

[13] Giocoli (2003, p. 392) specify that "[t]he tradition in SEUT is to circumvent the problem either by adding an extra postulate (the sure-thing principle: ibid, pp. 39—40) or by having recourse to an imaginary experiment, a lottery able to turn the preferences' domain into a probabilistic space. The former is the approach originally followed by Savage but it is quite involved, so it is the latter method, first proposed in the early 1960s by Anscombe and Aumann, that has gained acceptance."

issue of EGT, as already highlighted, is that the prior beliefs supposed to capture the rationality of the players through the assumption of common belief in rationality are defined ex post, as a mere representation of the players' choice – they cannot therefore be treated as an input of the decision process. While the terminology itself of 'prior beliefs' suggests that there exists an actual hypothetical prior stage from which, given the subsequent realization of the state of the worlds, the players can update their beliefs – and then maximize their expected payoff given those 'posterior' beliefs – it appears that the priors have no genuine substantive meaning in EGT. Gul (1998) labels the former interpretation of priors as the "prior view" and contrasts it with the "hierarchy interpretation", in which case "the priors are artifacts of a notational device to represent the infinite hierarchies of beliefs on [the players'] "posteriors" at the true state of nature." (Gul, 1998, p. 925) Aumann and Brandenburger (1995) explicitly endorse the latter interpretation, by emphasizing that "an interactive belief system … does not suggest actions to the players. Rather, it is a formal framework – a language – for talking about actions, payoffs, and beliefs." (Aumann and Brandenburger, 1995, p. 1174)

A major difficulty arises from this behavioristic interpretation of payoffs. Heidl (2016, pp. 26–44) indeed suggests that preferences and payoffs can be interpreted either in a mentalistic or in a behavioristic way. According to the mentalistic interpretation, "preferences are understood as scientific refinements of the folk psychological concepts of desire and preference" (Heidl, 2016, p. 26), while the behavioristic interpretation is that "preferences are not mental entities but consistent patterns of choices" (ibid, p. 27). Although the mentalistic interpretation leaves space for the integration of the players' reasoning process, the behavioristic interpretation – by understanding payoffs in games as vN/M utilities rather than material payoff – has attempted to erase the psychological dimension of preferences and choices (Hausman 2000, p. 115; Heidl, 2016, pp. 21-27). The behavioristic interpretation defining payoffs as vN/M utilities rather than material payoffs, means that the primitive of the game is the players' choices: the utility functions are defined ex post, as a representation of their choices. VN/M (1947) indeed showed that, if the choices of a player respects certain axioms, then it is as if the player was maximizing an expected utility function. Similarly, Savage (1954), Anscombe and Aumann (1963), and Aumann and Drèze (2008) show that, if the player's choices respect certain formal conditions of consistency, then we can define a utility function and subjective beliefs such that the action that maximizes the expected subjective utility of the player is precisely the choice we observed. Defining the payoffs in games as vNM utilities was an attempt to get rid of psychological and unobservable variables such as the (subjective) tastes and beliefs of the individuals (Sen, 1973, p. 243; Hausman, 2000, pp. 100-101; Lehtinen, 2011, p. 289; Grüne-Yanoff and Lehtinen, 2012, p. 505; Heidl, 2016, pp. 33-40; see also Bacharach, 1986, 1989).

The behavioristic interpretation raises serious methodological difficulties regarding the status of prior beliefs. Hausman (2012, pp. 28–33) indeed highlights that, if choice is jointly caused by preferences and beliefs, then we cannot simultaneously deduce preferences and beliefs from the choice of the players. If it were the case,

> "the payoffs [in the behavioristic interpretation] would say how individuals would choose. They would already incorporate the influence of belief, and belief could play no further role. If the revealed-preference theorist were right and payoffs already represented what

> strategy was chosen, there would be nothing left for game theory to do." (Hausman, 2000, pp. 111–112)

This means that the players' priors would be settled before the game, and the result of the other players' deliberations (i.e. their choice) would be comprised in the descriptions of the world provided by the prior stage. For a given equilibrium strategy profile (e.g. a correlated equilibrium for Aumann, 1987), we can build a prior belief such that the expected utility maximizing choice of the players based on their posterior beliefs (which result from the realization of the state of the world) corresponds to this equilibrium strategy profile. In this case, the players' priors are settled before the game, and the result of the others' reasoning process, i.e. their choice, is comprised in the descriptions of the world provided by the prior stage. Recall that Levi (1998, p. 175) refers to this approach as the "Joint Action Prior Space Assumption" and argues that endorsing it makes deliberation vacuous (ibid, p. 181), since players' beliefs are deduced from the equilibrium profile of strategies (Battigali, 1988, p. 705; Bicchieri, 1993, p. 81; see also Roth and Schoumaker, 1983). That is why, claims Binmore (1993, p. 330), players are "frozen" into a profile of strategy at the equilibrium: their respective choices describe the equilibrium choice as defined by the theorist. Accordingly, epistemic game theory cannot explain how the player reasons toward this equilibrium. In the same way, classical game theory cannot explain why players play a Nash equilibrium. As Gilboa (2011, p. 304) puts it in different terms; if players are empty shells as Aumann (1992) claims, then their preferences must be "purely hypothetical." There is thus an ambiguity around the term "utility" for him. In the same way players' beliefs are thus also purely hypothetical; they "are induced by equilibrium strategies with Bayes formula" (Battigali, 1988, p. 705) They are not the grounds of the players' strategic reasoning as would supposedly be the case in a Bayes analysis of games (see Lecouteux, 2018b, p. 22)

Therefore both in the EUT and SEUT the players' payoffs as described in terms of utilities are a mere description of their choice. The payoffs represent both the choices and the beliefs of the players. In this case it means that it is not players' epistemic states, i.e. their beliefs and preferences that are "the primitive of the analysis" (Lecouteux, 2018b, p. 314). As Aumann (1998, p. 936) claims "a common prior derives naturally from the axioms."

This means that with such a behavioristic flavor, the players' choices in games cannot be explained. For instance, cooperation in a PD or coordination and a Hi-Lo game cannot be justified because in this revealed preference interpretation of payoffs, the payoffs are already a representation of the player's choice. That is why sustaining cooperation and coordination would be contradictory within the matrices specified (see Binmore, 2007, 2009; Lecouteux 2018a). Sugden however does not agree with Binmore's "strong" interpretation of payoffs, which enable anyone to say something on player's behavior from descriptive and normative perspectives: "Binmore uses a particularly strong form of revealed preference theory, in which it is a matter of definition that an individual's choices always reveal her preferences. Thus, once a game has been specified, with utility indices for the various possible outcomes, certain propositions about what a player will do (for example, that she will not choose a dominated strategy) are necessary truths, and not merely the implications of particular solution concepts which game theorists are free to dispute (I: 104-110). I am not convinced that this is the most useful - or indeed the conventional - way of interpreting utility in games." (Sugden 2001, p. 222)

6. Conclusion

The main aim of game theorists throughout the history of the discipline, even in its modern form of epistemic game theory, has been to get rid of psychology and mental variables. However, it appears that the main improvements for non-cooperative game theory, i.e. explaining how and why an equilibrium occurs, can come only from the inclusion of such mental variables and psychological dimensions which entail the introduction of effective reasoning processes in games. Indeed, both from a normative and a positive perspective, game theory would strongly benefit from an explanation of how and why a solution occurs. The remainder of this thesis will first provide examples with contributions from two game theorists showing that the introduction of real reasoning processes and the incursion of psychological dimensions in game theory induce these improvements. Second we will explore theories from cognitive psychology and from philosophy of mind that can provide fruitful thinking on the elicitation of player's beliefs in order to overcome the indeterminacies of players' priors as understood in terms of real beliefs, i.e. in terms of mental variables. And we then integrate one of these theories into a game model to furnish a line of inquiry and discussion regarding a new methodology to explain how, why and whether a specific solution occurs.

As shown in this chapter, if the payoffs matrix describes the players' expected utilities or subjective expected utilities then the game is settled, the equilibrium known and players' beliefs prior to the game are at the equilibrium. Everything is settled before the game and the game just offers a representation of a possible way in which the interaction has occurred (Hargreaves Heap Varoufakis, 2004[1995], p. 274; Lecouteux, 2018b, pp. 3-14)

This is mainly explained by the norms of research into game theory, driven all along by mathematical concerns: "the research standards in game theory are those of mathematics research (Kadane and Larkey, 1983, p. 1377) which induce the focus on solution concepts, but the interaction process through which a specific solution is reached is left undefined (e.g. see Giocoli, 2003; Lecouteux, 2018). From that perspective "what we need is an empirically supported psychological theory making at least probabilistic predictions about the strategies people are likely to use, … given the nature of the game and their own psychological makeup." (Kadane and Larkey 1982b, p.124, quoting Harsanyi 1982a, p.122). Game theory understood from such perspective necessitates the introduction of real mental states such as players' beliefs but not in a static and equilibrated way as is the case in epistemic game theory. Chapter 2 and 3 will explore two authors who introduce the idea of players as real individuals with their own "psychological made up", and who introduce reasoning processes and beliefs understood as mental variables. Indeed as Rubinstein (2001, pp. 617-18) underscore,

> "A 'game' varies according to whether the players are human beings, different 'selves' of the same person or bees. A strategic game changes entirely when payoffs are switched from utility numbers representing von Neumann and Morgenstern preferences to sums of money or to measures of evolutionary fitness. The formal model is identical with any of these interpretations but the models per se are not."

Chapter 4 of this thesis will provide such "empirically supported psychological theory". It will expose the psychological mechanisms that can help to explain how the players' beliefs may become consistent so that they sustain an equilibrium profile of actions, i.e. the existence of a solution. Indeed "[t]he missing part in the transition from assuming rational behavior and beliefs to deriving an equilibrium configuration of actions is precisely an account of what grounds players' mutual expectations and a description of the process by which players' beliefs come to agree." (Bicchieri, 1993, p. 34)

Chapter 5 will integrate such theory into game theory. Besides, as suggested in the quotation above, by changing the payoffs of the games and no longer assuming that they describe vN/M utilities we can free ourselves from the behavioristic interpretation of preferences and beliefs which led to the stalemates described in this chapter.

Morris (1995, p. 228) emphasizes that the partisans of the CPA often claim that if the hypothesis of common priors is dropped then game theorists argue that "any outcome … is consistent with heterogeneous prior beliefs." But he shows that this justification is not enough to preclude heterogeneity and this is not completely true: there can exist a device that can allow postulation that not all of the potential heterogeneous beliefs are acceptable. Chapters 4 and 5 will explain how to provide such a methodological device.

We argue that epistemic game theory should explain how the players' beliefs can converge to a specific solution but without postulating such a state of consistent beliefs prior to the game, i.e. before the game, and before any reasoning of the players on their decision problem. This necessitates an explanation of how players form their beliefs about others' reasoning and beliefs, and therefore an explanation from a player's perspective, of how she assesses the game she is to play with her co-players. The remainder of the thesis will be devoted to this explanation.

In addition, the indeterminacy problem is not solved by the epistemic models of game theory (Brandenburger, 2010, p.6; Lecouteux, 2018a, p. 21) which, according to us, show that the resolution process of games in classical game theory are not investigated. Investigating how and why the players of a game can reach a solution provides the only possible attempt to bypass this indeterminacy problem.

The remainder of the thesis will thus show how changing the ontology of game theory and of games can lead to fruitful answers to the main theoretical and methodological difficulties that game theory faces and that have been exposed in this chapter. We defend the idea that standard game theory has remained a mathematical theory which prevents the improvements needed to overcome such difficulties. We thus progressively show how to renew the object of study and the purpose of game theory as a framework of analysis. Such a framework of analysis should have as a purpose the investigation of the players' strategic reasoning process which ultimately necessitates a focus on the kind of interdependence among players involved in strategic interactions. Reasoning about someone else and anticipating her choices generally requires more than common knowledge or common belief of rationality. We investigate such dimensions and ultimately show how the integration of new individual and collective determinants in players' reasoning offer new perspectives on this intersubjective dimension of gaming. One of the main claims of the thesis is the prime importance of investigating and identifying the determinants

involved in the intersubjective dimension of strategic reasoning. The change of ontology asked for in this thesis brings up the prime importance of putting thinking on intersubjectivity in games at the forefront of game theory.

This calls for a new epistemology of games in which games must be understood as real processes of reasoning and not mere descriptions of individual and independent choices as is the case in standard game theory. We will show that this necessitates in the first instance a departure from a purely mathematical account of games and game theory in order to consider games equilibrium and games solution as a result of the games and not as an hypothesis of the models of games. The analysis of games should not be focused on the analysis of the equilibrium from a mathematical point of view but on the process that lead to such equilibrium and to the solution. This calls for a major shift in the standard epistemology of games.

Such a new epistemology entails a real form of interdisciplinarity as it requires integration of the analysis of game players' mental states. If, as we claim in the thesis, the reasoning process of the players must be integrated, players' epistemic states such as their beliefs must be considered as real mental states. Players' beliefs, intentions and perceptions must be taken into account. Players as real individuals in flesh and blood must be integrated into games and game theory contrary to what is done in standard game theory. This necessitates real interdisciplinary intercourse among game theory and other social sciences: cognitive psychology, sociology and philosophy, and in particular the philosophy of mind.

We will show that Schelling and Bacharach, with the integration of players' reasoning processes in games and, from a methodological point of view more precisely, of players' frame?, open up a new research program which ultimately brings a deep thinking and a deep renewal of the intersubjective dimension involved in strategic reasoning. But we show, once more, that this requires a new ontology of game and of game theory.

Chapter 2

Schelling's reorientation of game theory: towards a theory of interdependent decisions

1. T. C. Schelling: a dissent economist?

T. C. Schelling appears in the 1950s as a 'strategic thinker' for the Nuclear Age (Ayson, 2004). His work on nuclear weapons and arms race during the Cold War sets his notoriety and his now widely known original contribution to game theory. At that time he lays the foundations of his *Theory of Strategy* which is the cornerstone of his thinking (e.g. see Ayson, 2004). In this perspective, he intended to *reorient* game theory. The prime objective was to apply game theory to practical strategic decision problems (e.g international negotiations, arm race, and nuclear war). However, what Schelling was doing in this perspective was also deeply formally challenging. He ultimately offers conceptual and methodological solutions for some of the main game theoretic shortcomings even if such solutions were not formalized.

We will show that Schelling actually anticipates some of the major weaknesses of contemporary game theory: namely the inadequacy of its tool to account for the convergence of players' beliefs, i.e. for their capacity to coordinate, which ultimately justifies the existence of an equilibrium. His work invites to an ontological thinking of the nature of games and of strategic interactions. He assessed the mathematical constraints imposed on game theory and called for the opening of game theory not only to economics but also to other social sciences such as psychology, sociology, history, law, and so on. This interest in interdisciplinarity is very much modern: we can witness such a new approach to game theory very recently (the chapter 5 will be an example of this new interdisciplinarity). It has indeed taken some decades for influential theorists to see economics and game theory as an 'open-system' like Schelling. Conceiving economics as an open system means acknowledging that "outside factors influence economics" (Mearman, 2003, p.15; see also Chick, 2004, p.5). Besides I will show that Schelling was building a theory of strategic interdependence which was for him the core of economics instead of a set of tools – a view that is again very innovative for his time (see Chassonnery-Zaïgouche and Larrouy, 2017).

Schelling (1984) characterized himself as an "errant" economists and has always be widely recognized as a "dissent" economist, at least, as far as he became famous with the first publication of *The Strategy of Conflict* (e.g. see Latzko, 1998).

Schelling shows his originality quite early by considering game theory as a mere '*framework for analysis*' (in Swedberg, 1990), he sees its utility "at the most elementary level" (Schelling in Aydinonat, 2005, p. 1). Schelling considers that game theory is "a way of thinking about a problem" more than a mathematical device (in Swedberg, 1990, p. 189).[14] Having in mind that game theory in the 1960s is dominated by the contributions of von Neumann and Morgenstern on the one hand and Nash on the other hand, and therefore that game theory is considered as belonging to the realm of mathematics more than economics, we understand why Schelling appears as a dissenter (see Giocoli, 2003). By presenting his theory in its mere simplicity, by multiplying clear and concrete situations of real game-like situations stemming from everyday experience, Schelling nevertheless indisputably contributes to the fragile and yet onward expanding interest in game theory from economists.

> "The publication of von Neumann and Morgenstern's (1944) pioneering book on game theory was immediately hailed as a milestone in the social sciences. Its promise took a long time to fulfillment, but is now essentially complete. Game-theoretic concepts, terminology and modes of analysis have come to dominate most areas of economics, and have made large inroads into political science and other social sciences. Thomas Schelling deserves a considerable part of the credit for this." (Dixit, 2006, p. 215)

This partly explains why, when awarded the Nobel Memorial by the Royal Swedish Academy of Sciences in 2005 Schelling is presented as a "pre-eminent pathfinder" (in Colman, 2006, p. 607). He upgraded game theory to concrete situations of strategic interactions other than conflicts, which means – from a theoretical point of view – broadening game theory to non-zero-sum games, and in terms and style approachable by social scientists (contrary to Nash (1950, 1951)'s work for instance).[15] As emphasized by Rivzi (2007, p. 404)

> "Because of his broadly appealing style, wide-ranging choice of topics, and his willingness to step outside the orthodoxies of the profession, Schelling has had great influence outside of economics and, within economics, in encouraging a distinctive style of research."

According to Kenneth Boulding's (1957) classification, Schelling appears as a dissenter from the orthodoxy of economic thought, i.e. he disagrees with the ideas, customs, and methodology of mainstream economics (see Latzko, 1998). He indeed shows that economics offers tools that in many cases seem inadequate for understanding economic reality (see Schelling, [1960]1980, 1984, [1978]2006). He also appears to assess the application of the principles of the natural sciences to the social sciences. He intends to show that the methods of natural sciences must be coupled with the methods of social sciences such as law, economics, psychology, etc. According to him, scientific methods imported from natural science are valid for objective facts that are external

---

[14] Schelling's reluctance for using sophisticated mathematics is often cited (e.g. see Ayson, 2004; Rivzi, 2007; Aydinonat, 2005; Colman, 2006; Innocenti, 2007). According to him there is no need to refer to mathematics.

[15] See Dixit (2006, pp. 215-16) who claims that even if Nash considerably broadened game theory when proving the existence of equilibrium for many players non-zero sum games (compared to the predominance of two players zero games in von Neauman and Morgensterm's work), "it was at too high a level of abstraction for application" and economicts did not had a sufficient training in mathematics to understand this work.

from human perception and psychology. On the contrary, as demonstrates his work on human decision-making, most economic matters concern mental states, i.e. people's perceptions of facts or beliefs which are subjective: they depend on players' perceptions. Subjectivity is here understood in a broad sense; it merely implies that it is primarily the players' perceptions that determine their beliefs, intentions and then their behavior.

Economic facts are rarely accessible at first sight; they are first and foremost contingent on the economic subjects' apprehension of them. Appraising people's mental states requires psychology, i.e. to resort to cognitive sciences and more specifically to cognitive psychology. Economics thus needs to be an 'open science' that does more than importing mathematical tools. In other words, to understand players reasoning and mental states, economics needs to be open, i.e. conceived in connection with the other social sciences (such as cognitive psychology, sociology, history or law) and not be immune to theories and methodologies coming from these other disciplines.

Schelling can also be considered as a dissenter when having a look at the scope of the topics he investigates. He explores domains that are not traditionally studied by mainstream economic analysis and applies some of the economics tools, like game theory, to domains that were considered as belonging to the realm of psychology or sociology for instance (e.g addiction, organized crime, or segregation).[16] This is linked to Schelling's conception of economics as an open science. He does not see economics as a confined practice relevant for studying market forces only. In the same way theories and methodologies from other social sciences are relevant to understand strategic situations and players' reasoning in strategic contexts, economic theory and economic methodology can be relevant for studying phenomena such as addiction, organized crime or residential segregation. These domains are all about interdependent decisions (between different people or between different selves: the me of today and the me of tomorrow in case of addiction; they are about rationality and lapse of rationality, etc.). Hence they should be the concerns of game theory and decision theory.

Schelling sees economics as a framework of analysis with at its core a theory of interdependent decisions. Economics is therefore not only a set of tools – some of which need to be amended – and is more than a particular topic of interest – i.e., basically, more than the market. He indeed proposes to reorient game theory or consumer theory so that they can be more congruent with the actual behavior of the real world individuals. It implies for him that game theory must be contextualized to the subject under study. Using game theory for understanding a specific phenomenon not only requires the application of a mathematical tool but the broad apprehension of the phenomenon and the adaptation of the tool for the specificity of this phenomenon. His observations of the real world help him pointing out the insufficiencies of mainstream economic theory. He uses them to explain in simple terms and with illustrations the analytical improvements he proposes, and at the same time, they serve as an experimental setting to calibrate and verify the applications of his theoretical propositions and analysis. This specific methodology explains the impact of his work on game theory outside the community of mathematicians: in economics, in law and international trade just to mention these three domains.

---

[16] However the way Schelling investigates domains that are traditionally considered as non-economic, is very different from Economic Imperialism (see Swedberg, 1990)

To understand how dissent Schelling is from the standard topics of interests in economics, a brief recall of his carrier showing the variety of domains of research in which he contributed, and from the methods of classical game theory in particular, the RAND Corporation context and the landscape of game theory at the time Schelling breaks into, seal the matter. Schelling's originality is indeed for a great part the result of his carrier. While becoming known after the first publication of the *Strategy of Conflict* in 1960, Schelling's interest in strategic thinking dates back to his early work on international negotiation when working as an economist for the Marshall Plan, in the late 1940s. As he claims

> "in 1948 I was just finishing graduate study. And I had an opportunity to go to Washington to join the new Marshall Plan which began in April of 1948. And I continued working on foreign aid negotiations until 1953, at which time I went off to Yale University on a faculty but decided I think negotiation is the most interesting thing I've ever been involved in, and I'm going to make that my study, so that bargaining, conflict, cooperation, all led me into the studies of strategy. And eventually I got into game theory." (Schelling in Jeffrey Brown's interview, 2005)

Military strategy was just an application of his thinking of negotiation, i.e. an application of strategic interdependence like negotiation. As Ayson (2004, p. ix) emphasizes "Schelling the strategist is a dimension and application of Schelling the social scientist." In fact, Schelling's different involvements in national and international council, i.e. his extra academic work, constitute the raw material for his academic contribution, i.e. his analytical and conceptual contribution. They all set the underpinning of his work on Strategy. He indeed sees economic matters as situations of interdependent and strategic decisions quite early, and enriches his work on interdependence throughout his carrier whichever the domain of interest and application. He progressively modulated his analytical framework and searched for tools to account for his broad conception of interdependent interactions (see section 3 of this chapter on the dynamic models of residential segregation).

Schelling worked at Washington for the United States Bureau of Budget and for the Economic Cooperation Administration, which led him to negotiate the allocation funds of the Marshall Plan in which he was involved from 1948 to 1953. When working on international negotiation for the Foreign Aid Bureau, Schelling became increasingly interested in bargaining situations. Then, while serving two years at the White House and the Executive Office of the President, he was re-affected from the Foreign Aid Bureau to the Mutual Security Program. He joined the RAND corporation discussion group during the summer 1956. After that, he became adjunct fellow at the RAND Corporation during the summer 1957 and finally a staff member of the RAND Corporation from 1958 to 1959. He was recruited to work on "new strategic thinking" (Sent, 2006, p. 3). These periods at RAND Corporation played an important role in the development of his thinking on strategy and stability. Schelling increasingly enhanced his training to game theory. And he came with the first publication of the Strategy of Conflict (1960), the Strategy and Arm Control with Morton Alperin (1961) and finally Arm and influence (1966).[17]

---

[17] See Ayson (2004) and Sent (2006) for a more detailed version of Schelling's academic and extra academic carrier.

The RAND Corporation period plays a significant role in Schelling's apprehension, conceptualization and uses of game theory. Indeed, "RAND became unquestionably the leading center for the development of game theory between 1946 and 1962" (Hounshell, 1997, p. 253). Famous game theorists such Shapley, Dresher, McKinsey, Flood, Savage, Nash, and von Neumann were involved in the RAND Corporation.[18]

The 'mathematical foundations' of game theory have been layed within the RAND Corp. (Innocenti, 2007, p. 417). We must mention the role played by the RAND Corp. in the development some of the most important analytical tools on which neoclassical economics has been grounded after the Second World War such as "linear and dynamic programming, system simulation, game theory, and artificial intelligence" (Hounshell, 1997, p. 244). Such mathematical revolution played an active role in "the changing nature of mathematics and mathematical economics" (Sent, 2007, p. 459) which gave rise to the success of neoclassical economics.[19] Besides, one of the consequences of such deep transformation is that economics became related to a 'mathematical tool-kit' rather than a specific domain of interest (ibid, 458). This echoes Schelling's use of the economic toolbox for non-traditional areas of study for economists (see Schelling, 2006 [1978]). However, as Innocenti (2007, p. 417) points out, "he addressed some criticisms to game theory that conflicted with some of the principles endorsed by that same community." Schelling did not conform to the mathematization of economics and game theory in particular. His concern for the application of game theory to real strategic situation and his penchant for real gaming led him to such criticism (Leonard, 2010, p. 340). Schelling indeed regularly claims that game theory should not be considered as a 'mathematical toolbox' but as a conceptual framework for social scientists.

In the late 1960s Schelling extends his range of research from military strategy and arm control to topics such as residential segregation, organized crime, environmental issues and energy. While he began his academic carrier on a traditional economic matter, namely international economics and in particular trades and international pricing, as Samuelson (in Zeckhauser, 1989, p. 157) points out "[o]nce the vital game of survival in a nuclear age challenged Schelling's attention, mere economics could no longer contain him."

Before the end of the Cold War, the RAND Corp. evolves from a "pure Cold War institution" to a major think-tank working on a wide range of subjects. The budget allocated for military

---

[18] The vast amount of production of the RAND Corp. in game theory can be evaluated by looking at the RAND Memorandum compiled in the *Index of the selected publication of the RAND Corporation, 1946-1962* and in the *Project RAND publication index*. Besides, as Hounshell (1997, p. 253) emphasizes a look at the bibliography of Luce and Raiffa's book: *Games and Decisions: Introduction and Critical Survey* (1957) shows the vast legacy of publications in game theory offers by the RAND Corp.

[19] The 1960s saw changes in the way to characterize the economic world in terms of a "system of forces" into an explanation in terms of "relations" (Giocoli 2003, p. 4). In the "system of forces" account of economics, the explanans is the markets—i.e. the forces of supply and demand on such markets—while in the "system of relation" the explanans became the individuals' choice and rationality, where rationality is understood in terms of consistency of choices (Giocoli, 2003). Neoclassical economics became a system in which "[i]f the rules of mathematics have been obeyed, the model exhibits internal consistency, which is the criterion of acceptance of such models." (Chick, 2004, p. 8) and in which "internal consistency is the only test of rigour and criterion of theory acceptance (ibid, p. 10) The methodological implication is "that only closed-system theorising is acceptable." (ibidem)

researches was redistributed in other domains and in particular social and civilian domains like racial issues. From that period "the analytical methods, tool, and penchant for research that RAND had manifested at the height of the Cold War were actively engaged in the "War on poverty"" (Hounshell, 1998, p. 267)

This switch of interest translates in Schelling's work. As Ayson (2004, p. 113) emphasized "Schelling's insight was to transfer certain ideas from one body of thinking to another, but he also adapted and developed this theoretical framework to suit the particular situations typically encountered in strategic studies." Schelling came to apply his conception of interdependent choices and behaviors in his strategic analysis to the topic of residential segregation.

> "[F]or a variety of reasons, Schelling was looking elsewhere for the raw material for theory building. He found useful sources in such questions as organized crime and racial segregation. However, these were not quite so new and distinct from his earlier interests as they may seem. Schelling had not only raised many of these issues in his work on the strategy of conflict, but at time did so in a way which suggest the underlying unity of all of the strategic behaviour which he studied." (Ayson, 2004, p. 37)

Schelling, at the same time, developed and applied his conceptual framework for and to the analysis of a large scope of topics involving interdependent choices and behaviors implying a mix of negotiation, conflict and cooperation. These specificities of real strategic reasoning, i.e. of strategic interaction in the real world cannot be formalized in non-cooperative game theory. Recall, from the chapter 1, that the only relevant information in games is objective facts, i.e. the specification of the rules of the games: the individual payoffs and strategies for vN/M and the individual preferences and rationality of the players for Nash. The use of negotiation, conflict deterrence, bluff, or cooperation relies on some subjective perceptions of the situation, on personal evaluation that are contextual and dependent on the process of interaction that cannot be incorporated in a payoff matrix as standardly defined. Sent (2006, p. 3) lists some of these domains: "military strategy and arms control, energy and environmental policy, climate change, nuclear proliferation, terrorism, organized crime, foreign aid and international trade, conflict and bargaining theory, racial segregation and integration, the military draft, health policy, tobacco and drugs policy, child rearing, taxi driving, investing in the stock market, tax collecting, house buying and selling, voting, playing charades, striking, price wars, traffic jams, kidnapping, daylight savings, etiquette, Lot's wife, selecting Miss Rheingold, as well as a variety of ethical issues in public policy and business".[20]

As a consequence, in the 1960s

[20] Each of these domains of interest leads to academic works. We can mention Schelling (1960c, 1966) and Schelling and Halperin (1961) for nuclear strategy and arms control, Schelling (1967, 1984g) for organized crime and extortion, Schelling (1980, 1992a) for addiction and self-control, Schelling (1971a, 1972) for racial segregation, Schelling (1990b, 1992b) for global warming, Schelling (1979) for energy policy, Schelling (1983) for environmental protection, Schelling (1955) for foreign aid, Schelling (1968) for the value of human life and Schelling (1980, 1984b, 1984d) for self-command.

> "The picture emerges of a strategist thinker whose ideas are linked into many of the intellectual currents of his day from the whole host of disciplines. This extends from economic and game theory (where Schelling's interest is well known but at times rather misunderstood and underappreciated) to theories of organization, information and social psychology. As a result, Schelling's thinking needs to be seen as more than a theory of strategy for the nuclear age, and more in terms of a quest for a general social-scientific theory within which problem of strategy exist as an important category. This underscores the importance of considering the history of strategic ideas in terms of the history of ideas as a whole." (Ayson, 2004, p. 2)

While at the same time challenging the mere foundations of the game theory as developed within the RAND Corp., some of its methodological underpinnings pervade in Schelling's thinking. Indeed, since its early days, one of the main features characterizing the RAND Corporation methodology is its interdisciplinary (Sent, 2007, p. 461). The RAND Corporation favored cross exchanges among mathematics, natural sciences like physics, social sciences like sociology, economics, philosophy and psychology. The experimental methods and the exchange between economics and psychology in this perspective, has been strongly enhanced (see Flood, 1952; Weintraub, 1992; Leonard, 2010; Innocenti, 2005). We will see that the interdisciplinarity and the experimental methods both play an important role in Schelling's theoretic building. Schelling ultimately built a theory of strategy, encompassing many dimensions of situations of interdependent decisions by appealing to an interdisciplinary thinking. He is in particular one of the first to formally claim that contextual and analytical analyses are deeply interconnected.

2. Schelling's reorientation of game theory

2.1. What is the essence of game theory and what are the limitations he identifies in classical game theory

When Schelling grasps the essence of his conception of game theory he asserts in an interview with Jeffrey Brown (2005)

> "The simplest explanation is that game theory is the study of any situation in which two or more people make decisions that impinge on each other. And each individual must take into account what the other is likely to do either simultaneously or in succession."

While lots of game theorists would define game theory as the *mathematical theory* of games, *representing* the choice of rational agents, Schelling merely sees situations involving the interdependence of individuals' choices, i.e. involving a process of reasoning. When making their decision, the individuals – or players – *must* take into account the others' eventual choices. This entails for Schelling a mode of *contingent* behaviors (Schelling, [1978] 2006, p. 17). Players have in consequence to adapt their goals to the others (ibidem). This statement contradicts with the static classical appraisal of game theorists which states

> "Game Theory is the formal study of rational decision in situations [in which] two or more individuals have choices to make, preferences regarding the outcomes, and some knowledge of the choices available to each other and of each other's preferences." (Schelling, 1984, p. 214)

When laying the consequences of players' interdependence of choices and actions in his own definition of game theory, Schelling claims that the main task of game theorist should be the identification of the *mechanisms* through which players could adapt each other and not the imposition, a priori, of the consistency of players' beliefs as it is the case in the mathematical view of games. The purpose is for him to explain how such state of consistency can be reached and the reasons, the determinants, explaining such consistency. The existence of a *determinate solution* for games – which is the purpose of game theory (Sugden, 2001) – can only be ensured by this mechanism of adaptation. He thus attempts to explain "how" and "why" a specific outcome occurs, contrary to standard and epistemic game theory (see chapter 1), i.e. how and why the existence of a solution is guaranteed (see Giocoli 2003). Recall that the "neoclassical" account of equilibrium that comes with the rationality the equilibrium is conceived in a static manner (Giocoli, 2003, p. 138). It corresponds to a steady state, i.e. "a state of no motion" (Weintraub 1991, p. 18). Nothing can explain the process that leads to such steady state situation, i.e. how and why a specific equilibrium occurs (Giocoli, 2003, p. 208); it is only supposed that the plan of every individual composing the society is congruent with each other—everybody has maximized her expected utility—and no one has an interest in changing her plan. This is the underlying premise for the Nash equilibrium and more generally for standard and epistemic game theory. One of the major interests of Schelling's contribution for the thesis is thus this conception of equilibrium that cannot be considered as a hypothesis of the model or of the theory of games, as is standardly assumed in non-cooperative game theory (being classical or epistemic) but that must be explained by the theory, the model.

The solution of games is necessarily dependent on the choices made by each player.

> "The outcome depends on the choices that both of them make, or all of them if there are more than two. There is no independently "best" choice that one can make; it depends on what the others do … Any "solution" of a problem like this is necessarily a solution for both participants." (Schelling, 1984, p. 214)

According to Schelling, in real 'game like' situations, the players are perfectly able to understand the consequence of this interdependence of choice, they are accordingly aware that the best outcome they can get greatly depends on the others so that they have to adjust to each other, even to compromise. The existence of a solution is dependent on this propensity to adapt to each other and to compromise in particular in coordination game (as standardly defined in game theory) in which no solution exists because of the multiplicity of equilibrium).

They have therefore to guess what these others may want, what are their goals, their purposes, etc. Such situations generally involve a high level of *uncertainty*. This is linked to a form of complexity which implies that the individual interactions create a specific order and that this social order in turn influences these same individuals when interacting. This echoes the theory of complex systems largely expanding nowadays.

Schelling ([1960]1980, p. 116) mentions "[t]he uncertainty that can be usually presumed to exist about each other's value systems" in games. In his work, the players' 'value system' is a concept grasping the players' preferences, intentions, beliefs, modes of reasoning, patterns of behaviors, etc. In other words, it captures the players' background and can explain the players' apprehension of the games they face. By implication, players' 'value system' can explain the way they solve the games. We will see in section 2.2 how such value systems, which imply that players follow certain patterns of behavior, orient the solution they anticipate and the beliefs they hold regarding how the other players may behave (also following these patterns of behavior). We will exploit such social dimension in the chapter 5 to propose a theory of game in which a game is considered as an open-system.

Another dimension then appears in Schelling's definition of game theory: the players' expectations.

> "Game theory is the formal study of the rational, consistent expectations that participants can have about each other's choices" (Schelling, 1984, pp. 214-15).

Within this uncertain context that are the strategic games, the mechanism insuring the existence of a solution relies on the convergence of players' expectations. When common knowledge of rationality is not a priori assumed, reaching this state of convergence can be very hazardous; this explains the many determinants that Schelling identifies that intervene in the players' reasoning process towards this state. In Schelling's work, such state of convergence is the main requirement for the existence of a solution. This essential dimension of strategic contexts is very contemporary. It preempts the epistemic turn in game theory long before the problem of players' expectations crystalized the interest of game theorists and led to the new development of contemporary game theory: the epistemic theory of game. Though, as explained in the first chapter, epistemic game theory appears to be very disappointing on this dimension. The analysis of games remained static, the process of the convergence of players' beliefs is never incorporated as players' beliefs are a priori – i.e. their prior beliefs are the beliefs they hold at the equilibrium.[21] Recall that the existence of an equilibrium is already assumed in epistemic game theory as the game describes the choice of the players at the equilibrium. The players' beliefs are thus in that perspective, mere artifact. This required convergence justifies for Schelling the investigation of the interaction process. For him, the way players form their beliefs about each other is *endogenous* to the game, i.e. the players' expectations are generated during the interaction process (see Innocenti, 2007). This insight is even more pioneering, considering that even nowadays none of the epistemic game theorists sustain such endogenous process of the formation of players' beliefs. We will formalize this process in the theory of games proposed in chapter 5 as well as the process of the convergence of players' beliefs. Bacharach's work also provides an example of such endogenization of players' beliefs (see chapter 3).

One of the main arguments for Schelling to appeal for a "reorientation of game theory" is to investigate the conditions under which a specific solution occurs. According to him, game theorists have been much more interested in the mathematical properties of games solutions

---

21 See Larrouy and Lecouteux (2017), Gul (1998), Levi (1998) and Morris (1995).

without investigating the structural elements of strategic interactions that may influence such solutions. The process ensuring the convergence of individual beliefs is one of these structural elements. Schelling accordingly intends to overcome this stalemate. For him, the solution concept is the masterpiece of game theory since it provides "a way of thinking about the problem" (in Swedberg, 1990, p. 189) of strategic interactions and provides a clear interpretation of the result of these interactions. It should however not subvert the search for an explanation of its existence. Investigating the conditions under which the existence of a solution can be guaranteed is the main task of a game theorist for him.

Taking a full account of the process through which a particular solution emerges needs a thorough examination of (i) the players' cognitive process that ultimately lead to a convergence of expectations and (ii) the context of the interaction that is a ground for having a hint on players reasoning, and again, which can ultimately explain how can players' expectations converge. We will rely on those two dimensions in the theory proposed in the chapter 5. We will indeed (i) integrate players' mental states in games, their perceptions and beliefs, and (ii) explain how the context of the game impinge on the players' beliefs and ultimately on the outcome.

Schelling incriminates game theorists for being not interested on those aspects of games. His explanation for that is twofold. This is first and foremost explained by the predominance of interest, in game theory, for zero-sum games, i.e. 'divergent-interest games' – in Schelling's own words – symbolizing a situation of pure conflict among the players. The primacy of the study of zero-sum games, and in particular the minimax theorem, precludes for any consideration on the need for convergence of players' expectations. .

> "In the theory of games of pure conflict (zero-sum game) randomized strategies play a central role. It may be no exaggeration to say that the potentialities of randomized behavior account for most of the interest in game theory during the past one and one-half decades. The essence of randomization in a two-person zero-sum game is to preclude the adversary's gaining intelligence about one's own mode of play – to prevent his deductive anticipation of how one may make up one's own mind, and to protect oneself from telltale regularities of behavior that an adversary might discern or form inadvertent bias in one's choice that an adversary might anticipate." (ibid, p. 175)

Schelling's criticizes the minimax theorem – involving strategies of randomization – which relies on the reduction of players' decisions in strategic contexts to individual decision-making. It is a mathematical artifact to avoid the epistemological consequences of the interdependence of players' choices and behaviors. With the minimax theorem, no need to account for the endogenous formation of players' beliefs, no need for players to adapt to each other. On the contrary, Schelling asserts that "a 'game-like' situation can be viewed as a collective decision-process – a process by which two or more individuals jointly decide on an outcome" (Schelling, 1984, p. 236).

> "In a zero-sum game the analyst is really dealing with only a single center of consciousness, a single source of decision. True, there are two players, each with his own consciousness; but minimax strategy converts the situation into one involving two essentially unilateral decisions." (Schelling, 1980[1960], p. 163)

Even in situations with apparently pure conflict of interest Schelling sees the need for the coordination of expectations. Players in this type of situations would be better off if they find an agreement[22], i.e. if they recognize that in this situation they have a common interest in the avoidance of such conflict. They have to coordinate on a profitable issue for each of them. The example of the games of war is straightforward and is indeed the example chosen by Schelling to highlight how profitable it is for the parties involved to avoid a destructive conflict and instead to recognize that they have an aligned interest in finding a solution that would be better for each of them (ibid, p. 100).

Schelling therefore denounces the misrepresentation of the fundaments of zero-sum games. Because they do not involve a pure conflict of interests and because they essentially involve a need for coordination, Schelling proposes to re-label the zero sum games in "mixed-motive games". This term can grasp the essence of these 'game-like' situations, i.e. the nature of a strategic interaction: the interdependence of the players.

> "These are "games" in which, though the element of conflict provides the dramatic interest, mutual dependence is part of the logical structure and demands some kind of collaboration or mutual accommodation – tacit, if not explicit – even if only in the avoidance of mutual disaster." (ibid, 83)

Indeed Schelling claims that in the mixed motive game like in every game-like situation,

> "two or more centers of consciousness are dependent on each other in an essential way. Something has to be communicated; at least some spark of recognition must pass between the players. There is generally a necessity for some social activity, however rudimentary or tacit it may be; and both players are dependent to some degree on the success of their social perception and interaction. Even two completely isolated individuals, who play with each other in absolute silence and without knowing each other's identity, must tacitly reach some meeting of minds."

Understanding how such meeting of minds can be reached necessitates the incursion into psychological theories, into cognitive and social psychology as suggested in this quotation.

The prevalence of zero-sum game is particularly harmful for game theory since it also leads to the insufficient interest for non zero-sum games while they bear the fundamental determinants of the mechanism insuring players' coordination. Schelling is well known for his study of coordination games, i.e. games showing an alignment of players' interest.[23] This focus of attention is in his own words an attempt "to identify that certain qualities of the mixed game that appear most clearly in the limiting case of pure coordination" (ibid, p. 90). The application of the methods and the concepts developed for pure conflict games is responsible for the shortcoming that classical

---

[22] At least a *tacit* agreement. Such tacit agreement is nothing more in Schelling's account than the coordination of players' expectations.

[23] Schelling is responsible for the now widely use label of 'coordination games': "coordination game seems a good name for the perfect sharing of interests." (ibid, p. 89)

game theory faces (Schelling, 1980[1960], p. 83). This leads to the under determination of a theory of strategy, i.e. a theory of interdependent decisions.

> "It is to be stressed that the pure-coordination game is a game of strategy in the strict technical sense. It is a behavior situation in which each player's best choice of action depends on the action he expects the other to take, which he knows depends, in turn, on the other's expectations of his own. This interdependence of expectations is precisely what distinguishes a game of strategy from a game of chance or a game of skill. In the pure coordination game the interests are convergent; in the pure-conflict game the interest are divergent; but in neither case can a choice of action be made wisely without regard to the dependence of the outcome on the mutual expectations of the players." (ibid, p. 86)

Therefore, as part of his reorientation of game theory, Schelling attempts to get rid of this standard dichotomy between divergent-interest games and coordination games. He attempts to rethink the common determinants involved in each game whatever the importance of the potential conflicting dimension. Divergent interest games and pure coordination are the two *abstract* opposite poles of the theory of games. However, true is that in reality a game is more likely to be within these two poles. The traditional categories of games should not be distinguished. Coordination games (i.e., non-zero-sum games), and divergent interest games (i.e., zero-sum games) are all games of common interest. All of these types of games require "some kind of collaboration or mutual accommodation" (ibid, p. 83). It means that all of these games are about coordination, i.e. are about the identification of the process of convergence of players' perceptions, intentions, beliefs and choices. The prevalent opinion within the game theorists' community is to the contrary that coordination games are at the opposite extreme of divergent interest games. Basically, if a player is rational in the sense of the EUT, he does not cooperate or coordinate with the other(s); players do not adapt each other. By acting like this, she does not maximize her vN/M expected utility. It implies that a game is a static representation of a choice but not an interaction process. The ontology of a game is completely different in Schelling's approach compared to the standard one, i.e. both classical and epistemic game theory. Understanding how and why a game can be solved in Schelling's view requires to understand what strategic interdependence involves and thus what coordination involves.

A last dimension justifying Schelling's proposition for reorienting game theory is that it remained confined to the universe of mathematics. For Schelling, that specificity leads game theorists to be insufficiently interested in the non-mathematical features of game situations, although they are valuable for players to solve a game. Schelling's criticism of the minimax theorem and the interest for zero-sum game is not distinct from the prevalence of mathematical concerns in game theory. This is indeed a consequence of it for him. He postulates that "a randomized strategy … is a means of expunging from the game all details except the mathematical structure of the payoff, and from the players all communicative relations" (ibid, p. 105). He therefore claims "nonzero-sum game theory may have missed its most promising field by being pitched at too abstract a level of analysis." (ibid, p. 119): the mathematical level. The process leading to the resolution of games, i.e. the process sustaining the convergence of players' expectation is inherently *empirical* and more specifically *contextual* (see Zeckhauser, 1989; Ayson, 2004; Aydinonat, 2005; Colman, 2006; Dixit, 2006; Sugden and Zamarrón, 2006; Innocenti, 2005, 2007; Rivzi, 2007). Such

contextual dimension relies on players' perceptions, social backgrounds, and so on, i.e. on multifactorial elements that refer to psychology, sociology, history, law, and other social sciences. Schelling's main criticism against classical game theory is to abstract games from all empirical facts and from all contextual information, even though they are highly relevant and valuable in the resolution process of games. All of these determinants impinge on the player's capacity of coordination and on their reasoning process towards the states of consistency characterizing the existence of a solution. The identification of the extent to which all of these determinants intervene in the player's reasoning process is of particular importance, if, as it is nowadays more and more acknowledged, the imposition of common knowledge (or common belief) of rationality is insufficient to lead to a solution or is methodologically problematic.

Besides, it presupposes that players can act according to the same rules of play as the game theorists, i.e. as mathematicians, and can understand games' mathematical structure. As he declares

> "We must avoid assuming that everything the analyst can perceive is perceived by the participants in a game, or that whatever exerts power of suggestion on the analyst does so on the participant in a game. In particular, game characteristics that are relevant to sophisticated mathematical solutions (except when that same solution can also be reached by an alternative, less sophisticated route) might not have this power of focusing expectations and influencing the outcome. They might have it only if the players perceived each other to be mathematicians. This may be the empirical interpretation of such "solutions" as those of Braithwaite, Nash, Harsanyi, and others. It is that the mathematical properties of a game, like the aesthetic properties, the historical properties, the legal and moral properties, and all the other suggestive and connotative details, can serve to focus the expectations of certain participants on certain solutions. If two players are themselves mathematical game theorists, they may mutually perceive and be powerfully affected by potential solutions that have compelling mathematical properties. Each may transcend, and know that the other will transcend, various adventitious details that, to non-mathematician game players, might be more relevant to the focusing of expectations than some of the quantitative properties of the game." (Schelling, [1960]1980, pp. 113-114).

Such presupposition has for consequence the abstraction of games for any non-mathematical dimension. However Schelling claims that "the principle relevant to successful play, the strategic principles, the propositions of a normative theory, cannot be derived by purely mathematical means from a priori considerations." (ibid, p. 163). Again as emphasized in the previous quotation Schelling clearly argues that many dimensions outside the realm of mathematics: legal, social, economic, historical, and so on, can help players to identify one solution and therefore induce the convergence of their beliefs. This refers to an 'open-system' theorizing (Chick, 2004, p. 8): a mode of theorizing that does not only rely on mathematical concepts, that admits interdisciplinarity, that admits the limit of comprehension and apprehension of the real world in modeling and that acknowledges that a theory must be evolutive and adaptable.

This implies in Schelling's view that, contrary to the classical premise, the game matrices are not the only necessary devices for players to come to a solution. He asserts that even when players

can only count on the information contained within the game matrices, some non-mathematical aspects like some perceptions or some power of suggestion conveyed by such matrices can lead the players to expect one potential outcome or at least to guess the more likely to be realized. Such aspect of strategic reasoning in games when it is acknowledged that the way players appraise their decision problem is manifold, i.e. when games are open-systems and no longer self-contained matrices, is an essential element of the theory proposed in the chapter 5. Every matrix carries some suggestive and perceptual details that can help player in their decision process. The particular labeling of strategies within the matrices is one of such suggestive details, and this is clearly the type of information that may be considered as relevant (ibid, pp. 95-96). For Schelling, strategies labeling can provide reliable elements for decision-making. A purely mathematical game theory however cannot account for that. Yet, a payoff matrix should simply be considered as "an analytical device" (ibid, p. 95), i.e. a translation by the theorist of a real decision problem, in his own language. The presentation of the way we see a game in the chapter 5 will echo this statement.

That is why for Schelling:

> "the mathematical structure of the payoff function should not be permitted to dominated the analysis … there is danger in too much abstractness: we change the character of the game when we drastically alter the amount of contextual detail that it contains or when we eliminate such complicating factors as the players' uncertainties about each other's value systems. It is often contextual details that can guide the players to the discovery of a stable or, at least, mutually non destructive outcome." (ibid, p. 163)

Basically, everything is a source of information for the players to solve a game. By expunging games from all kinds of non-mathematical aspects, game theorists may deprive the players of their capacity to solve games (Schelling, 1980 [1960], pp. 95-96). Enhancing the predictive power of game theory require to identify all of the potential relevant elements on which players can count to be successful. Rational players should use every parcel of information allowing them to be better off and to reach a higher payoff. The particular labels of strategies is one of these potential elements while the mathematical label may not at all be useful in particular when considering players that are not familiar either with the classical game theoretic fundaments or its mathematical and technical aspects.

All of the critics mentioned above constitute what Schelling identifies as the main limitations of classical game theory. His reorientation is an attempt to overcome those limitations. He suggests a new epistemology of game theory. His contribution is analytical and offers a new methodology: interdisciplinarity.

2.2. His reorientation of GT

What Schelling offers in his reorientation of game theory is

> "[an] attempt to enlarge the scope of game theory, taking the zero-sum game to be a limiting case rather than a point of departure. The proposed extension of the theory will be mainly along two lines. One is to identify the perceptual and suggestive elements in the formation of mutually consistent expectations. The other … is to identify some of the basic "moves" that may occur in actual games of strategy, and the structural elements that the moves depend on; it involves such concepts as "threat", "enforcement", and the capacity to communicate or to destroy communication." (ibid, pp. 83-84)

This necessitates, contrary to classical game theory which is according to Schelling (1984, p. 239) "concerned with *outcomes*, not intermediate processes", acknowledging that it is no longer the outcome of game *per se* that is the focus of attention but the process leading to such outcome.[24] The purpose of Schelling is accordingly to identify (i) the suggestive and (ii) the structural elements on which players may rely during the interaction process. The former more specifically rely on individual perceptions, while the latter more specifically rely on the context of the game. Those suggestive and structural elements thus rely, on the one hand, on the way players interact, i.e. on the way their respond each other through communicative behaviors (individual behaviors convey information for Schelling) and, on the other hand, within this interactive process, on the identification of "clues", "hints", and so on that orient players on a particular solution over the other potential ones. Those clues are given either by the player themselves and their behavior or by the context of the games. Those clues are constitutive of what Schelling calls a *focal point*. Since the starting point of his thinking is that in each – *real* – game-like situation there is a wide range of possible outcomes (ibid, p. 101), Schelling apprehends the concept of solution within his framework with the concept of focal point. It translates the empirical fact that there is no *a priori* determined equilibrium and that the outcome that will effectively occur is determined during the interaction process. For Schelling each game in its overall dimension (i.e. the context and the interaction process) provides some clues or signals helping players to find a focal point, i.e. to find a mean to solve the game (e.g. see Schelling, 1980[1960], p. 108). We will detail and explain these elements in the next subsections.

2.2.1. The solution concept: the focal point

In Schelling's framework, the solution of a game is informally and formally characterized by *stabilized convergent expectations* (ibid, pp. 114-115). This is in his own words an "empirical fact" (ibid, p. 114). Analytically this requires integrating the players' beliefs in the realm of game theory. This description of the games' solution in terms of convergent expectations, i.e. in terms of common beliefs, is characteristic of the epistemic game theory. However, it is interesting to note that none of the propositions that Schelling makes concerning the uncertainty prevailing in strategic context and the impact of the incompleteness of information in games has been cited

[24] See also Walliser (2004, p. 183) who claims that in classical economic theory the "substantive individual rationality" entails "no mental deliberation process leading to the chosen action." He however considers that epistemic game theory is an attempt to overcome such insufficiency.

when the introduction of incomplete information in the realm of game theory has been settled. None of the epistemic game theorists mention Schelling's early work on the epistemic conditions sustaining the solution concept.

> "The outcome is determined by the expectations that each player forms of how the other will play, where each of them knows that their expectations are substantially reciprocal." (ibid, p. 107).

As already emphasized, the interest that Schelling shows concerning the coordination games, as he states, is that they serve as a paradigmatic case of the psychic phenomenon of the convergence of expectations (ibid, p. 65). They help to define a new solution concept compatible with every type of games.

> "The pure common-interest game, or coordination game, may add insight into the reasoning behind certain solution concepts in game theory, particularly that of solution in the strict sense for the "non-cooperative" game." (ibid, p. 291)

In coordination games, a solution exists as soon as each player "does exactly what the other expects him to, knowing that the other is similarly trying to do exactly what is expected of him." (ibid, p. 100) It means that a coordination game is solved when there is between players' a "meeting of minds", i.e. when players "read the same message in the common situation" (ibid, p. 54). The same mechanism prevails for every type of games: the existence of a solution for mixed-motive games is ensured when there is among the players a meeting of minds (ibid, p. 114). This will be later identified as a state of common reasoning.

To reach this meeting of minds, "one line of actions" must serves as a *coordinator* (ibid, p. 65). One course of actions must induces the convergence of players' expectations (ibid, p. 54). This is the essence of coordination and of strategic thinking in Schelling's work. When there is a meeting of mind between the players, their expectation converge and a solution emerges. The existence of a solution is ensured only if such meeting of mind exists, i.e. if players' beliefs are consistent with each other's. It means that the players' perceptions, intentions and plans of actions must be consistent with each other and congruent. This dimension will be of particular importance in the chapter 5 of the thesis in the model of game developed. Many components can provide such device, i.e. can have the power to focus players' expectations. Even the standard mathematical solution can have this power, but it is only one possible coordinator and it supposes at least some minimal knowledge either in game theory or in mathematics (ibid, p. 114). However, as already exposed, limiting this power of focusing expectations to mathematical solutions restrains the possibility of the existence of a solution to very specific players and it therefore becomes very restrictive as a solution concept.

Indeed, for Schelling, "any solution is "correct" if enough people think so" (ibid, p. 55), accordingly, any coordinator can be used by the players to finally agree – tacitly – on a common solution. Therefore, as Schelling argues

> "The problem is to find some signal or clue or rationalization that both can perceive as the "right one", with each party prepared to be disciplined by that signal or clue in the event that it appears to discriminate against him." (ibid, p. 100)

This signal must be recognized by every player as the one inducing one common line of actions and therefore as having the power of inducing the convergence of players' expectations. It corresponds to a point in time about which "everyone expects everyone to expect everyone to expect observance, so that non-observance carries the pain of conspicuousness." (ibid, p. 91). This signal is for Schelling a focal point and it induces common reasoning (Hédoin, 2014, 2016).

Focal points arise in situations in which "two or more parties have identical interests and face the problem of … coordinating their actions for their mutual benefit, when communication is impossible." (ibid, p. 54). Focal points thus arise indifferently in coordination games and divergent interest games. Focal points have for consequence that players ""mutually recognize" some unique signal that coordinates their expectations of each other." (ibid, p. 54)

A focal point is a solution that is *qualitatively* different from all of the other potential solutions.

> "If then we ask what it is that can bring [players'] expectations into convergence … we might propose that it is the intrinsic magnetism of particular outcome, especially those that enjoy prominence, uniqueness, simplicity, precedent, or some rationale that makes them qualitatively differentiable from the continuum of possible alternatives." (Schelling, 1980[1960], p. 70)

A focal point is a solution that necessarily has a discriminatory power (ibid, p. 300). Such power is ensured only if some *qualitative* characteristics like 'prominence', 'uniqueness' or 'simplicity' allow differentiating one outcome as being a focal point, from all of the other potential outcomes. The task of the players is therefore, during the interaction process, to identify such focal point and to be prepared to be committed to the 'line of action' induced by this focal point. This aspect of Schelling's work will be formalized in the chapter 5 through what will be called strategic framing. More specifically players are aware that what is collectively focal may be different than what they individually perceive as focal; they acknowledge that to coordinate they must follow what is collectively focal and possibly adopt a behavior different from what they would have adopted without considering the collective.[25] This statement is highly important since it means that there is no way to define axiomatically a solution concept or an equilibrium point (ibid, p. 163; see also Innocenti, 2007, p. 418). It is not possible to define a priori the solution of a particular game, nor to determine the properties possessed by this solution. Players indeed discover this solution during their interaction process. For Schelling, the existence of a solution is a matter of fact. According to him, in the real world, individuals interact and coordinate everyday. In the real world the context of the interactions provides an abounding informational basis. If individuals are perfectly able to coordinate it is because some lines of action become obvious in a specific context which enable them to successfully reach a meeting of minds. However, such solution is unlikely to be optimal or to be characterized by joint maximization. In mixed-motive games, players are ready to be committed to the focal point they mutually identify in order to avoid a situation in which there is a default of coordination and therefore no – tacit – agreement, which would be prejudicial for each of them (ibid, p. 100). They are prepared to conform to the line of action conditioned by the focal point even in cases it discriminates them.

---

[25] See Orléan (2004).

But, for Schelling, the fact is that in reality players are particularly inclined to follow the line of actions they associate to the focal point. For instance, in a bargaining game, if an agreement like fifty-fifty is reached, it must be a focal point, i.e. a rule of sharing on which players converge.

The very nature of strategic interactions has for consequence that every player engaged in such interaction knows that everybody is searching this focal point. Everybody attempts to find the focal point that will ultimately induce a meeting of minds, i.e. the existence of a solution.

> "People can often concert their intentions or expectations with others if each knows that the other is trying to do the same. Most situations – perhaps every situation for people who are practiced at this kind of game – provide some clue for coordinating behavior, some focal point for each person's expectation of what other expects him to expect to be expected to do." (ibid, p. 57)

Finding the focal point is "the key" to solve the game (ibidem). The players can perceive these clues – i.e. the focal points – only if they pay attention to the context of their interactions. In this context, players acknowledge that they share a common capacity, to find and to understand, these clues. Schelling argues

> "A prime characteristic of most of these "solutions" to the problems, that is, of the clues or coordinators or focal points, is some kind of prominence or conspicuousness. But it is a prominence that depend on time and place and who the people are." (ibid, pp. 57-58).

If players are embedded in different contexts, and face an identical problem of coordination, the way they solve this problem and the solution will be different. In a comparative way, different players facing a common coordination problem in a common context will potentially identify different focal points, and therefore come to different solutions.

In fact, the focal point can rely on "exogenous" devices such as conventions, institutions, regular pattern of behaviors, "precedents," and so on, if players perceive it as a reliable device of coordination in the conditions they face. It means that some behavior that previously led to successful coordination serve as device for a meeting of minds: players believe that everybody will conform to the same successful behavior and that everybody believe so, so that players' beliefs are congruent. Or a focal point can be built endogenously through the players' interactions and through communicative behavior. When there is no reliable pre-existent focal point, players determine, together, during their interaction process one prominent solution.[26] One pattern of behavior orienting toward the convergence of players' subjective beliefs may emerge through their interactive actions.

This may explain why Schelling asserts

> "Sometimes the focal point itself is inherently unstable. In that case it serves not as an outcome but as a sign of where to look for the outcome." (ibid, p. 112)

---

[26] *Emergence* is a concept leading to numerous investigations in the philosophy of sciences which again shows how Schelling's work strongly relies on interdisciplinarity.

This instability partly relies on the fact that the focal point is extremely sensitive to the dynamic of interactions and in the process, some potential outcomes may appears as well as disappears; these appearance and disappearance depend on the specificity of the players' interaction during the game and the context of the game. The focal points are extremely sensitive to the potential clues of the games and their reliability at the moment players make decisions (ibid, p. 111). In this perspective a focal point can never be determined a priori and given exogenously – i.e. determined before the game independently of the way the game is played – but always assessed in a specific context, in a specific interactional decision problem when players are confronted to the choice they have to make. These clues depend on suggestive details which hinge on players' perceptions which are both context-dependent – and therefore specific to the interaction process and the context of play – and socio-culturally determined. The prominent character of a focal point may evolve during this interaction process. It depends on the elements that players consider as reliable when making a decision and on the way players react each other. As Schelling (2006[1978], p. 26) emphasizes, there may be "shifts in the parameters that determine the equilibrium." This explains why Schelling dedicates a lot of space in *The Strategy of Conflict* to describe the elements that have to be identified during this process to provide a theory of strategy, and why the resolution of games is of such importance in his reorientation of game theory. The evolutionary dimension of focal point that Schelling sketches shows that he does not see the eductive, i.e. epistemic side of game theory incompatible with its evolutionary, i.e. historical side. This confirms again how his account of game theory is deeply interdisciplinary and how he resorts to an open-system theorizing (Chick, 2004).

2.2.2. The resolution process of games

The starting point of Schelling's framework is that it is the *resolution process* of a game that ultimately leads to a solution. Again this contradicts with the standard conception of games in both classical and epistemic game theory in which the game is a representation of a choice, and in which therefore, the existence of a solution is assumed (see the chapter 1 and the introduction). As already emphasized this is an empirical fact (ibid, p. 71). For that purpose,

> "The players must jointly discover and mutually acquiesce in an outcome or in a mode of play that makes the outcome determinate. They must together find "rules of the game" or together "suffer the consequences." (ibid, p. 107)

Recall that a solution is determined when players "read the same message in the common situation", when they "identify the one course of action that their expectations of each other can converge on", i.e. when they mutually recognize one focal point which serves as a "unique signal that coordinates their expectations of each other" (Schelling, 1980[1960], p. 54). To reach such state of convergence players have ultimately to have aligned interests, aligned purposes, and accordingly perceive the same thing from the same situation. This refers to common understanding and then common or symmetric reasoning: "common understanding . . . defines the set of events that have the properties to generate a common knowledge set of consistent expectations" (Hédoin 2014, p. 385). Common reasoning induces common inductive inferences (ibidem). Hence, the resolution process of games is a search for shared rules leading to this final

tacit agreement. This is particularly demanding for players who may have a great uncertainty on who the other is and on what she wants. In this perspective, players may use whatever coordinators may help them to finally reach such state of agreement.

Reaching such state of convergence of perceptions, minds or expectations, when no communication is allowed and no enforcement system exists is for Schelling the result of a multifaceted process that can be real (when players' moves are sequential) or psychic (for one-shot games or simultaneous moves). In other words the outcome of the game is dependent on the multiple elements conveying information when players interact.

In that perspective, the players can rely on two dimensions (i) the way the other behave, and the way everybody react to each other, and (ii) the context of the game. In Schelling's work, this context is not only a conceptual tool but has a semantic content as well. This context is actually twofold. Schelling distinguishes the physical environment surrounding a game from the social and cultural environment. And both can matter in the resolution process of game.

In the case of the physical environment, "some of the objective details of the situation can exercise a controlling influence" (ibid, p. 71). We can mention his example of paratroopers (1980[1960], pp. 54-55). Paratroopers are dropped in an unknown territory in which one characteristic emerges: a bridge crossing a river. In the absence of any other possible – seemingly reliable – meeting point, the paratroopers will converge toward the bridge. This bridge, in Schelling's words, enjoys 'prominence' and is therefore 'conspicuous'. Thus, in this specific case, the bridge is the "objective detail" which allows coordination. In the case of the social and cultural environment, players exploit their experience of social interactions or their social background. Players' social/cultural backgrounds let them define what kind of coordination device should be use to insure convergent expectations. For instance, players can rely on "analogies," "precedents," "incidents" (ibid, p. 90), "clichés," "conventions" (ibid, pp. 84-85), "institutions," "traditions" (ibid, p. 91), etc.; basically, everything that can be perceived and interpreted as a successful, and accordingly, as a stable coordination device (ibidem). Players must think that these devices are reliable in the sense that everybody believes that everybody else can conform to it, etc. (ibid, p. 91). For that purpose, players must share a common background, i.e. players must ultimately know that these coordination devices belong to the value system of the other players (Hédoin, 2014, 2016).

It means that in Schelling's work salience and focal points can be understood either in a naturalistic way or in a social – i.e. community-based – way (Hédoin, 2014, 2016). In the former case salience is linked to "an objective and natural property of some entities (events, strategies, outcomes)" (Hédoin, 2014, p. 366). This is for instance the case of the geographical elements, of the "precedent" that Schelling mentions or of historical elements: "salience is essentially natural as far as it derives from cognitive mechanisms which are themselves the ultimate product of our evolutionary history" (Hédoin, 2016, p. 3). In the latter case, "community membership is the basis for salience" (Hédoin, 2014, p. 385). Salience becomes explained by the history of a specific community (Hédoin, 2016, p. 3; Orléan, 2004). Salience is therefore linked to what Schelling

identifies as conventions, traditions and institutions, which form the identity of a community (see Orléan 2004).[27] A recurring practice, or behavior, that has led in the past to successful coordination has indeed every chance to work in the future; players can be confident about the fact that such practice or such behavior can make consensus. They become obvious and therefore salient. This generates common expectations towards the other behavior and a symmetric reasoning. In a similar way, regarding conventions and institutions, in a given community, players will interpret such "social facts": a "shared institutional heritage . . . brings common understanding into a population" (Hédoin, 2014, p. 383). When belonging to a common community, players "have confidence in the fact that they interpret the same institutional fact in a similar fashion" (ibidem). This is what Hédoin (2016, p. 4) calls "community-based salience" and which is induced by "community-based reasoning." Community-based reasoning entails that the players believe that they are all members of the same community and it ultimately leads to a common belief among the players that everybody will follow the same pattern of behavior (ibid, p. 16).

This explains why Schelling claims that "[a]n important characteristic of any game is how much each side knows about the other's value system" (ibid, p. 139). Why the knowledge about each other's value system is of such importance for Schelling? Because such knowledge gives them some hints on some patterns of behavior that their co-players can follow, on their eventual purpose, and ultimately on the way they will be able, *together*, to reach a tacit agreement. It gives information to the players about the way they may accommodate each other in this perspective. It involves the same principle as in explicit bargaining games except that in this case such bargain process is implicit: it relies on the players' capacity to make compromise through behavior and not language. A value system entails patterns of behavior, i.e. behavioral routines or specific practices, that orient towards some specific expectations as these routines imply some way of doing and thinking**.** It helps the players to reach a tacit agreement: it induces the convergence of their beliefs. Patterns of behavior as rules of behavior are sustained by a value system and therefore entail the knowledge of the specific condition and intention triggering such behavior. This can therefore entail specific expectation concerning the reasons that are behind such behavior and the intention concerning the issue of such behavior. This social dimension that impinges on players' thinking is hardly compatible with a mathematical conception of games in which games have been expunged from all of these details to become so abstract that the only relevant information is the player's payoff function or expected utility. This means, as defended in this thesis and as demonstrated in the chapter 1 in particular, that a restrictive conception of game theory solely based on the mathematical conception of games, exhibits considerable limits (with respect to the indeterminacy problem for instance) and methodological inconsistencies (the double imposition of common belief of rationality and common priors for instance) which can be bypassed by a change of epistemology and of ontology with respect to the apprehension of games. Games must be considered as interaction processes and not only as mathematical representations of individual rational choices. A game must be the investigation of how and why a specific solution occurs. In that perspective, the thesis appeals for a change in the ontology of

[27] For Orléan (2004, p. 200) the identity of a community is shaped by its "cultural and historical context."

games. Again, the solution must be a result of a game, of a reasoning process and not a hypothesis of the model of game.

For instance, individuals' social position – or 'social role' – belongs to their value system. These social positions underlie *patterns of behavior* which in turn may orient toward convergent expectations (ibid, p. 92).

> "The concept of role in sociology, which explicitly involves the expectations that others have about one's behavior, as well as one's expectations about how others will behave toward him, can in part be interpreted in terms of the stability of "convergent expectations", of the same type that are involved in the coordination game. One is trapped in a particular role, or by another's role, because it is the only role that in the circumstances can be identified by a process of tacit consent." (ibidem)

Therefore, the way players are responding to each other provides 'clues' in the determination of a solution. Schelling argues that "moves can reveal information about a player's value system or about the choices of action available to him; moves can commit him to certain actions when speech often cannot" (ibid, p. 102). Players have to pay attention to the others' behaviors. These behaviors reveal the players' underlying value system and in turn their eventual perceptions and intentions. Behaving is a way to communicate intentions, to give some hints about a purpose. This process of apprehension of others' mental states will be detailed in the chapter 4 of the thesis. The mindreading capacity and the direct social perceptions at play in individual interactions will clarify this mechanism of attribution of mental states to others, as well as the extent of the intersubjective knowledge that individuals can acquire when interacting. This is of particular importance because of the uncertainty surrounding who the other is and in particular concerning her value system which, again, is one of the basis of her eventual perceptions, intentions, purposes, goals, preferences. And for Schelling, behavior can be more communicative than language and commit players more than cheap talk for instance. That is why the way players respond to each other through communicative behavior is of such importance: "if we deny them any form of speech, they must convey their intentions and their proposals by their patterns of behavior. Each must be alert to what the other is expressing in his maneuvers, and each must be inventive enough to convey his intentions when he wants them conveyed." (ibid, p. 104) A behavior can influence the others' perceptions; it can reveal, a commitment, or a threat, and therefore can orient the others' choices.

> "They must find ways of regulating their behavior, communicating their intentions, letting themselves be led to some meeting of minds, tacit or explicit, to avoid mutual destruction of potential gains. The "incidental details" may facilitate the players' discovery of expressive behavior patterns; and the extent to which the symbolic contents of the game – the suggestions and connotations – suggest compromises, limits, and regulations should be expected to make difference. It should, because it can be a help to both players not to limit themselves to the abstract structure of the game in their search for stable, mutually nondestructive, recognizable patterns of movement. The fundamental psychic and intellectual process is that of participating in the creation of traditions; and the ingredients out of which traditions can be created, or the materials in which potential traditions can

be perceived and jointly recognized, are not at all coincident with the mathematical contents of the game." (ibid, pp. 106-107)

However, when no reliable pattern of behavior emerges by only observing the way players interact and behave, again, the context of the game can provide a reliable source of coordination devices. And in this case it can be either the physical context or the social context as emphasized in the quotation above. This physical or social or cultural environment provides conventional way of interacting and coordinating. Players can refer to the knowledge of which arrangements, conventions, rules, etc. are followed in a specific cultural or social environment. Such arrangements entail convergent expectations toward the kind of behavior expected with respect to such rules. Schelling is one of the first in game theory to claim that conventions or norms or institutions create focal points. Contemporary game theorists rarely give Schelling credit for this.

The resolution of a game, like its outcome is specific to the set of players involved in the game, the time and the place. The solution is extremely sensitive to the players' subjective perceptions at the moment they have to make a decision.

> "The "obvious" outcome depends greatly on how the problem is formulated, on what analogies or precedents the definition of the bargaining issues calls to mind, on the kinds of data be available to bear on the question in dispute." (ibid, p. 69).

This is accordingly extremely difficult to define axiomatically – like the solution concept and the equilibrium of games – a rational play in his account of strategic interactions (ibid, p.163). Recall that solving a game is a discovering process.

> "Taking a hint is fundamentally different from deciphering a formal communication or solving a mathematical problem; it involves discovering a message that has been planted within a context by someone who thinks he shares with the recipient certain impressions or associations. One cannot, without empirical evidence, deduce what understandings can be perceived" (ibid, pp. 164-165).

2.3. The social ontology behind Schelling's theory of strategy

Throughout the presentation of Schelling's reorientation of game theory it appears that he resorts to concepts – asides that of interdependence – like stability (i.e. stable pattern of behaviors), traditions, emergence, value systems, social role, and so on, which are some of the fundaments of social reality. Some of the underpinnings of his work are topics that clearly refer to social ontology. Basically, social ontology deals with the mode of being of society, its structure and organization. Schelling does not aim to investigate, explain and define what are social facts, or what are the mode of existence of society, institutions and collectives. His approach does not belong to social ontology although his conception of focal point challenges the usual closed-system account of games and game theory and instead refers to an open-system. Nevertheless, he makes very strong claims about these social objects and their mode of existence**.** By describing reality and the social world as open-systems Schelling lays the foundations of a theory of

interdependence as an open-system and a methodology congruent with it.[28] It means in his approach that (i) understanding interdependence relies on the acknowledgement that individuals when interacting create a social order that in turn impinges on their decisions and on the way they interact, and (ii) a theory of interdependence requires a strong form interdisciplinarity.

Schelling makes statements both on the creation and the mode of existence of "institutional facts" (Searle, 1995, 1998, 2010) through the concept of focal points, which are voluntarily shaped by individuals through their strategic interactions.

From focal points "traditions" emerge (Schelling, 1980 [1960], pp. 106-107). This implies that traditions, i.e. institutional facts, are emergent 'objects'. They are emergent objects in the sense that they do not correspond to any individual characteristics (there is no aggregation mechanism in Schelling's work and the solution of games do not correspond to joint maximization). They emanate from the process of interaction, i.e. from the transitory collective entity composed by the individuals who manage to coordinate. This phenomenon is mainly explained by individuals' need to coordinate and to conciliate their subjective and potentially divergent will (in the case of mixed-motive games). In this perspective, Schelling defines institutional facts – and then institutions – as conventions allowing individuals to coordinate. They correspond to human artifacts. Humans first intentionally create these institutional facts in order to compensate either the uncertainty surrounding coordination (because of their heterogeneity, because of their different value system, etc.) or the eventually divergent interests among individuals or collectives.

In Schelling's account of strategic interdependency, institutions induce convergent and stable patterns of behaviors. Accordingly, they are associated with *rules of behaviors* which point toward consistent and convergent expectations (see Hédoin, 2017, p. 49). It means that individuals ascribe 'constitutive rules' to institutions in order for these institutions to serve as coordination devices (Searle, 1998, 2010). Rule following is defined by Hédoin as

> "a behavioral event that finds its roots in expectations that are constitutive of a practice. More exactly, the persons' behavior corresponds to a practice that entails the knowledge or the belief that some rule holds and where the rule's meaning is defined by the very practice it is constitutive of." (ibid, p. 50)

When players follow a rule, they know what they should do it in the context in which the rule prevails and know or at least believe that every other individual will also reason in the same way. They infer the other's behavior from the knowledge that the rule prevails. "In the case of rule-following, the rule indicates something to be done to everyone, and there is a common reason to believe so because everyone reasons the same way on the basis of some mutually accessible event" (ibid, pp. 62–3).

In outline, referring to Searle (1998, 2005, 2010), it implies that institutional facts are ontologically subjective and epistemologically objective. Focal points, as institutional facts, are first ontologically subjective because they depend on players' subjective perceptions. Focal points

---

[28] See Chick and Dow (2005), Chick (2004) and Mearman (2003).

are specific to a particular set of individuals, a particular time and a particular place. This is the first step or the creation *per se* of these institutional facts. But then, after a certain amount of time these institutional facts progressively acquire an autonomous status. They become anchored in individuals' social background – or social knowledge.[29] They are collectively accepted as devices of coordination as soon as they enter in individuals' social background. This is in this manner that they lead to the convergence of individual subjective expectations. For Schelling, institutions (and their associated patterns of behavior) exist as soon as they are perceived and understood in a common way even by different individuals; i.e. even by heterogeneous individuals who do not have the same value system but belong to a common community in which exist, or which is constituted by, such institutions (see Orléan, 2004). Institutions are therefore objective entities in the sense that they may be taken for granted by these individuals. They become objective facts taken abruptly as such, without the need of interpretation (see Hédoin, 2014, 2016).

In some readings of Schelling's conception of focal points, it becomes an institutional fact if it recurrently serves as a successful coordination device. This explains why he asserts that focal points have a content that is inherently dynamic or evolutionary (see Schelling, 1980[1960], pp. 111-12). The persistence of an institutional fact is dependent upon individuals' acceptance. Subsequently, Schelling's conception of emergence is diachronic. In order for institutional facts to act in turn on individuals' behavior by orienting toward some pre-determined patterns of behavior, they have to be anchored in individuals' knowledge of the functioning of society. As Sugden and Zamarrón (2006) emphasize, focal points, as institutional facts, are conceived in a pragmatic way. In other words, they have to recurrently allow successful coordination. These institutional facts are therefore reproduced through interactions by their practical dimension. In such reading salience is made by precedents. This is accordingly similar to Rawls's notion of "rules as summary view" and which echoes to Searle's account of institutions as "regulative rules." Such rules "emerge from the agents' actions but do not play any causal or functional role in the rise of these actions" (Hédoin, 2015, p. 6). They "regulate an existing practice or activity" (ibid, p. 4). They even "facilitate a social activity or even contribute to enhance its efficiency"; this is exactly the role that the focal points play: they facilitate coordination and eventually arbitrate between individual and collective divergent will. Salience and focal points as institutional objects in this case emanate from individual behavior but do not explain such behaviors: "given the history of past plays in a game, each agent is incentivized to reproduce the behavioral pattern, thus leading to a further reinforcement of the pattern" (Hédoin, 2015, p. 6).

With another reading of Schelling, institutional facts like conventions, norms, etc. are associated with Searle's "constitutives rules" (Searle, 1995, 2010). "Constitutive rules literally create a new institutional reality" (Searle, 2010, p. 97). When players together decide that one suggestive detail, one key becomes the key, to paraphrase Schelling, making accordingly a focal point, they ascribe constitutive rules to such focal point. They decide that such focal point induces common reasoning and convergent expectation, i.e. serve as a coordination device. Focal points indeed

[29] Social or cultural knowledge is a knowledge "of which arrangements are salient or traditional in that culture and so provide coordination" (Bacharach and Hurley, 1991, p. 3). It corresponds to a "shared institutional heritage that brings common understanding into a population" (Hédoin, 2014, p. 383) or to the knowledge of the set of "historical and cultural points of reference that define the identity of the group" (Orléan, 2004, p. 208).

create a "new institutional reality," as they are responsible for the creation of traditions or norms, conventions, etc. They establish new practices (see Hédoin, 2015, p. 2). Such rules change the game that the players are playing, making such game a coordination game and possibly defining new strategies, new beliefs and preferences:

> "to be constitutive, a rule must in one way or another define the game the players are actually playing . . . a constitutive rule is directly responsible for the game's structure (e.g., the players' strategy sets or payoff functions) and possibly for the players' beliefs and behavior. (ibidem)

Some focal points – as institutions – in Schelling's work indeed transform the game that the players are playing, making a divergent interest game into a coordination game, changing the set of strategies, preferences that initially defined the game, or modifying players' beliefs.

It means that the structure of social reality, as well as the institutional system, shapes the resolution process of individuals' interdependent decisions. Individuals can take for granted certain arrangements of society. Schelling claims that "[t]he solution depends on some kind of social organization, whether that organization is contrived or spontaneous, permanent or ad-hoc, voluntary or disciplined" (Schelling, 2006[1978], p. 126). Accordingly, institutions causally affect individuals. They orient and condition individuals' decisions and actions.

Subsequently, in Schelling's work, society is among others, constituted by a set of institutions.

> "A good part of social organization – of what we call society – consists of institutional arrangements to overcome these divergences between perceived individual interest and some larger collective bargain" (2006[1978], p. 127).

Society is constituted by 'institutional arrangements' facilitating and allowing the conciliation between both heterogeneous agents and collectives, since the aggregate is emergent and do not necessarily match with individuals' intentions. Society is an enduring entity supported by stable institutional and organizational systems. In Schelling's view, there are enduring and pervasive patterns in the social world. This is true for institutions.

Schelling's account of both strategic and social interdependent decisions is characterized by open-systems (Lawson, 2003). Schelling declares:

> "Abstractly speaking, stable organisation comes in at least two forms or 'systems', the first of which we might refer to as an environmentally closed, or equilibrium, system, and the second of which we can label an environmentally open, or far-from- equilibrium, ... equilibrium systems are stable if there are no disturbances from the outside environment; far-from-equilibrium systems require perpetual inputs from the environment to endure and be stable." (Schelling, 1984, p. 357)

Because micro variations exist constantly, the pervasiveness of social and institutional facts, when there is no legal enforcement, requires "perpetual inputs" or individuals' continuous acceptance of the underlined patterns of behavior.

Another aspect that appears in Schelling's work is that society is multileveled even if he does not build a theory of the interaction among these different levels. Several interdependent types of collective (like ethnic, religious, socio-professional categories, etc.) constitute it. Each one possesses its own underlying value system. We can mention for instance Innocenti (2007, p. 416) who asserts that Schelling, through observation, builds categories of individuals and that it constitutes the underpinning of his thinking and his theory building. Therefore, different complementary types of collective exist in Schelling's account of economic systems. First, there are collectives which correspond to social, cultural, ethnic, etc. categories. Those collectives enhance the identification of common characteristics; for instance, common perceptions and eventually common ways of reasoning. Second, there are collectives which are formed by specific interactional contexts, i.e. strategic contexts. In the latter category, a collective is formed as soon as there is a strong interdependency between individuals, i.e. a situation involving a strategic dimension. When players create a new institutional fact when making a focal point, they become a specific community, even transitory. As such, focal point induces common reasoning, it is associated with rule-following which is mainly a "community-based practice" (Hédoin 2014, 2015, 2017). And all of these types of collective participate to the construction of social reality.

In summary, the ontological commitments behind Schelling's theory of interdependency emphasize that he attempts to incorporate in economics a rich account of social reality. The richness of the social ontology behind Schelling's theory of interdependence has important methodological consequences, especially regarding (i) the role of *methodological individualism* within social sciences and a fortiori in economics, and (ii) the relation between economics and other social sciences. His theory of strategy requires a view of economics and economic theory as open-systems.

3. The models of residential segregation

The dynamic models of residential segregation involve similar fundamentals with Schelling's strategic analysis in game theory. In fact, some of the main methodological statements for strategic interactions are used in the formal approach that underlies the dynamic models of residential segregation. However, one major difference relies is that the intersubjective dimension that prevails in his reorientation of game theory disappears. This is because "strategic analysis typically involves a small number of interacting decision units" (Schelling, 1984, p. 203) Nevertheless Schelling sees the importance of the evolutionary aspect of game theory which should not be opposed to an eductive or epistemic approach. Understanding how a particular interaction occurs in a specific context requires in first instance to understand the dynamic of evolution which led to this time and place. And again this means having a vision of game theory and games as open-systems.

Both his reorientation of game theory and the dynamic models of residential segregation indeed exhibit coherence in Schelling's thinking. They underlie a "theory of interdependent decisions" (Schelling, 1980[1960], p. 83).[30] They provide (i) the methodological innovations required to account for "realistic" strategic interactions and (ii) the conceptual tool to formalize these interactions. It is often claimed that Schelling criticizes standard game theory without proposing any formal solution. To the contrary, we assert that – as the precursors of the agent-based models (e.g., Aydinonat, 2005, 2007; Kirman and Vinkovic, 2006; Epstein and Axtell, 1996; Epstein, 2006) – the dynamic models of residential segregation are a methodological answer to the constraints Schelling early identifies in standard game theory. The social ontology previously underlined in his conception of a "theory of interdependent decisions" helps to understand this statement.

According to me, these two contributions cannot be dissociated. They jointly allow for (i) a broad understanding of Schelling's conception of the interdependence of individual choices and actions, (ii) an appreciation of Schelling's methodological innovations for game theory, and (iii) an illustration of how there are two intertwined aspects in strategic interaction for Schelling: an epistemic one and an evolutionary one. Considering these two sets of contributions as part of an overall project allows to delineate the scope for onward enhancements of the methodological innovations he proposes within an enriched conception of game theory.

3.1. The purpose of the models

In several interviews after the Nobel Prize, Schelling tells the story of the conception of his dynamic models of segregation. He describes a context in the United States in which neighborhoods are too often composed of either almost only 'whites' or almost only 'blacks' (e.g. see Schelling, in Steelman, 2005, p. 40). Schelling was dubious about the most unsophisticated explications that would be in his own words "rampant racism" (ibidem), and critical about Becker's analysis. He severely criticizes Becker's treatment of discrimination in *The Economics of Discrimination*, when he claims

> "[Becker] had a piece of machinery that was cranking out results, and that he wasn't sufficiently interested in racial segregation to look and see what was going on. He just decided to throw a parameter into a preference function, giving everybody a "taste" for being with or not being with people of another color ... What he is primarily interested in is showing that traditional economic models are all you need ... he doesn't appear to think there is anything to learn from outside economics. He is not interested in coupling the methodology of economics with the methodology of sociology." (Schelling, in Swedberg, 1990, p. 194)

---

[30] The model of "dying seminar" (Schelling 2006[1978]) is another example of the kind of social interdependence involved in the dynamic model of residential segregation; I am however more interested by the simulation aspect of the latter and its presentation by the author in a more systematic way.

In this pervasive characteristic of American society, Schelling sees the *collective result* of interdependent individual decisions such as where to live, with whom, and therefore, the outcome of a dynamic of interactive decisions and behaviors. When telling the story about how he came to his models of residential segregation, Schelling indeed says "I had a strong intuition that you can get a lot of things like fairly extreme segregation through the dynamics of movement" (Schelling in Aydinonat, 2005, p. 4).

His main interest was therefore to explain how segregation in the United States could rise from mild-discriminatory behaviors (Schelling, 2006, p. 254), i.e. from individuals preferring to leave in mixed area however without being within a minority.[31] Schelling's intuition was that this kind of behavior could lead to residential segregation. His objective was then to look at the process of interactive behaviors and to verify if such process can eventually lead to a segregated area (Schelling, 1984, p. 254).

> "To understand what kinds of segregation or integration may result from individual choice, we have to look at the processes by which various mixture and separations are brought about. We have to look at the incentives and the behavior that the incentives motivate, and particularly the way that different individuals comprising the society impinge on each other's choices and react to each other's presence" (ibid, p. 259).

Schelling attempted to come with a simple and tractable model for studying this mechanism but at the same time, general enough to be applied to comparative social phenomena (see Schelling, in Steelman, 2005, p. 41).

3.2. The models

Schelling (1969; 1971a; 1971b; 1972) develops three models of residential segregation: (i) the spatial proximity model, (ii) the bounded neighborhood model and (iii) the model of tipping – which is an application of the bounded neighborhood model.[32]

Each type of models is characterized by the *interdependence* of the individuals, in the following sense:

> "people responding to an environment that consists of people responding to an environment that consists of people who are responding to each other. As people respond they change the environments of the people they associate with, and cause

---

[31] Schelling (2006[1978], p. 138) defines a discriminatory behavior as follow: "an awareness, conscious or unconscious, or sex or age or religion or color or whatever the basis of segregation is, an awareness that influences decisions on where to live, whom to sit by, what occupation to join or to avoid, whom to play with, or whom to talk to."

[32] "'Tipping' is said to occur when a recognizable new minority enters a neighborhood in sufficient numbers to cause the earlier residents to begin evacuating" (Schelling, 2006, p. 302).

further responses. Everybody's presence affects, if only slightly, the environment of everybody else." (Schelling, 2006 [1978], 169)

Then Schelling postulates a twofold population (presently blacks and whites, but the same models can be applied for boys and girls, young people and old people, etc.) (ibid, p. 138). Within this twofold population, he assumes that individuals have a preference for a certain ratio of the two colors within the population, i.e. "some ratios or average or percentage of the total", and specifies "the dynamic of response" (ibidem). The individuals have mild-segregationist preferences.[33] Schelling insists on the fact that "absolute numbers do not matter, only ratios"(ibid, p. 156).

The main difference among these models lies in the characterization of the neighborhood under scrutiny, i.e. in the way each individual making a decision of location apprehends her environment.

3.2.1. The spatial proximity model

In the *Spatial proximity model*, each individual defines her own neighborhood according to her own location: "there are no objective neighborhood boundaries" (ibid, p. 260). Individuals evaluate the color ratio of their own neighborhood. If this ratio fits with their requirement they stay. If not, they move from their initial location to another appropriate place fitting their requirement. Schelling assumes a rule of motion: from the left to the right, unsatisfied individuals move to the nearest place where they can be satisfied by the color ratio and they continue to move as long as necessary for finding their satisfying location.

The equilibrium of the model is reached when everybody is satisfied with his own neighborhood, i.e. when the collective structure is stabilized.

Schelling first experiments this process with individuals placed on a line (cf. figure 1). Then he changes the distribution area with individuals placed within a square like a checkerboard (cf. figure 3). In each case he positions the two types of individuals randomly (blacks and whites are represented by "pluses" and "zeros"). He compares the possible outcomes with different initial distributions of "pluses" and "zeros", with equal and different numbers of "pluses" and "zeros", with different rules of motion, with different sizes of neighborhood, and with different

---

[33] Schelling however affirms that "integrationist preferences" could be applied to the same models. For him, it only implies to "postulate a preference for mixed living and simply reinterpret the same schedules of tolerance to denote the upper limits to the ratios at which people's preference for integrated residence is outweighed by their extreme minority status" (Schelling, 2006 [1978], p. 165). Integrationist preferences only means "to assume that members of both colors have certain minimum demands for neighbors of like color, but no maximum demands" (Schelling, 2006[1978], p. 282). The same kind of conclusions can be drawn in both cases: "[t]he same model fits both interpretations. The results are as pertinent to the study of preferences for integration as to the study of preferences for separation" (ibidem). It is merely the same kind of dynamic phenomenon. In each case, the prime matter is the issuing dynamic of interactive decisions and successive individual motions.

preferences regarding color ratios (for instance, one of the two types of population is endowed with more segregationist preferences).

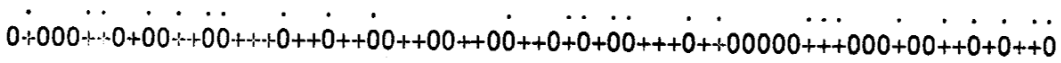


Figure 1

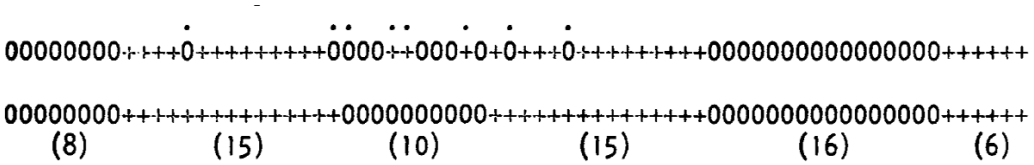


Figure 2

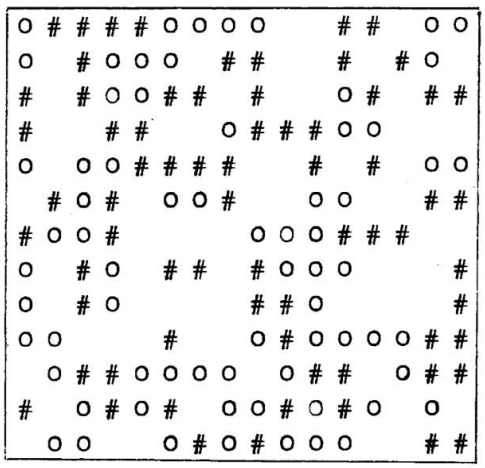


Figure 3

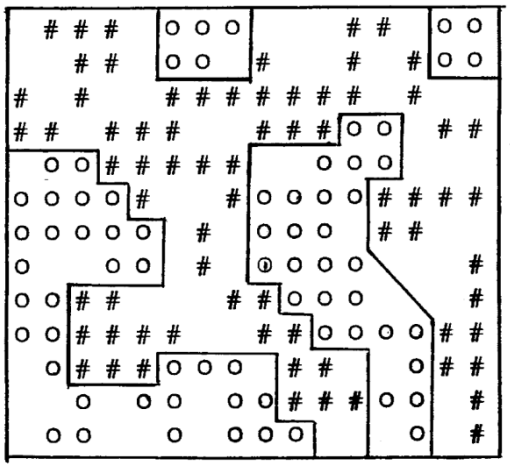


Figure 4

These figures (entitled figure 1, 2 and 3 and 4) are taken from Schelling's *RAND Memorandum RM-6014-RC* (1969).

In each type of configuration (i.e. either with lines or with squares), the conclusion is straightforward. A general pattern appears: the resulting distribution of pluses and zeros shows a segregated area (cf. figures 2 and 4). This phenomenon is quite robust to different initial conditions and to different parameters of individual choices. There are always clusters of 'like-color' even if variations of the initial conditions have some effects on the resulting distribution.[34]

---

[34] See Aydinonat (2007, p. 441) for a presentation of the different theoretical and empirical tests made on Schelling's checkerboard model. Except for Bruch and Mare (2003) they all emphasizes that Schelling's results

For instance, with individuals in line, reducing the size of the neighborhood induces more clusters of 'like-color' (ibid, p. 265). At the opposite, enlarging the area that the individuals consider as their neighborhood can attenuate the segregation (ibid, p. 281). If the initial number of the two populations is not equal, the minority tends to be less clustered than the majority (ibid, pp. 266-67). Even changing the rule of motion can lead to different distributions. Nevertheless as Schelling asserts "the order of moves makes little difference" (ibid, p. 265).

With individuals positioned within a square, increasing the ratio of like-color required by the members of the population (ibid, p. 274) induces according to his experiments the following results: the number of initial discontent raises, the dynamic of motion increases, and the density of like-color clusters increases (ibidem), and (iii) "the greater the demands the more movement is induced by those that move on the part of those that were originally content" (ibidem.). When there are unequal numbers of pluses and zeros, but equal demands of like-color in their neighborhood, the minority tends to constitute "larger cluster" (ibid, p. 277). Furthermore, in this case, the density of the clusters of the minority are denser than the clusters of the manque un mot (ibid, p. 280). And with different demands of like-color of the two categories of population, even if they are in equal number, the more demanding category ends with a neighborhood in which the like-color ratio is finally higher than initially demanded (ibid, p. 279).

As a consequence, whichever the initial conditions, the dynamic of interactions leads to segregation and none of the outcome corresponds to the preferences of the individuals composing the population.

3.2.2. The bounded neighborhood model and the tipping phenomenon

In the "bounded neighborhood model", the definition of neighborhood changes "instead of everyone's defining his neighborhood by reference to his own location, there is a common definition of the neighborhood and its boundaries" (ibid, p. 284). Individuals are either 'in' or 'out' of a given neighborhood (ibid, p. 260). In order to move either in or out, each individual evaluates the color ratio within the neighborhood in question (ibidem). Each population is represented by a curve symbolizing the cumulative frequency distributions of the individual's tolerance thresholds.[35] This tolerance level is the upper limit of the color ratio that an individual can tolerate. Beyond this tolerance threshold, the individual is dissatisfied and, either decides to move out when she is already located within the neighborhood, or when this is not the case, to stay outside the neighborhood.

In this model Schelling experiments the distributions of tolerance thresholds compatible with 'dynamically stable mixture'. He attempts to identify if some initial conditions are compatible with a stable mix equilibrium, if the initial conditions influence the outcome and what kind of

---

hold. Sugden (2000, 2009) also recognizes that Schelling's results are robust, even if he adopts a critical assessment of Schelling's checkerboard model.

[35] More precisely: "the cumulative form measures, for any number of anticipated attendance, the number of people for whom that number is large enough" (Schelling, 2006 [1978], p. 103).

numerical constraints can alter this outcome (ibid, p. 260). Again Schelling tests different initial conditions: with equal and unequal numbers of blacks and whites, and with different shapes of the curves symbolizing the frequency distribution of individual's tolerance (ibid, p. 285). The dynamic here has much more importance than previously for Schelling. Even if the initial conditions have an impact on the outcome, as Schelling claims "it is the dynamics of motion, though, that determine what color mix will ultimately occupy the area" (ibid, p. 288).

In the different configurations tested there are several possible outcomes or equilibria. Two types of equilibria generally compete: an equilibrium of all blacks or of all whites, and a mixed equilibrium.

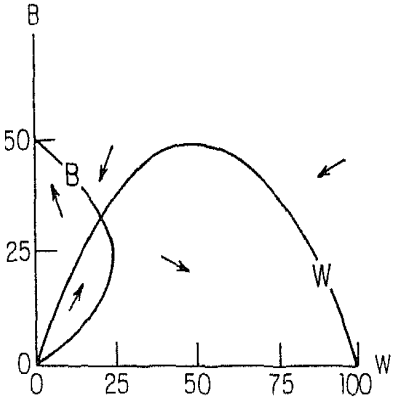


Figure 1

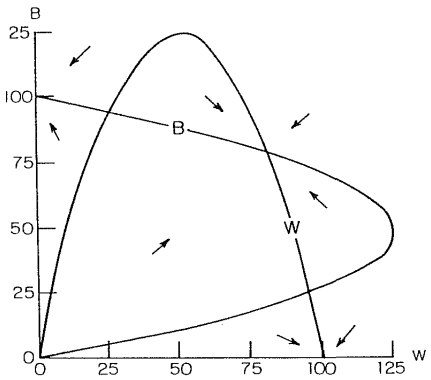


Figure 2

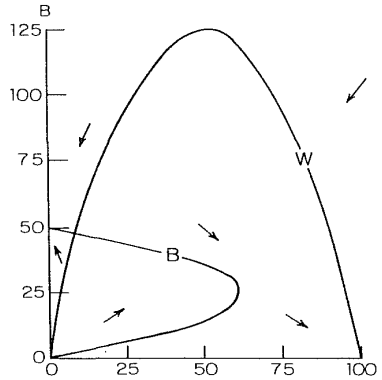


Figure 3

These figure are taken from *Micromotives and Macrobehavior* (2006[1978])

Each abscises represents one category of the population (B for blacks and W for Whites). Each curve represents the frequency distribution of the tolerance thresholds for each category. The arrows symbolize the attraction points.

As shown in the figures 1, 2 and 3 each intersection of the two populations' curves is a potential equilibrium in which blacks and whites can live together. The different shapes of the curves in the figures above show however that the two types of 'mono-color' equilibria are attractive points. The 'mix-color' equilibrium is very sensitive to perturbations, and tightly relies on both the initial conditions and the process of interactions:

> "the occurrence of several mixed-color stable equilibria is usually sensitive, though, to small changes in the shapes and the positions of the curves. It is the extreme one-color equilibria that tend to be least disturbed by shifts in the tolerance schedules or changes in the aggregate numbers; and the occurrence of a single mixed stable equilibrium may be fairly immune to shifts in the curves" (ibid, p. 296).

For instance, in a situation where there are two possible equilibria (a total occupancy by either blacks or whites), which one of the equilibrium will occur depends on where the motion process starts (i.e. is there a sufficient number of one of the color in the neighborhood?) and on the relative speeds of entrances and departures. If one color is dominant within the neighborhood, this color will obtain a total occupancy. If the two colors are initially present in the neighborhood

and in 'statistically viable numbers', and if the relative speed of the entry of one the color is higher this is very likely that again, this color will obtain the full occupancy (ibid, p. 289).

Now, if initial conditions allow for a mixed equilibrium, the problem is that such mixture induces perpetual motions. The mixture necessarily attracts outsiders that are satisfied by the ratio color. However, by doing that, they attract potentially new outsiders and create unsatisfied insiders that leave, and so on,. so that the ratio is constantly modified which creates new motions. For instance, if one color begin to be numerically superior, the insiders of the opposite color start to leave; they reduce again their numerical inferiority, which create new dissatisfied insiders; they move out; the minority then became more and more smaller so that this minority finally leaves the neighborhood, and there is ultimately, again, a full occupancy of one color only (ibidem). If and only if the initial size of both populations is below the two curves, a mixture of blacks and whites could be an equilibrium.

Finally what is called the "tipping" phenomenon, which is a case comparable to the "bounded neighborhood model" – since tipping occurs when the entrance of a minority induces the evacuation of the former residents. Therefore, it is exactly a case that fits to the previous model, even if it is more complex than the situations he tests in his "bounded neighborhood model".[36]

Therefore, as for the spatial proximity model, the dynamic of interaction mainly explains the outcome and such result is not representative of the preferences of the individuals.

3.3. Some methodological insights

The dynamic models of residential segregation participate to Shelling's conception of game theory, as a framework of analysis and as an open-system. His understanding of interdependence implies a non-mathematical and non-static view opposed to aggregation, to exogenous and *a priori* defined preferences. His conception of game theory requires the determination of how and why an outcome occurs. An historical and evolutionary explanation is necessary. The models of residential segregation provide an explanation of how and why segregation occurs, and a theoretical framework.

The two types of models presented by Schelling are based on a set of variables with respect to the individuals' decisions and to their environment, and encompass some variations of them, allowing for different complex dynamics. Schelling experimented some of these variations, but much more could have been tested (ibid, p. 164). In fact, his models are conceived like experiments. Schelling (2006, p. 261) claims that the results are "experimental". Besides, his models present several limits. To set some of these limits: Schelling postulates that information is perfect; every individual knows the color ratio within the neighborhood she evaluates at the

[36] For a detailed version of the tipping phenomenon, see Schelling in *Racial Discrimination in Economic Life* (Ed. by Pascal, 1972). It is also important to note that the mechanism involved in the bounded neighborhood model and tipping, like in the dying seminar, is the critical mass. This is not the case for the spatial proximity model.

moment she makes her choice (Schelling, 2006, p. 285); she does not anticipate the phenomenon of motion and does no take into account the intentions of the others (Schelling, 2006 [1978], p. 156). If anticipations were allowed the segregation phenomenon be aggravated (Schelling, 2006, p. 307).

For these reasons, Schelling's models are incomplete and he was perfectly aware of that fact. Being aware of these insufficiencies in the models is characteristics of an open-system methodology (Setterfield, 2016; Chick and Dow, 2005; Chick, 2004; and Mearman 2003). It corresponds to a "conditional closure" and a "temporary closure"; it is a "partial analysis within [an] open-system theory" (Chick and Dow, 2005, p. 369). They do not translate all the possible individuals' incentives and processes leading collectively to residential segregation. Hence, Schelling (2006, p. 268) recognizes that "this is too abstract and artificial to be a motion picture of whites and blacks ... but it is suggestive of a segregating process and illustrates some of the dynamics that could be present in individually motivated segregation." He justifies the need to abstract because "there are many different incentives or criteria by which blacks and whites ... become separated" (Schelling, 2006 [1978], p. 142). He acknowledges that in reality "it is not easy to tell from the aggregate phenomenon just what the motives are behind the individual decisions, or how strong they are ... the dynamics are not always transparent. There are chain reactions, exaggerated perceptions, lagged responses, speculations on the future, and organized efforts that may succeed or fail." (ibid, p. 146) This complexity is explained by the fact that the phenomenon studied is characterized by continuous feedback loops, i.e. as individuals react to their environment they change this same environment, this causes other reactions, the environment continues to change, it causes other reactions, and so on (Schelling, 2006, p. 169). Such real-world system characteristic of an open-system ontology therefore requires partial and temporary closure to be studied (Chick and Dow, 2005; Chick 2004).[37]

Accordingly, this situation is too complex (i) to accurately characterize all the possible explanatory variables, and (ii) to predict the outcomes of these individual interactions – or interdependent decisions. This may explain why Sugden (2000, p. 12) claims that Schelling's models are "similar to the neoclassical model … in their use of highly simplified assumptions". He considers that Schelling's contribution is mainly *conceptual*. Schelling is for him "pointing to an error in an existing theory" (Sugden, 2000, p. 9; my emphasis).

What Schelling has offered is however a general framework allowing to formalize complex interactive systems in which people are dependent on each other and on their environment and in which as the system evolves the individuals' decision function modifies. As we show below this is exactly the kind of framework that Schelling informally develops in his reorientation of game theory. Indeed, we claim that Schelling's dynamic models of residential segregation incorporate some of the main postulates prevailing in his reorientation of game theory. Besides, as Schelling argues in the preface of *Micromotives and Macrobehavior* (2006, p.4) his models of segregation are n-players games.[38]

---

[37] See Chick and Dow (2005, p. 366) for the conditions characterizing a real-world system as an open-system.

[38] He explicitly says: "I do appreciate that it is easily construed as multi-person game theory."

Here are the main arguments sustaining such statement. First, as he claims for game theory, the outcome of the interaction process is the result of the dynamics of interactions. He is more interested in the examination of the intermediate process than in the equilibrium *per se*. And while in standard game theory the outcome is foreseeable – since it directly relies on individuals' preferences,[39] in the dynamic models of segregation different possible equilibria are possible from the same initial conditions (it depends for instance on the different speeds of motion). It becomes therefore difficult to predict which one of the competing equilibria will be reached. The existence of a particular equilibrium cannot be *a priori* stated. Again this is exactly the position taken within his reorientation of game theory.

Second, even if according to Schelling (2006 [1978], p. 176) his models are "an example of "equilibrium analysis"", the equilibria are neither optimal nor the mere aggregation of individuals' preferences. He insists on the fact that "there is no presumption that the self-serving behavior of individuals should usually lead to collectively satisfactory results" (ibid, p. 25): "an equilibrium division is not likely to have any optimal properties." (ibid, p. 182) Equilibria are interpreted in terms of stable patterns of behavior. We find the same idea in Schelling's account of focal points. His idea is to show how different configurations can lead to regular general patterns which are, in the present case, segregated area. And, again, instead of having a static conception of stability like in classical game theory, the conception is dynamic (Ayson, 2004, p. 186).[40] Thus, the link between individuals and the collective in Schelling's models, in which social patterns are not the simple summation of individual characteristics – i.e. individual preferences – differs from standard game theory.[41] But it is again characteristic of an open-system reality (Chick and Dow 2005; Chick 2004).

Third, because individuals' decisions are based on the environment they contribute to modify by acting, rational choice theory cannot be a tool for formalizing individuals' decisions. Indeed, Schelling asserts that "the person has a preference about [a] statistic, and the person contributes something to that statistics" (2006 [1978], p. 186). This is rather far from any standard premise in individual decision-making. But again we found a comparative basis in his account of strategic interactions. Players' preferences contribute to influence the game, the others, in the same way as the environment, the game, and the others influence players' preferences. As we will explain, the hypothesis of "context-independency" does not prevail in Schelling's contribution. The standard conception of individual rationality in terms of consistency of choices cannot prevail. The agent-based model can incorporate such requirements. The dynamic models of residential segregation are indeed a very rudimentary form of ABM. However contemporary ABM, with their spread and subsequently their improvements provide, in our opinion, the methodological tool suitable for Schelling's conception of interdependence both in strategic and social interaction. Contemporary ABM can indeed conciliate the two perspectives that are developed in Schelling's

---

[39] When preferences are understood as mental states. For a detailed discussion of this debate regarding the understanding of preferences and beliefs either in a behavioristic or mentalistic way see chapter 1.

[40] For a discussion of the pervasiveness of Schelling's concept of stability and its evolution throughout his work, see Ayson (2004)

[41] For Schelling (2006 [1978], p. 14) segregation is about the kind of situations "that usually don't permit any simple summation or extrapolation to the aggregates" since the "aggregate results ... sometimes have no recognizable counterpart at the level of the individual" (Schelling, 2006, p. 256).

account of game: the "epistemic" or "eductive" one and the "evolutionist" one (Phan, 2004, p. 371). Even if it is mainly the evolutionist approach that is developed in the agent-based computational economics (ACE) a "learning dimension" is more and more integrated in a more complex way than a simple adaptive learning like generally in evolutionary game theory (Phan, 2004, p. 372). In ACE, the eductive perspective which involves learning implies either "belief revision" or "eductive coordination in the case of rational agents playing a game with their nearest neighbourhood" (ibid, p. 373). This is accordingly perfectly suitable for Schelling's conception of game theory.

We will discuss these last features in more details in the next part of the chapter to demonstrate how Schelling's reorientation of game theory is based on exactly the same postulates.

4. How Schelling challenges standard methodological individualism

There are strong methodological implications according to the way Schelling portrays players and according to his conception of strategic rationality. A game is a complex decision problem in his vision. In fact, as accurately highlighted by Innocenti (2007, p. 421)

> "The key hypothesis of [Schelling's] approach is that economies are not just collection of of homogeneous agents but complex dynamic systems characterized by dispersed interaction among heterogeneous agents acting locally on each other in some space"

Heterogeneity means here that players can have different perceptions of a common situation, different needs and desires and different beliefs, as their inductive inference when belonging to different communities or their personal and historical backgrounds of interactional situations, are different. Such heterogeneity however does not imply that players cannot ultimately come to a tacit agreement. Precedent, common interactional experiences, the belonging to common communities induce homogenization, i.e. a common intersubjective background which ultimately seal coordination.

Besides as will be argued in this section, Schelling is more inclined to a loose version of methodological individualism than the standard rational choice theory (RCT) and standard rationality in game theory – which is a direct extension of the RCT (e.g. see Bacharach and Hurley, 1991; Mariotti, 1995).

Boudon (2004) identifies six postulates – among which three are characteristic of methodological individualism – underlying RCT. First, every social or economic analysis has for purpose to understand the individual intentions, beliefs, decisions or behavior that led to social or economic phenomena. Second, understanding the individual beliefs, decisions or behavior is to explain the meaning that they have for the individual. Third, there are reasons behind individual beliefs, decisions and actions, i.e. there is a rationality postulate behind those attitudes. The three other postulates characterize such rationality principle. First, the individual is concerned by the consequences of her decisions and actions; second, she is concerned by the consequence only for her; and third, the individual is a maximizer. While Schelling subscribes to the first and second

postulate, as will be demonstrated, he deviates in some respects from the standard way that the other postulates are implemented in economics and game theory.

4.1. What is a player

Schelling's breaks with the traditional *representative agent* of classical game theory (see Rivzi, 2006; Colman, 2006; Sugden and Zamarrón, 2006; Rivzi, 2007; Innocenti, 2007). Considering that the purpose of Schelling's reorientation of game theory is the analysis of real game like situations, he is indeed interested by players that are flesh and blood people. His players can no longer be portrayed as the economic representative agent. The heterogeneity and the subjective or personal identities that shape the players are the essence of players' subjectivity in Schelling's work.

It is widely acknowledged that one great part of Schelling's originality is linked to his peculiar conception of players (e.g. see Innocenti, 2007, p. 410). For instance, according to Innocenti (2007), Schelling's account of the players in his games or models, is indeed the main explanation of his empiricism (ibid, p. 409).

We will focus in this section on the way Schelling conceives the agents in his games. The players are heterogeneous – and this heterogeneity bypasses the standard game theoretic account of individual heterogeneity; they are socially skilled, and they are constituted by multiple identities (collective or individual).[42] Besides, Schelling's conception of a player can in turn clarifies his conception of *interactive rationality*.

The first element characterizing the players is their heterogeneity. Different aspects in Schelling's thinking explain players' heterogeneity: the perceptions, the mental representations, the mode of reasoning, the social role, the individual and collective identities, the personal experiences, and so on. Therefore, contrary to game theory, heterogeneity it is not a matter of information (and incompleteness of information), or of preference in Schelling's work.

Therefore, players' heterogeneity first relies on the difference among players' mental representations, perceptions, frames, etc. (Innocenti, 2007, p. 410). Every player has a subjective interpretation of the games she faces and the potential clues surrounding the games (ibid, p. 414). This statement is undeniable when remembering that players have (i) to communicate their intentions – and their underlying perceptions – through their behavior, and (ii) to pay attention to others' behavior in order to have a bit of insights into their value systems, which in turn can reveal their eventual perceptions. If players, confronted to a common situation, have to

---

[42] I will not detail his investigation of individuals as multiple selves, since it is a methodological device mostly linked to his account of individual choice theory and I focus in this paper on interactive decision- making. For an overview of this aspect in Schelling's thinking see his book Choice and Consequence (1984) or Innocenti (2007). However it is worth noticing that "[o]nce again, notwithstanding his efforts to be conciliatory with the mainstream, Schelling departs from standard decision theory in a way that asks for rebuilding of its main foundations. By also taking the hypothesis of heterogeneity within the self, Schelling makes it the keystone of his methodology." (Innocenti, 2007, p. 424)

progressively discover and understand the other, it is precisely because they are heterogeneous and that they may not perceive the same things from a common situation.

> "player heterogeneity is not restricted to preference orderings or available strategies, but extends to include the players' labels and attributes, the mental representations of the game ad the processes of preference formation. This method implies not only that players may perceive the same game differently but also that they are not assumed to think that other players reason symmetrically to them." (ibidem)

In addition as Innocenti emphasizes, "[h]eterogeneous players can follow different processes of belief formation, which become individually specific acts." (ibid, p. 415)

This heterogeneity of individual perceptions and mental schemes has for consequence the reintroduction "in the analysis of those psychological elements removed by the assumption of the representative agent" (ibid, p. 414) long before psychology made its way back in the realm of economic theory and in particular within decision theory and game theory (see Bruni and Sugden, 2007). To portray more "realistic" players, Schelling draws on psychology (and especially on Gestalt psychology) and on social psychology (Schelling (1980[1960], p. 108; Ayson, 2004). He reincorporates in individuals' decision making some psychological assumptions.

Schelling's references in psychology are however quite light. Yet, all the new game theoretic arguments that he develops are merely about individuals' psyches. He often refers to a game as involving the 'psychic phenomenon' of the convergence of player's expectations. Players' psychology and in particular players' cognitive processes is the ground of a game. Schelling also often refers to the players' perceptions, to their specific and subjective labeling of strategies or game matrices, and to the way these matters of cognitive psychology determine decision-making and accordingly the outcome of a game. Recall that it is the suggestive details of a game or surrounding a game that orient toward one solution among all of the possible ones, and even when players can solely rely on the game matrices, they label these matrices in a specific way; i.e., some strategies have a power of suggestion. Again the fact that game theorists do not admit that individual and strategic rationality cannot derive solely from the axiomatic of choice and from exogenous preferences, but primarily from the way individuals perceive their decision problem, may prevent the resolution of a game.

If such incursions into some domains belonging to cognitive psychology are of prime importance in Schelling's account of strategic interaction, he does not develop so much this aspect. Yet, many scholars identify Schelling's work as a psychological contribution to game theory. For instance, Colman (2006) mentions Schelling's "psychological decision theory".

Another aspect concerning players' psychology is that players have limited cognitive abilities. According to Innocenti (2007) this is precisely because players have limited rationality that they can be successful in coordination games. Indeed, players are not necessarily able to draw complicated conclusions from a mathematical structure; they are not over-skilled computational entities. Some problems of attention exist. Players may not pay attention to all of the potential relevant elements that matter for their decision-making. They can simply focus on some prominent or conspicuous elements surrounding the game.

In addition to the players' subjective perceptions of the game *per se* or of the context, it is the perceptions of whom the others are (i.e. their ethnic and socio-economic categories) that primarily matter. The importance of this context will be developed in the model of games we propose in the chapter 5 to outline a theory of games in which coordination is generated by focal point.

> "The subjects populating Schelling's thought or real experiments do not suppose that other players follow identical or symmetrical rules of logical inference to make their choices." (Innocenti, 2007, p. 409)

This explains why players must put themselves in the other's shoes. This capacity to put oneself in the other's shoes, in particular through a process of simulation (Goldman, 2006) will be developed in our model of games in the chapter 5. From that prospect Schelling anticipates some of the major and more recent developments of game theory that incorporate news insights from cognitive sciences and neurosciences (Guala, 2018, 2016; Lecouteux, 2018a,b; Larrouy, Lecouteux, 2017; Devaine, Hollard and Daunizeau, 2014; Schmidt and Livet, 2014; Morton, 2012). This incorporation is moreover the result from critical thinking on the philosophy and methodology of game theory. For Schelling, players' *strategic reasoning* is first and foremost explained by their capacity to see the problem from the other's angle, to try to understand this other by putting in her place in order to come to a prediction of her behavior. Here is one of the elements that portray a player as a *socially skilled* individual. And this aspect of players' strategic reasoning is rarely underscored in the reflexive literature on Schelling's contribution to game theory.

As Schelling declares, as individuals

> "We figure out what a person might do by putting ourselves into his position, adopting for the purpose as much as we know about his preferences, and deciding what he ought to do or what we would do if we were he. By « what he ought to do » we merely mean what he should decide in accordance with his own aims, values and objectives, given the alternatives that he faces. The supposition is that he has a choice to make, can reflect on that choice taking into account things that we perceive or infer, and will decide not in a haphazard way and not Under the dominance of some unrevealed drives but in a way that suits his apparent purposes. If we have some clue to what his purposes are, and some appreciation of the alternatives he faces, we anticipate his choice by figuring out what one would do in that situation with those aims and values. In its pure form, this is cheap theory. We rely on empirical information about a man's objectives and what alternatives are available to him; but from there on we must imagine ourselves in his position, see how we should proceed if we were he, and conclude that they may go ahead and do just that." (Schelling, 1984, pp. 205-06)

The *cognitive mechanism* that Schelling describes in this quotation, when players have to find a tacit agreement, i.e. when they have to come to a meeting of minds, is exactly what is defined by Goldman (2006) as the simulation process. Goldman's account of simulation will be extensively detailed in the chapter 4 of the thesis. In that perspective, it is interesting to mention that when Zeckhauser relates some of his personal interaction with Schelling he says:

> "And what better time than a reunion to play out Schelling's widely discussed lessons on tacit communication (1960, pp. 83-118): "[I]f you reflect hard, and place yourself in another's shoes, you can figure out where and when to meet on campus, say after initial arrangements have broken down." … Those who read Schelling and participate in his games learn a more general principle: In any interactive situation it is vitally important to look at matters from the side of the other party. The other-people's-shoes approach is often recommended by soft-hearted promoters of compromise. The core principle, however, is that by understanding the other party's perspective you will improve your comprehension of the situation dramatically and will come out better yourself." (Zeckhauser, 1989, p. 155)

Schelling's conception of heterogeneity is also linked to individuals' personal identity. This personal identity is shaped both by individuals' personal experience and by their multiple collective identities. The stable patterns of behaviors that Schelling mentions in game theory, which are determined by social roles, traditions, institutions, conventions, etc. rely on players' collective identities. By taking into account these social dimensions (i.e. the institutions, conventions, etc.), players think of themselves as members of a group. It is because of this identity feeling (i.e. the acknowledgement that they belong to a community) that they can follow some established patterns of behavior, and that they can ultimately rely on these coordination device knowing that they offer some 'clues' that are understood in the same way by the other members of the same community. Some institutions, that are preexistent to the game, give to the members of the community 'clues' to coordinate. Players accordingly integrate into their reasoning *social components* which are recognized by themselves as members of this group. In fact, the way members of a group interact is specific to this group. Besides, as we saw in the previous section, complementary types of collectives exist in Schelling's account of economic systems and in a large scope in society: the social, cultural, or ethnic categories, etc. Individuals are constituted by their membership in these multiple collective categories. Individuals' personal identity is therefore at the cross point of their multiples collective identities. We can mention that Schelling numerously refers to the influence of players' value system, or of their social role, in their decision-making. Again, this value system is both determined by players' collective identities and personal experience. In players' subjective identity there are at the same time personal dimensions that derive from their personal trajectory of life, their personal experience, and some collective dimensions because of their belonging to collectives, to communities; but the multiplicity of collectives to which they belong also participate to the creation of a personal and specific identity.

This explains the importance of psychology and not only cognitive psychology but also social psychology to understand how all of these personal or collective dimensions structure players' mind, players' reasoning and perceptions about others. This in turn explains why he is interested in the cultural and ethnic dimensions that shape players' identity.

In addition, when players cannot rely on some pre-existent focal points that are determined by conventions, institutions, etc. they build themselves, during their interactions, a device of coordination. In this condition the focal point is emergent. If players can rely on this focal point to coordinate it is precisely because they acknowledge that they both have a common purpose, that they both have to solve the game *together*. They have to bypass their own individuality to adapt to each other. To understand such statement it is worth citing, again, the following

quotation: "[t]hey must together find "rules of the game"." (ibid, p. 107) Players accordingly acknowledge that they form a transitory collective just during the resolution time of the game.

To sum up, Schelling's account of players is multidimensional. He integrates players' mental states (their perceptions and beliefs), their personal experience, and in that manner, a dimension of subjectivity that is however perfectly compatible with an intersubjective and social dimension. Players are indeed able to draw on common interactional experiences, on possibly common institutional heritage which induce common perceptions, common reasoning, to ultimately share an intersubjective background. In these cases their subjectivity can converge to understand each other, i.e. to tacitly agree. In the theory of games that will be developed in the chapter 5 we will show how such subjective dimension is compatible with an intersubjective background necessary for coordination. To portray realistic players, he introduces heterogeneity. This heterogeneity depends on the one hand on the integration of subjective dimensions like the players' perceptions, beliefs elicitation, modes of reasoning. On the other hand it depends on players' personal history and combination of collective identities.

As soon as these subjective dimensions are integrated in the realm of game theory, Schelling justifies players' capacity to reach a meeting of minds by a cognitive process endogenous to the games. And this cognitive process mainly relies on players' capacity to put themselves in the others' shoes. Players resort to social dimensions like institutions conventions etc. to coordinate, however to think of these coordination devices as being reliable, they have to consider the way the other players perceive the situation and accordingly to try to see the problem from the others' angle. They have in some respect to see themselves as belonging to a common community. In other words, the socially skilled side of the players is the *necessary counterpart* of the subjective side that Schelling takes into account. The use of the empathetic capacities and the conventions, institutions, etc., are the necessary determinants of strategic reasoning when the players are no longer like a representative agent but on the contrary subjective entities.

4.2. What is strategic rationality?

The conception of players' rationality underpinning Schelling's reorientation of game theory is very far from the standard strategic rationality in contemporary game theory. As will be developed in this section, players' rationality is neither instrumental nor Bayesian (as standardly defined). There is no way in Schelling's vision of game theory, to define deductive principles of rational decision based on an axiomatic of choices (see, Sugden and Zamarrón 2006, p. 620).

As Schelling claims,

> "There is … no way that an analyst can reproduce the whole decision process either introspectively or by an axiomatic method. There is no way to build a model for the interaction of two or more decision units, with the behavior and expectations of those decision units being derived by purely formal deduction. An analyst can deduce the decisions of a single rational mind if he knows the criteria that govern the decisions; but

he cannot infer by purely formal analysis what can pass between two centers of conciousness." (ibid, p. 163)

Instrumental rationality entails that players adopt a best reply reasoning having perfect knowledge of the situation, i.e. of whom the others are, their rationality, their available strategies, their preferences, their payoff, and of the consequences of each combination strategy profile. Bayesian rationality entails that, again, the players adopt a best reply reasoning according to their beliefs – furthermore assumed to be rational – of the others' choices and beliefs. In each case players' preferences, strategies and beliefs are exogenously given and settled before the game. In each case, providing such conditions of knowledge and beliefs, according to a set of axioms defining rationality allows to identify the set of admissible strategies for each player, i.e. the equilibrium profile of actions. However, as we have seen in this chapter, the way Schelling presents strategic interaction is very far from any of this conception of games and rationality. In Schelling's work players do not resort to an instrumental best reply reasoning. Even the concept of strategy is challenged. In standard and Bayesian game theory the whole strategy profile is settled before the game and involves the plan of action for each player for the whole game. In case of extensive games the whole plan of action for every nodes is fully specified. This is absolutely not the case for Schelling. New strategies, new plans of actions may appear and disappear during the game as some new information arises (see Dixit, 2006, pp. 216-217). Such configuration is ruled out by Bayes updating in which new information cannot contradict the prevailing system of beliefs that can only be updated with congruent information.

Yet, Schelling believes in *rational choice*.[43] He however militates for an inclusive conception of rationality. There is "no restrictions of validity on players' reasoning" to quote Sugden and Zamarrón (2006, p. 620). In Schelling's work, it seems that rationality should better be understood in terms of having reasons for acting according to specific contexts (Schelling, 1984, p. 205). This is explained by the fact, as already emphasized, that players have to adapt to each other, and they must evaluate this need to adapt to each other *in situ*, when interacting. This is also explained by the fact that the players belong to collectives, to communities, with ways of doing, ways of interacting that impinge on their apprehension of a situation, on their decisions and behavior. All of these dimensions can impact their decisions and what is judged rational to do in the specific situation in which they are embedded.

We however do not subscribe to this vision of rationality. The theory of games that we will propose in chapter 5 will be laid on a conception of strategic rationality that sensibly diverges from this interpretation. We will amend Bayesian rationality to break with some of its standard hypotheses (i.e. the ratifiability and the action-independence of players' beliefs hypotheses) that are questioning regarding the mere appraisal of strategic rationality.

Besides, when facing coordination problems, the standard best reply form of rationality in game theory leads to an infinite regression of beliefs (i.e. I believe that you believe that I believe that

[43] Schelling declares: "We will better understand the uses and limits of rational choice if we better understand those exceptions ... It is a wonderful tool if used when appropriate, but it may not work all the time. I consider myself in the rational-choice school, absolutely. But I am more interested in the exceptions than many other economists tend to be." (in Steelman, 2005, p. 38)

you believe that I believe, and so on, ad infinitum) and the incapacity for both players and theorists to provide a determinate solution (see Sugden and Zamarrón, 2006, p. 614). The focal point is precisely a methodological device to induce the selection of one outcome in case of multiple equilibrium and therefore in case of the failure of the existence of a solution in standard game theory. Nevertheless, justifying that the focal point is a rational solution in case of multiple equilibria requires some important methodological enhancement of the standard account of rationality in strategic context and in particular it requires to break with the axioms of symmetry. It is indeed impossible, in the standard framework, to justify the existence of a form of rationality which allows individuals to reach the focal point in a coordination game or in a cooperation game (see Gilbert, 1989, 1990; Colman, 1997, pp. 13-14). In reality, in a formal game theoretic apparatus there is still an infinite regression concerning players' beliefs about other players' beliefs, and there is no rational basis to firmly imply that one of the equilibrium point is the solution.

However, the way Schelling presents the focal point is that coordinating on a focal point is a rational play, i.e. the behavior implied by the focal point is a rational play. Here is the pragmatic or empirical side of his account of rationality (Innocenti, 2006; Sugden and Zamarrón, 2006). Rationality implies an effective coordination of expectations. Here lays the pragmatic dimension of Schellling's account of rationality: being rational means finding a solution, i.e. reaching a meeting of mind, and thus, correctly inferring each other's beliefs and actions. Rationality implies being successful in solving a game, whatever the mode of reasoning and belief elicitation that lead the players to such result, and whatever the information on which players rely for that purpose. As Schelling declares, "a normative theory of games, a theory of strategy, depending on intellectual coordination, has a component that is inherently empirical" (1980[1960], p. 285). And in coordination games, standard rationality leads to the inexistence of solution. That is why the principles of rational play are to be found outside the realm of mathematics and the axiomatic of choices.

> "It is an empirical question whether rational players can actually do what such a theory denies they can do and should consequently ignore the strategic principles produced by such a theory."(ibid, pp. 285-86)

A normative theory, i.e. a theory of strategic rationality, is a pragmatic theory that works in practice and that can help players to solve such coordination games (see Sugden and Zamarrón, pp. 612-618-620). And in that perspective Sugden and Zamarrón (2006, p. 619) asserts that Schelling is offering more than an "assimilassionist theory",

> "the assimilationist theories adapt classical game theory to take account of properties of games (their "labelling" or "framing") and players (their bounded rationality) that have been discovered by empirical investigation. But, we suggest, Schelling means something deeper than this. His understanding of rationality itself is empirical in the following sense: a principle of decision is rational for an agent just to the extent that, by using it, the agent tends to be successful in achieving her ends."

This is contrasting with standard game theory in which the "principle of rational determinacy" prevails (Sugden, 1991).[44] The rationale of a decision is judged *a posteriori* and by its success, i.e. by players' capacity to coordinate. There is no possibility to define a priori and deductively a rational decision in interactional systems. These statements are very far from any claim within the rational choice theory and standard modern game theory.

More concretely, neither of the standard requirements of strategic rationality prevails in Schelling's propositions to reorient game theory.

First, Schelling does not believe in the optimality of an equilibrium in a game, or on the fact that individual self-serving behavior can lead to a social desirable state (demonstrating this statement was one of the purpose of his dynamic models of residential segregation). There is no maximization or best reply reasoning in Schelling's account of strategic interactions. Players may even choose 'dominated' strategies just for signaling some intentions of being committed to a peculiar solution, for a threat, for gaining credibility or simply for respecting some values or ethics (see Zeckhauser, 1989, p. 159). This may explain why according to Ayson (2004, p. 126), Schelling's conception of rationally is comparable to Simon satisficing. Besides, the rationality of a decision depends greatly on the individual perceptions. Some options are beyond players' control even if from a game theoretic point of view those options could have been optimal. From those perceptions however, some strategies or elements of the context of the game or of the play itself have a symbolic content, and from this symbolic content a focal point can emerge. In this condition, a game can be solved. For Schelling, "[t]he assertion … is *not* that people simply *are* affected by symbolic details but that they *should* be for the purpose of correct play", i.e. a rational play (ibid, p. 98).

Second players' preferences are neither exogenous to the decision problem nor stable (Innocenti, 2007, p. 423).[45] Those preferences are explained by the game. They evolve with the interaction process while some opportunities appear and others disappear. As soon as Schelling asserts that solving a game is a discovering process, players' preferences are built during the decision process. They are defined and evolving during interactions. Recall that the players must find the rule of the game while playing (i.e. the set of beliefs, preferences and payoffs). Since rule following and its associated mode of reasoning, community-based reasoning, is constitutive of practices, i.e. has the power to change the game they are playing, the norms that are followed condition the players' preferences, utility functions and eventually, beliefs. Collective belonging therefore impinges on the rationality of a decision. The 'rules-as-normative-expectations' (Bicchieri, 2005) – a specific form of rule following – induce "a conditional preference for conformity entering directly into the agents' utility functions. As a consequence, their preferences are a function of their beliefs, and hence of the norms" (Hédoin, 2015, p.16). The other form of rule-following induced by 'rules-as-correlated-devices' set the stage for the fact that the agents have a common prior and lead them to behave in a particular way. In other words, the norm defines a specific epistemic

---

[44] "The principle of rational determinacy" states that there is only one way to decide and to act in all of the possible circumstances, e.g. in all of the types of game.

[45] This is also true for Innocenti when referring to decision theory and the multiple selves account of the individual.

game and a change in the norm would change not only the agents' behavior but also the whole characteristics of the epistemic game (starting with the agents' prior). (ibid, p. 17) It therefore defines the players' beliefs. Symmetrical beliefs as implied by rule following must be connected with the mindshaping approach that will be developed in the chapter 4 of the thesis. Besides those preferences are also determined by the context of the game and the other players. The eventual consequences that a player has to compare and order cannot be purely subjective and independent of the context of the game and the other. As Schelling (1984, p. 205) claims "[t]he critical question is not whether a person is "rational" according to any particular definition, perfectionist or merely approximate, but whether his choice is determined in large part by the situation he is in."

As previously stated players draw their reasoning, their expectations, both on the objective and the institutional environment surrounding the game. Players' reasons for acting are partly determined by this multidimensional environment, they are no longer intrinsic to the individuals (e.g. see Innocenti, 2007, p. 416). For instance, it can be more rational for an individual to rely on conventions than to follow what could be identified as her preferences, even if following these institutional devices will discriminate against her (see Schelling, 1980[1960], p. 100). The standard conception of individual rationality in terms of consistency of choices cannot prevail. Players' preferences are not context independent. Besides what is rational to do for players depends to a great extent on the others with whom they interact, on what they perceived about these others and their intentions. It implies that conversely to standard game theory the other is no longer "naturalized" (see Hargreaves Heap, Varoufakis, 2004[1995], p. 37; Lesourne et al, 2006; Mariotti, 1995). The other is not considered as an event of nature as in both complete and incomplete information game theory. The other's identity, mental schemes and mental states, perceptions, feelings, and intentions all matter.

Players have therefore the faculty of reflexivity: they can "reflect on their mode of reasoning" (Hédoin, 2014, p. 381). They can reflect on their beliefs and those of their co-player. For instance, as emphasized by Orléan (2004) players are able to distance from their 'personal' belief to ascribe to what they guess to be the beliefs of the collective. He considers that in Schelling's work, group beliefs are possibly disconnected from the beliefs that the individual inside those collective believe (Orléan, 2004, p. 200): "Each individual can believe P and, at the same time, believe that the group believe Q" (ibid, p. 199). Such account does not correspond to a strict methodological individualism as the group beliefs or collective beliefs are no longer the aggregate of the individual beliefs (ibid, p. 211).

Third, contrary to standard and contemporary game theory, for Schelling rationality does not mean symmetry. For him being rational does not imply following identical modes of reasoning, or paying attention to the same information, or having identical beliefs from the same information (like it is implicitly supposed with the common prior assumption for instance). His work breaks with the symmetry assumption, i.e. players' homogeneity, imposed by Nash, and then by Harsanyi and Aumann. For the latter, if players have different beliefs it is because they have different informati**on**. But if they are rational and an asymmetry of information exists the players will learn the new information because players cannot agree to disagree (cf. the Harsanyi-Aumann doctrine). So that if they know that they are disagreeing with each other they know that some information is missing and they reassess their beliefs provided the new information.

Rationality understood in such perspective also means that confronted to a common decision problem, a common situation, players would have to have the same preferences. Schelling is extremely critical of Harsanyi (1956) who states symmetry as a 'fundamental postulate' of rationality: he quotes him declaring that "[b]argaining parties follow identical (symmetric) rules of behavior (whether because they follow the same principle of rational behavior or because they are subject to the same psychological law)." (Harsanyi, 1956, p. 149) Therefore, Schelling claims

> "the justification for symmetry postulate has not been just that it leads to nice results; it has been justified on grounds that the contradiction of symmetry would tend to contradict the rationality of the two players. This is the underpinning that I want to attack …What I am going to argue is that, though symmetry is consistent with the rationality of the players, it cannot be demonstrated that asymmetry is inconsistent with their rationality, while the inclusion of symmetry in the *definition* of rationality begs the question." (Schelling, 1980[1960], p. 278)

There can be symmetry "in move structure" because ultimately the players tend to a common solution, however there cannot be symmetry "in the configuration of payoffs" (Schelling, [1960]1980, p. 268), this would mean that players have identical preferences. As a consequence Schelling claims that "symmetry in the solution … cannot be supported on the notion of rational expectations" (ibid, p. 267). Payoffs are not necessarily symmetrical even when players have identical interest. Such identical interest cannot be translated to common payoff. They are just a matter of solving a game, of finding a solution; even if such solution entails different individual payoffs.

However, the concept of symmetrical beliefs prevails with respect to Schelling's account of game solutions. Since the outcome exists when there is between the players a meeting of minds, it requires the convergence of individual subjective beliefs. This symmetry is *ex post*. Different individual modes of reasoning, different preferences, can lead to such symmetry of beliefs; different initial subjective beliefs, can ultimately converge. There is no unique rational way to come to such symmetry of beliefs.

A last aspect underscored in Schelling's account of game theory is the collective dimension of rationality.[46] The meeting of minds, i.e. the convergence of players' modes of reasoning, and common understanding takes place inside a given community: "the basis for common understanding is community membership" (Hédoin, 2014, p. 384). Players' reasoning is therefore determined by their collective identities and no longer intrinsic preferences, or beliefs. In the same way, rule following is linked to the belonging of a given community (Hédoin, 2017, pp. 60–1). It therefore entails a "non-individualistic conception of salience" (Hédoin, 2014, p. 366) which contradicts methodological individualism. Due to the strong interdependency of players in any type of games, the set of players become a collective entity. Players form a collective in the

[46] "Collective rationality" refers to several types of literature. In philosophy to Hodgson (1967), Regan (1980), Gilbert (1989, 1996, 2000, 2006, 2013) Hurley (1989, 1998), and Hollis (1998). It is also closely related to collective intentionality (Tuomela, 1995, 2000, 2002, 2007, 2013; Tuomela and Miller, 1988; Searle,,1995, 1998, 2010; Bratman, 1993). And in game theory to the "team reasoning", i.e. to the work of Sugden (1993, 2000, 2003, 2007) and Bacharach (1995, 1997, 1999, 2001a, 2001b, 2006).

sense that they have a common purpose which is solving the game and this requires a certain proclivity for cooperative behavior.[47] This is true that players do not become a team in a restrictive sense. They do not decide by a common and explicit agreement to become a "plural subject" in the logic of Gilbert (1989, 1996, 2000, 2006, 2013) or a "team" as conceptualized by Sugden (1993, 2000, 2007) or Bacharach (1999, 2006) in game theory. They become a collective entity by virtue of the problem they have to solve. Players recognize that they face together a common problem and that they will be able to solve it only by their mutual acknowledgement of their strong interdependency. By virtue of the situation, i.e., of the collective decision process they face, players know that they will have to learn together in a very unique and specific way, how collectively they can manage to coordinate. This aspect tends to argue in favor of collective rationality. Players become a unique and transitory "team" during the time of the resolution process, they adopt as a consequence, a distinctive mode of reasoning which is driven by the collective they form. It does not mean that players' purposes or 'preferences' are deduced from those of the collective. It means that the way players decide and act is defined by their interaction and is specific to whom they are. In this sense this is a collective form rationality.

4.3. Epistemological implications regarding the status of theories and models

Considering Schelling's enhancement of standard game theory, considering the type of modeling he proposes in the dynamic model of residential segregation, considering the social ontology behind his conception of interdependence, considering the way he portrays the economic agents in his games and considering his conception of individual rationality in interaction, his epistemological stances seem particularly challenging. Considerable questions arise about the way Schelling apprehends the status of theories and models in economics, and in particular the link between economics and the other social sciences and the role he gives to the experimental method.

It must be noticed that understanding economic and social phenomena requires at some point the use of model and theories for Schelling, even if the scientists must keep in mind that their models are just abstract representations of reality so that they must be extremely cautious when dealing with the conclusions they draw from their models and theories.

Schelling claims that by being so attached to the universe of mathematics and being too abstract, the validity of the conclusions derived from economics and game theory may be endangered

> "Actually economists do not usually make careful observations, compare what they observe with alternatives they can imagine, and judge the results to be good. What they do is to infer, from what they take to be the behavior characteristics of the people, some

[47] Sugden and Zamarrón (2006, 614) point out this 'we' dimension in coordination problems that can be understood as a collective agency. However they assert that this does not involve team reasoning (ibid, p. 615).

> [of the characteristics of the system as a whole, and deduce some evaluative conclusions." (Schelling, 2006 [1978], p. 23)

When confronted to real-world system that is an open-system such deduction as Schelling emphasizes is impossible (see Chick, 2004).

> "Game theory run the same danger as any theory in being too abstract, even in the propensity of theorist to forget, when they try to predict or to prescribe, that all their theory was based on some abstract premises whose relevance needs to be confirmed" (Schelling, 1984, p. 240).

By considering game theory as a 'framework for analysis' and grasping its essence, Schelling attempted to build his own theoretical system, adaptable for the analysis of specific and real matters. As already underlined, for him, "game theory is more than a "theory", more than a set of theorems and solutions; it is a framework for analysis. And for a social scientist the framework can be useful in the development of his own theory." (Schelling, 1984, pp. 221-222) For Schelling "a model is a tool; to be useful, it has to be adjustable or to consist of a set from which we can select the appropriate member." (ibid, p. 90) Schelling has selected in game theory the concept of interdependence of individual choice and behavior and the way game theory analyzes and represents this interdependence of choices and behaviors. He grasped this concept and then developed his framework to incorporate within this theory the underlying mechanism of strategic interactions.

> "I have mentioned only the rudiments of game theory, and none of the subtle or elaborate analysis that has attracted the attention of mathematics. But what may be of most interest to a social scientist is these rudiments. The rudiments can help him to make his own theory, and make it in relation to the particular problems that interest him." (Schelling, 1984, p. 221)

The use of this abstract framework and modeling is however necessary especially when dealing about social reality and the interdependence of individuals' choices and action. Those interactions involve complex phenomena that cannot be analyzed and understood without the help of a mediating model simplifying such complexity. For him,

> "[s]implified models of artificial situations can be offered for either of two purposes. One is ambitious: these are "basic models" - first approximations that can be elaborated to simulate with higher fidelity the real situations we want to examine. The second is modest: whether or not these models constitute a "starting set" on which better approximations can be built, they illustrate the kind of analysis that is needed, some of the phenomena to be anticipated, and some of the questions worth asking. As we add dimensions to the model, and the model becomes more particular, we can be less confident that our model is something we shall ever want to examine. And after a certain amount of heuristic experiments with building blocks, it becomes more productive to identify the actual characteristics of the phenomena we want to study, rather than to explore general properties of self-sorting on a continuous variable." (Schelling, 2006 [1978], pp. 183-84)

According to us, the reorientation of game theory that Schelling offers and the dynamic models of residential segregation constitute a starting point. They both offer new conceptual and analytical framework for dealing with strategic and social interactions and in particular offer some of the fundaments of strategic interactions that were not investigated and are by the way still not investigated by game theorists.

That is why Schelling is extremely inclined to the use of experimentation (Zeckhauser, 1989, p. 158; Ayson, 2004; Innocenti, 2005, p. 7; 2007, p. 410; Aydinonat, 2005; Rivzi, 2006; Colman, 2006; Dixit, 2006; Sugden and Zamarrón, 2006, pp. 612-13; Sugden 2000, 2009). In addition to his well known informal experiments, in the chapter 1 of the *Strategy of Conflict* laying down the concept of focal point, Schelling dedicates a full chapter (the Chapter n°6) to propose some experimental methods and some experimental designs to study the influence of the interaction process on the outcomes of games and to test the influence of the elements he incorporates in his strategic analysis. According to Schelling, his experimental design would "give an operational representation of the theoretical system that the author has in mind in referring to the convergence of expectations and to suggest that the convergence that ultimately occurs in a bargaining process may depend on the dynamics of the process itself and not solely on the a priori data of the game." (ibid, p. 111)

Such experimental design could give answers, according to Schelling, to the following list of questions:

> "Is a stable, efficient outcome more likely between two players of similar temperament and cultural background or between two quite different players? Is a stable, efficient solution more likely with two practiced players, two novices, or one novice and a practiced player; and in the latter pair, who has the advantage? In a game of this sort, how crucial are the opening moves? If stable patterns of behavior, that is, "rules of the game", are not discovered early, will they be discovered at all? Is mutually successful play more likely if the general philosophy of each player is to begin with "tight" rules or highly "limited" weapons and resources, loosening them a little only as the occasion demands it, or if each player sets himself wider limits at the outset in order to avoid to establish a practice of loosening rules as he goes?" (ibid, p. 167)

The interest that Schelling puts on experimental methods expresses his conception of game theory and economics as an open-system. He helps delineate the scope of exchange that are needed to appraise all the dimensions of strategic interaction that lay outside economics and even more outside the realm of mathematics. His use of experimentation and the importance of experimentation he sees in theorization show how he distances himself from a pure mathematical conception of games and the need for more empiricism in games. It is a way to assess the need to refer to other social sciences such as psychology, sociology, law, history etc. In the elements that Schelling purports to investigate, as emphasized in the quotation and in the chapter more generally, some of them rely on social facts on social dimensions that are not traditionally investigated and integrated in the realm of game theory. This explains on the one hand why Schelling questions the role of methodological individualism in economics and game theory, and on the other hand why economics should not be appraised independently from the other social sciences.

> "A pervasive question for social phenomena is the role, or the exclusive role, of "methodological individualism", the notion that the ultimate analysis is a rational, or at least a purposive, individual. Some believe that any social phenomenon that cannot be reduced to the behavior (choices) of individuals is a black box and therefore unsatisfactory. There is some notion that what is inside a black box must be social mechanism, or several social mechanisms." (Schelling, 2006, p. 235)

Such statement is true for the model of residential segregation and for game theory as demonstrated in the previous sections. Some of the social mechanisms to which Schelling refers traditionally belong to the realm of sociology. In addition, they interfere in the psychology of the players, in the cognitive mechanism leading to the players' meeting of minds. In that perspective, the role of institutions or conventions in determining a focal point can be mentioned. For instance, Schelling mentions that players' tacit agreement are in someway "based on something psychologically and sociologically akin to *tradition*" (ibid, p. 168). That is why economics need the other social sciences. Schelling's inclination of interdisciplinarity is widely acknowledged (see Ayson, 2004; Innocenti, 2005, pp. 7-8).

Schelling often refers to the psychological phenomenon of coordination of expectations. When setting his experimental design he refers to the "robbers cave experiments" largely used in psychology and especially in social psychology and in that perspective he mentions the work of Bavelas (1950) and Sherif (1958). Schelling also refers to Gestalt psychology in his writings. Schelling emphasizes how through his work, "the question arises whether the game theory trail ramifies indefinitely over the whole domain of social psychology or leads into a more limited area particularly congenial to game theory." (ibid, p. 165)

Experimental psychology could in particular play a great role for the enhancement of game theory. Schelling indeed declares:

> "This process, by which particular moves in a game or offers and concessions achieve symbolic importance as indicators of where expectations should converge in the rest of the game seems to be an area in which experimental psychology can contribute to game theory." (ibid, p. 113)

The most evident links in Schelling's framework are with psychology and sociology, however Ayson argues that Schelling's interdisciplinary approach does not confine to these two disciplines. Ayson underscores "the importance of aspects of contemporary theory from other social sciences with which Schelling was familiar, including social psychology, communication and organization theory. It demonstrates how these theories set the context for Schelling's interest in the coordination of expectations in group situations, in a dynamic conception of stability based on an organic idea of 'feedback' as means of understanding stability" (Ayson, 2004, p. 8); he also mentions that Schelling's works "encompass such areas of study as Gestalt and social psychology, group dynamics, information theory, all of which Schelling uses to understand the prospects for stability in processes of interaction" (Ayson, 2004, pp. 160-61). As already emphasized this pleads for a conception of economics and game theory as an open-system. This is rather clear in the following quotation he appeals to bring interdisciplinarity in economics:

> We can only hope there is some theory capable of dealing with the multiplicity of phenomena that disturb our simpler theoretically established regularities. This additional theory may already exist as part of another discipline, so that it does not make part of the existing background knowledge of the economist. It may also be completely absent. It seems that interfield theorizing is needed to deal with policy relevant phenomena that escape our explanatory hypotheses and their clauses that serve to hedge them. (Schelling, 2006a, pp. 90-91)

5. Conclusion

As we emphasized in the introduction of the chapter Schelling provides very fruitful conceptual and methodological answers to some of the major difficulties that contemporary game theory faces. However, he does not propose theoretic enhancement, i.e. formal solutions, to bypass these difficulties, at least within an analytical game theoretic framework.

The next chapter is however an attempt to show, through the work of Bacharach, that most of the Schelling's insightful propositions can in fact be incorporated in a game theoretic framework.

The methodological literature on Schelling's thinking generally explains the lack of progress made by using some of the major insights made by Schelling because "psychological phenomena and influences based on cultural factors are difficult to formalize" (Rivzi, 2007, p. 405), or because Schelling's methodology is too far from or even disconnected from the standard game theoretic one and probably because game theorists are reluctant to use different mathematical tool (they still privilege maximization, EUT, Bayesian decision theory, etc.).

> "In Strategy of Conflict, Schelling gives an account of how people coordinate, and of how to go about coordinating with others, which the total drift of thinking continues to confirm. In developing this account, he uses research tools – experiment, inductive generalisation, imaginative introspection, metaphor – which evidently work, but which are different from those normally used in game theory; and he explains why these are the most useful tools for the project in hand. However, game theorists have persisted in using their customary research tools; and they have been continually frustrated by their inability to create a theory of focal points. Schelling's "vision", his proposal for the reorientation of game theory, remains unrealised. The process of realising it has hardly begun." (Sugden and Zamarrón, 2006, p. 620)

The next chapter is an attempt to prove the contrary. Drawing on the example Schelling legs we will show that in fact, the psychological and sociological side of Schelling's work can be formalized.

As a conclusion, we wish to explain why we choose to stick to the game theoretic side and do not attempt to integrate the most heterodox insights of Schelling into game theory.

First, because the psychological explanation of the elicitation of player's beliefs provides a methodological ground to overcome some of the main shortcomings of epistemic game theory that still lacks fruitful ideas and solutions. Schelling's reorientation of game theory provides a methodological ground to endogenize the players' beliefs in games and provides an explanation of such elicitation while most contemporary game theorists remain silent. The Chapter 5 will be an example of this incorporation.

Second, his work provides again fruitful insights to compensate the integration of subjective dimensions in players' reasoning (mainly the individuals' perceptions and expectations, furthermore endogenous to the game) through an intersubjective dimension, i.e. through players' capacity of empathy. Again, the chapter 5 will be an example of how to make compatible these two dimensions thanks to Goldman's simulation theory (which will be extensively exposed in the chapter 4).

Third his explanation of the selection of equilibrium among multiple equilibria is still the only serious attempt to overcome the problem of indeterminacy that is still unresolved even in contemporary standard game theory (except in evolutionary game theory). No analysis of strategic reasoning (which is not the case in evolutionary game theory) provides a convincing rational basis for the existence of a solution when multiplicity of equilibrium exists.

And finally because Schelling's "insights have not been fully developed yet" (Innocenti, 2005, p. 6); "little to no progress has been made in exploring Schelling's insights" (Kreps, 1990, p. 101). Similar claims have been made by Crawford (1991), Binmore Osborne and Rubinstein (1992), Bacharach (1991, 1993, 2006), Myerson (2001), Sugden (2001). His work is obviously widely known, his work on focal point often cited however Schelling's insight are rather misunderstood (Colman, 2006).

Only certain parts of his work are known and used but not the whole picture of his propositions, not the whole picture that we described in this chapter. And we claim that the major enhancement that can be done with Schelling's work within game theory, requires this broad picture. Especially considering that some analytical solutions, as we will show in the next chapter and in the chapter 5 could integrate Schelling's innovation. There are still very few attempts to integrate the focal point concept in a game theoretic framework; we can only cite six scholars in the history of game theory (Sugden, 1995; Casajus, 1998; Janssen, 2001, 2006). And some of them leave out some of the most important dimensions of focal point because they do not see the whole picture and miss some of the major innovative propositions that Schelling made (excepting Bacharach, 1993, 1997, 2006).

Chapter 3

Bacharach: How the Variable Frame and Team Reasoning Theories challenge standard non-cooperative game theory

Michael Bacharach's contributions are of interest for this thesis in so far as he integrates many of the dimensions of Schelling's reorientation of game theory. In particular he is, like Schelling, very much concerned by a strong interdisciplinarity in his methodology. Furthermore, he was at the same time very suspicious of a strictly mathematically centered game theory by pointing out its limits, and nevertheless contributed to enrich game theory from a theoretical and modeling point of view. Like Schelling, he showed the limits of standard game theory (both classical and epistemic game theory) and questions the role of common knowledge or common belief of rationality. In that perspective he integrates in games another intersubjective dimension that relies, as for Schelling's focal point, on both subjective and collective dimensions. Indeed, in Bacharach's contributions, focal points can rely both on subjective and personal apprehension of a game situation and on conventions or institutions that entail rule following behaviors (see chapter 2). This chapter will ultimately show that Schelling and Bacharach exhibit a common social philosophy centered on the role of focal points for coordination. A game is for both of them a process in which the players' apprehension of the strategic situation they face and their strategic reasoning is central for explaining the solution. And above all Bacharach suggests by his theoretical and formal contribution, potential theoretical solution for building a game theory based on a new form of intersubjectivity with the integration of players' mental states without the difficulties of standard and epistemic game theory underlined in the chapter 1 (with respect to common knowledge or common belief of rationality and the status of players' prior beliefs).

1. M. Bacharach: an interdisciplinary fellow

M. Bacharach studied mathematics and economics at Trinity College and graduated in 1959 in Economics. He went to Stanford for studying econometrics (1958-59) and then returned to

Cambridge in the Department of Applied Economics to complete a doctorate (1959-1965). He became Junior Research Fellow at Nuffield College (from 1966 till 1967) and then Temporary Teaching Fellow at Balliol College (from 1967 till 1969). He finally came to Christ Church as a fellow, became University Lecturer, and remained at Christ Church for the rest of his career. He was also a Research Associate of the Institute of Economics and Statistics from 1991.

M. Bacharach's Ph.D was on the mathematics of input-output analysis and was an attempt to extrapolate input-output matrices to structural and technological change. He saw the interdependence among the inputs of the producing sectors of an economy and each other's output, i.e. of the interdependence of industries' technologies for economic growth. He applied his analytical work for practical matters. For instance, he was involved in the 1960s, in the Cambridge Growth Project (under the direction of R. Stone and J. Brown) to work on inter-industry relations and economic planning. He was also involved, as a consultant, in the Food and Agriculture Organization in Rome (in 1965), in the Economic Commission for Europe (in 1967 and 1970) and in the UK Atomic Energy Authority (from 1975 till 1977).

M. Bacharach worked on a wide range of topics in economics: starting from input-output matrices, to the theory of economic planning and policy-making, rational expectations, decision theory, rational choice theory and Bayesian decision theory, economics of information, the methodology of economics and the foundations of game theory, the common knowledge hypothesis and its validity in game theory, semantic and syntactic logic, experimental economics, , and coordination and cooperation in game theory.

He moved away from the input-output matrices, and became interested instead in the interdependence among individual economic agents in their choices and behaviors, in markets for instance. He became very critical towards general equilibrium theory (see Bacharach, 1990) and in general towards economic theory (see Bacharach 1986, 1989; see also Arena and Larrouy, 2016), and then more specifically toward decision theory and game theory (e.g. see Bacharach 1987, 1992; Bacharach and Hurley, 1993). He appraised economic theory from an interdisciplinary perspective quite early. He draws on such interdisciplinarity to substantiate his criticism of a too mathematical and closed-system account of economics and later of game theory.

Bacharach criticized the assumption of Rational Expectations (in Bacharach, 1989), investigated the foundational shortcomings of decision theory and game theory (in Bacharach, 1979, 1985, 1987, 1992, 1994, 1997, and Bacharach and Hurley, 1993) using formal methods from economics, philosophy and logic. He challenged the epistemic foundations of choice theory and game theory in particular. He introduced new tools to analyze common knowledge such as modal logic and the underlying knowledge operators, or syntactic logic. Bacharach was one of the pioneers in the introduction of modal and syntactic logic in game theory before they became at the core of the analysis of epistemic problems in game theory.

Therefore after publishing a textbook in game theory in 1976, *Economics and the Theory of Games*, while game theory was in complete reconstruction and facing a deep turmoil for rethinking its mathematical foundations, when game theory finally became the core of standard microeconomics, Bacharach was already assessing its philosophical foundations, attempting to

improve even its new assumptions. As Sugden (in his memorial address) declares, "Michael was well ahead of his time in seeing the significance of game theory for economics. But from the outset, he saw that there were deep problems in the theory."

Recall that when Bacharach began to work on game theory, the context was that of the refinement program. Bacharach however addressed the problem of indeterminacy which led to the refinement program with a methodological and epistemological perspective while most of the contributions were focused on the production of new mathematical conditions and new mathematical restrictions on the solution concept. He did not contribute to the refinement program *per se*; but the effervescence of contribution that this program led and the deep questioning it could have generated on the nature of strategic rationality drove Bacharach's contribution. This research program generated methodological debates about the foundations of game theory among a narrow set of game theorists. The fact is that game theory was still considered to be too mathematized and too formalized, i.e. too abstract, without any "possible applications to concrete socio-economic problems" (Giocoli, 2003, p. 348). In addition, in a number of economic situations the strong epistemic assumptions imposed in the standard complete and then incomplete information games was insufficient to lead to a determinate solution such as in games where two Nash equilibria exist or where no strictly dominant strategies exist. And the fact is that, by the application of game theory to economic cases in the 1980s, situations in which there are multiple equilibria appeared to be more frequent than situations in which there is one unique solution (Sugden and Gold, in Bacharach, 2006, p. xv). For this kind of situation, Nash equilibrium had to be refined in order to identify one equilibrium as "the rational way to play" (Bicchieri, 1993, p. 64). As it is now acknowledged, the major part of contributions led to a negative conclusion in the late 1980s – all the new efforts to justify the validity of the Nash equilibrium as a *unique* solution were not successful. The refinement program has failed in this task (Sugden and Gold, 2006; Giocoli, 2003). As we will explain in the next section, Bacharach contributed to demonstrate that the faith putted in the Nash equilibrium as the only valid solution concept was misleading.

Thus, when in the late 1970s and in the 1980s, game theorists recurrently faced problems related to the "internal consistency" of classical game theory's assumptions and "the indeterminacy of its predictions" (Gold, Sugden, 2006, p. xiii). They had to deal with the problems (i) of equilibrium selection, i.e. how to select one equilibrium in cases of multiplicity and (ii) of rationality, i.e. of what a rational player ought to do in games. Bacharach proposed solutions to answers these two problems in a context in which "finding coherent foundations for game theory was seen as one of the most important theoretical projects in economics." (ibid, p. xiv) He drew on an interdisciplinary approach, relying on philosophy, psychology, sociology, and evolutionary biology, to propose "new coherent foundations" for decision making first in decision theory and then in game theory. Bacharach, in that perspective, showed that game theory should not be a purely mathematical theory but on the contrary open to the other social sciences, as players, in their strategic reasoning, rely on their perceptions so that their personal and social experience matter in the resolution process of a game.

Bacharach became more and more acquainted with disciplines at the frontier of economics such as psychology, sociology, or philosophy, or logic. His criticism of the orthodoxy is rooted in this interdisciplinary knowledge. While running a seminar on Mathematical Economics (with J. Enos

and M.Dempster) in the 1970s at Oxford, and being a founding member of the Game Theory Society, he started managing interdisciplinary seminars (two of the numerous visiting speaker were for instance Susan Hurley and Margaret Gilbert). He founded a group of thinking around the concept of rationality and the rational choice theory with scholars coming from every discipline in humanities (personal conversation with Sugden).

> "Economists tend to be resistant to the idea that other academic disciplines might have something to contribute to the understanding of economic issues. Not Michael. In developing his theories, he drew on ideas, discoveries and methodologies from an astonishing range of disciplines … Among academics, there is a tendency to look down on interdisciplinary research as unstructured or unrigorous. But no one could ever accuse Michael of lack of rigour. Rigour was his trademark." (Sugden's memorial address)

While being trained in mathematics and being particularly inclined to use formal models he purported to extend his analytical methodology beyond economics to provide a better understanding of economic phenomena and in particular of individual decisions. He was pushed by "the positivist urge to explain and therefore understand how individuals function in a social context." (Colman's memorial address) That is why Bacharach was unsatisfied with the standard game theoretic account of strategic interactions and why he spent the last 20 years of his life and carrier to enhance the philosophical foundations of game theory, laying new analytical foundations compatible with a revised philosophy of game theory.

This is in this perspective that Bacharach developed the Variable frame theory (VFT) in the 1990s. He attempted to incorporate in game theory psychological assumptions and to take into account the importance of the way players themselves frame their decision problem but without sacrificing the game theoretic formal rigor. He saw early how understanding and explaining coordination in games requires the integration of players' mental states and reasoning process. He is one of the first to see the strong link between game theory and psychology (and not only to experimental psychology like Kanhneman and Tversky's work, which was the main reference of psychology for economists in decision theory and for a long time even later after Bacharach's contribution). Unlike others, he refers to many frameworks in psychology, many theories and not only experiments on individual decision-making. In the development of the VFT, Bacharach however attached a considerable importance to the experimental method. He indeed conducted experiments to calibrate and test his theory (see Bacharach and Bernasconi, 1997; and Bacharach and Stahl, 2000). In such context, in 1991, he created the Bounded Rationality in Economic Behaviour Unit (BREB) to develop the experimental methods in economics, to provide empirical observations of individual behavior. Experimental methods are of particular importance in his research program to orient the enrichments of game theory. Experimental methods help to circumscribe the extent of the integration of new determinants coming from cognitive and social psychology in players' strategic reasoning for instance.

In line with the empirical evidences showing that individuals more successfully coordinate and are more cooperative than standard economic theory predicts, he thought that in game contexts, individuals might frame their decision problem not in terms of "what should I do?", but rather in terms of "what should we do?". Interdependence and common interests in coordination and cooperation contexts may enhance this switch of reasoning mode. This lays the philosophical

ground of his Team Reasoning theory (TR) and he drew on the methodological foundations of the VFT to develop this theory. TR is indeed a mean to justify coordination (Sugden 1991, 2000; Hollis, 1998; Janssen 2001; Gold and Sugden, 2007b; Gold, 2018; Guala, 2018; Lecouteux, 2018). He thus opens an onward strong dialogue between game theory and philosophy (especially in the field of social ontology) with respect to the link between collective intentions and TR and the nature of collectives, their mode of existence, their principle of choice and rationality.

Bacharach was working the last ten years before his death to publish a book to unify these research interests to show how they combine in a "unified research programme with clear long-term aims" (Sugden's memorial address). This was the posthumous book *Framing and Agency: Extended Game Theory*, published in 2006 by R. Sugden and N. Gold.

2. Setting the epistemological ground for Bacharach's contribution to game theory

The seeds for Bacharach's contribution to non-cooperative game theory thanks to the VFT and TR theory are to be found early in his thinking on market exchange and market interactions, on his criticism of the standard economic appraisal of agents' beliefs which are central for him in the realm of economic analysis and the functioning of the economic world. In Brief, Bacharach saw market interactions as decentralized and as strategic interactions. Markets are for him networks of interpersonal and possibly long term interactions in which experience, or trust for instance, matter. Information about goods and transactions is not specifically conveyed by prices but by people's perceptions. In other words, economics relations are first and foremost explained and circumscribed by subjective mental states like desires, perceptions or beliefs.

2.1. On the importance of the individual economic agents' perceptions

Quite early, Bacharach saw the importance of individuals' subjective perceptions for economic analysis and before his enhancement of standard non-cooperative game theory thanks to the VFT. The starting point of Bacharach's thinking on these subjective perceptions stems from his assessment of the standard account of agents' beliefs in economic theory and his analysis of consumer theory and market exchange. In this perspective, it is important to note that Bacharach conceives market exchange and in particular buyers and sellers' interactions like strategic interactions. This concern is deeply original for market theory in which no interpersonal interaction exists (see Arena and Larrouy, 2016). It also means that economic theory and markets are, contrary to market theory and General Equilibrium Theory (GET), open-systems. The history of interactions, trust, fidelity, reputation, collective and individual customs, and so on, all impinge on the occurrence of market transactions. All of these elements impinge on the way potential buyers perceive the goods they plan to purchase and the way such goods may satisfy

their desires. This sets the urge for the introduction of psychology and sociology in those theories and accordingly the urge for interdisciplinarity for understanding market functioning.

In Bacharach's thought: the rationality of a decision and a behavior must be appraised by accounting for individuals' perceptions. Individuals' knowledge and beliefs are understood only within the frame of their perceptions. We see in this assertion the fracture with Rational Choice Theory (RCT) as used in GET where a rational choice is simply the representation of choices in which such psychological determinants like perceptions are discarded from the analysis. It thus opens the door to psychological theories and a deep thinking on the role of perceptions in individual decision-making in a way that goes beyond the mere use of experimental results in psychology such as Kahneman and Tversky's (1979, 1986). We will see that Bacharach indeed cross many approaches in psychology to circumscribe his conception of individual decision-making. This concern deeply drives all of the theoretic contributions that Bacharach offers from the 1980s onward. We see how Bacharach's work echoes Schelling's with respect to the interdisciplinary and open–system dimensions in theorizing.

According to Bacharach, explaining market exchanges requires to emphasize (i) that people are motivated by desires for goods and more specifically by "notional desires" and (ii) that these "notional desires" are the actual triggers of market exchange. In this way, Bacharach (1990, pp. 346-347) opens the door to the introduction of subjective perceptions in consumer theory. It also opens the doors for his interdisciplinary thinking in which economics and psychology are strongly related.[48] Bacharach (ibid, p. 387) refers to "the psychological fact of the notionality of our desires." Such 'notionality' of desires means for Bacharach that consumers are driven by the perceptions they have about the way goods may satisfy their desires. Bacharach thus breaks with the mathematical conception of the equilibrium understood in the GET entailed by the consistency view of rationality and which is merely assumed a state of consistency of individual choices without explaining how and why the equilibrium is reached (see the introduction and chapter 1 of this thesis). In his approach all of the types of goods are 'experienced goods', consumers experience satisfaction only when consuming goods. Seeing goods as experience goods is extraneous to GET. It indeed implies that past consumptions influence actual consumption and therefore that history, habits, or cultural determinants matter. Before purchasing, consumers are motivated by the perceptions they have about the eventual satisfaction drawn by consumption which is partly determined by their experience of market and past consumptions but also, as emphasized above, by possibly many other social or cultural determinants**.** In Bacharach' thinking, the individual perceptions can be grasped by the concept of frames.[49] Individuals' frames shape their representation of commodities

48 Bacharach's interdisciplinarity does not end there, he will be strongly involved in the development of philosophical thinking in economics but this will come a bit later in his carrier.

49 Bacharach's explicit reference to frames will come later: in his work on the VFT. Referring to frames and not perceptions however does not distort this contribution as perceptions and frames are equivalent in Bacharach's thinking (if we left aside the formal aspect of framing that is developed in the VFT). Referring to frames for this contribution helps to see how Bacharach's concern for the impact of perceptions in individual decision-making began early in his carrier and that there is a common underlying epistemology in his work.

> "[I]n order to explain how someone acts, we have to take account of the representation or model of her situation that she is using as she thinks what to do. This model varies with the cognitive frame in which she does her thinking. Her frame stands to her thoughts as a set of axes does to a graph; it circumscribes the thoughts that are logically possible for her (not ever but at the time). In a decision problem, everything is up for framing. The preferences on which she acts, her alternatives … So far from finding herself with given preferences over outcomes, as traditional theory holds self-evident, these preferences depend upon the evaluative concepts that are uppermost in her mind." (Bacharach, in Gold and Sugden, 2006, p. 69)

Therefore, commodities are 'things' or 'objects' under description (Bacharach, 1990, p. 351), and description necessarily relies on language:[50]

> "Commodities depend for their existence on words to express the concepts under which fall consumers' desires. Call the set of the community's verbalized concepts its conceptual repertoire." (Bacharach, 1990, p. 366)

Individuals' frames, i.e. individuals' conceptual repertoire is based on their "everyday experience", e.g. on their experience of effective consumptions, of market functioning, of trade, of strategic or social interactions on markets, etc. This everyday experience in turn shapes individuals' "everyday theory", i.e. the way they theorize the world. Each context induces the (involuntary) instantiation of a set of pre-existing concepts (or families of concepts), (e.g. see Scazzieri, 2008, p. 197; 2011).

> "It is reasonable to think that framing would primarily be associated with the cognitive and linguistic ability to grasp specific problem situations through the activation of particular set of 'naturally connected' features."(Scazzieri, 2008, pp. 196-97)

A stock of concepts is progressively stored according to individuals' everyday experience and this in turn gradually shapes individuals' cognitive structure. In other words this process shapes agents' conceptual repertoire, i.e., agents' frames (Scazzieri, 2008, 2011). These "naturally connected features" are specific to each individual with respect to her experience and her corresponding cognitive frame. Besides, individuals are not "conceptually omniscient" (Bacharach, 1991, p. 29). Framing is a matter of attention. Because of their non-omniscience, depending on each individual, certain concepts (or families of concepts) are more salient than others in specific contexts (Bacharach, 2001, p. 5). In this perspective, congruence and similarity are two mechanisms focusing attention (Scazzieri, 2008, 2011). They may drive the attention on specific features of the world (ibid, p. 195). This mechanism impacts market exchange. Information is provided in markets by the existence of an offer of a commodity that is labeled a "type F" commodity. The description (by words) of this type conveys information like the context surrounding the offer – e.g. "the name of the supplier, the price, the glossiness of the sales blurb, the ideological position of the newspaper in which the offer appears." (Bacharach, 1990, p. 354) However, a part only of this available information drives the potential consumers'

50 Bacharach's reference to language and to the term "conceptual" opens the door to linguistic and philosophy of logic. - We will see below that Bacharach later resorts to semantic and syntactic logics to question the validity of game theoretic models and in particular the way players' knowledge is standardly defined.

attention. Again this depends on their everyday experience of market exchange and consumption, and accordingly of their cognitive structure. This relies on their "recognitional capacity" (ibid, p. 367). All the information provided by this description will not be considered as relevant. For instance, "the glossiness of the sales blurb" may be relevant for economists, marketers or distributors but not for other consumers.

Since framing is prior to reasoning and decision-making (Bacharach, 2001, p. 5), introducing the phenomenon in economic analysis allows circumscribing the rationale of individuals' decisions (regarding their frames), and in particular the "possibility space" of individuals' beliefs and expectations (e.g. of economic variables, of others' perceptions and eventual responses in case of interactions). Individuals' beliefs are defined within their frames and may be expressed through propositions that are included within those frames (Bacharach, 1986, p. 182). The incompleteness of frames therefore necessarily induces "truncated beliefs" (Bacharach, 2001, p. 9): "the space of propositions or event on which an agent's subjective probabilities are defined is always incomplete" (Bacharach, 2001, p. 5). As explained in the quotation below, this statement has important consequences for Bacharach's account of market exchange. This approach is completely different than the beliefs as standardly defined in epistemic game theory and in Bayesian decision theory. In the latter, the beliefs that the players hold are supposed to be rational (for more details on the way players' beliefs are defined in epistemic game theory see the chapter 1). The way he appraises individual's beliefs is very different form this view. In Bacharach's contribution, players' beliefs are indeed real mental states contrary to Bayesian game theory and not mere mathematical artefacts to justify the existence of the equilibrium in incomplete information games. He however ascribes to Bayes rationality in the sense that individuals and players will act rationally according to their beliefs, i.e. they will intend to maximize their expected payoff according to their beliefs. Here agents' subjective probabilities means that these probabilities are defined or determined according or by their perceptions, i.e. by their frames and more specifically by the structure of these frames. Thus, he does not follow what would be identified as "bayesianism" (Binmore, 1993, 2009): the view that players' beliefs are rational supposedly as it is imposed by Bayes rationality. To sum up in Bacharach's contribution players' beliefs are not mere representation of their choice under uncertainty, as defined by rationality but mental states which translate the way players subjectively appraise their decision problem and evaluate the possible consequence of their choice. Players' beliefs are based on subjective probabilities and not objectivized probabilities like in epistemic game theory (again see chapter 1 for more details).

> "Let [O] be a predicate [i.e. a concept] of the language of the economy E which denotes the type F of [commodity]. When a member of E believes of a thing that it is 'a O', she not only believes that it is an F but also, crucially, she believes (perhaps only implicitly) to apply to it numerous generalizations describing tendencies of Fs… If she does not acquire a sufficiency of commonplace, tendential beliefs of this sort, she does not understand the sentence 'it's an [O]' and she cannot be correctly ascribed the belief that the thing is an [O]. I shall call the network of propositions that embed the concept F and inform in this way competent speakers' beliefs about what they believe to be Fs, the everyday theory of Fs." (Bacharach, 1990, p. 357)

The above quoted generalizations are specific to each individual, in a given time and a given place: they depend both on the individual's personal experience, and history and on the situation in which she is embedded, i.e. on her perception of the situation. We will see later in this chapter and in the chapter 5 how acknowledging such role of personal history and experience in game theory, and therefore of this dimension of subjectivity in decision making is compatible with the necessary intersubjective dimension that is required in games to anticipate the others' choice. This paves the way for the introduction of the theory of mind which will be exposed in the next chapter. And since every individual's frame is incomplete, every individual's "[e]veryday F-theory is partial" (ibid, p. 358). Individuals' and eventually groups' conceptual repertoire(s) do not contain all the potential "properties" of the F commodity. However, individuals can learn new concepts, bringing about new families, new associations of concepts, etc. and this process modifies existing frames, or eventually shapes new frames (Bacharach, 1990, p. 367). Yet, in order to be memorized and enrich individuals' frames, the eventually added predicates – or concepts – must be activated by new experiences and then enter individuals' "everyday theory" (see also Scazzieri, 2008).

> "If a concept F is to enter the repertoire, people must learn the addendum to everyday theory which treats of Fs – they must "do" a new topic; they must often acquire new recognitional capacities; they must enlarge their vocabulary. All these impose cognitive strain. Once acquired, new concepts must be maintained, expensively, in working order. The larger the existing stock, the greater may be the marginal costs of acquiring and maintaining a new concept." (Bacharach, 1990 p. 367)

Such introduction of framing in the realm of economic theory has important methodological consequences. It paves the way to the consideration of agent's subjectivity and heterogeneity in economics. It is important to stress that Bacharach's account of subjectivity is merely explained by the introduction of player's perceptions, i.e. of player's frames, which are personal, and explained by their personal experience. But in no way it implies that any form of intersubjectivity is impossible. Individuals are determined by their experience of social and strategic interactions, by their belonging to multiple communities, so that individuals can in some circumstances understand each other. They can have some partial knowledge of the others, their perceptions, intentions, beliefs and then behavior. This will be explained in more details in the next section of this chapter. Framing also paves the way to the introduction of a psychological appraisal of individual decision-making and in particular of game theory, and to a new treatment of individuals' beliefs and rationality in strategic and social interactions, like in markets exchanges for instance. Indeed, Bacharach declares "The usual form of explanation of an individual action in everyday discourse is psychological; it runs in terms of the desires and beliefs of the agent." (Bacharach, 1986, p. 177)

Such statement is a very strong claim and resort to a deep interconnection between economics and psychology that is very rare for his time. The problem for him is that in economic theory "we speak of the preferences of the agent, rather than his desires, and of the probabilities he attaches to the sates of the world rather than of his beliefs" (Bacharach, 1986, p. 177). He however emphasizes "in all cases there is, at least implicitly, an interaction of what the agent is assumed to desire with what he is assumed to believe, to yield the explanandum or the predictand, the thing the agent does." (ibidem)

Understanding the link between perceptions and beliefs is therefore of the most importance for economics, especially considering that the "theory of beliefs" is at best underdeveloped at worst ill specified for Bacharach (1986). He indeed claims that the economic theory of beliefs is "seriously deficient" (Bacharach, 1986, p. 189). Since beliefs are explained by individual perceptions, and are therefore essentially a psychological matter, the influence of logical positivism in economics leads for him to the harmful methodological trend consisting in getting rid off the individual subjective beliefs, although they "are critical variable in determining economic outcomes" (Bacharach, 1986, p. 178).

When the economic agents' beliefs are however considered, this is done in a very specific way which does not fit with the psychological side of the believing phenomenon and its link with the personal assessment of the – economic – world (Bacharach, 1986, p. 182). This is rather clear in the following quotation in which Bacharach criticizes the rational expectation hypothesis.

> '"The rational-expectations hypothesis does, indeed, attribute to the agent expectations that are in accordance with a correct theory, but it does this in quite special senses of both 'theory' and 'correct'. A 'theory' turns out to be a set of (usually stochastic) numerically specified simultaneous equations in economic variables, of the kind estimated in econometrics. To be 'correct' is to be correct relative to the set of available statistical data on these variables, and by the standards of certain statistical norms for the acceptance and rejection of forecasting hypotheses. In this way, the rational-expectations hypothesis seems to liken the ordinary economic agent to the econometrician – his representations of experience are numerical, his cognitions are cool, his tools mathematical – statistical. In the measure that it does so portray him, this special theory is typical of a general tendency in the standard theories to liken the ordinary agent to the professional scientist. This is an important theme in standard theory" (Bacharach, 1986, p. 187)[51]

This quotation exhibits the limits that Bacharach sees in a purely mathematical theory. Beliefs are generally treated as exogenous determinants. The theorists endow the economic agents with a set of correct and rational beliefs, as if they were theorists, without explaining the ordinary men form such beliefs (Bacharach, 1986, pp. 186-187). Beliefs are also endogenous variable to economic models or economic reasoning (Bacharach, 1986, p. 178). This is even truer when the economic models concern strategic interactions: "among those that are critical and endogenous … are the beliefs about each other's strategy choices held by the players in economic games" (Bacharach, 1986, p. 179). Game theory is about the kind of situations in which a player's reasoning is about "the mental processes of another person" who is, herself, engaged in the same reasoning (Bacharach, 1986, p. 181). This leads to the cognitive mechanism of attributing to this other, some desires, intentions and beliefs (ibid, p. 180). In other words, this kind of situations gives rise to players' beliefs about others' mental states such as their intentions desires and beliefs. Again this means that players' beliefs are real mental states and not notational devices to describe a

51 Such critic echoes Schumpeter (1942)'s one who argued that an important problem of social sciences is to distinguish the rationality of the observer from the rationality of the observed. It also implies that Bacharach is cautious with mathematization in economics. He does not accept a purely mathematized conception of economics. We will see such concern later in game theory in the section 3.

rational choice made at the equilibrium. This again calls for the introduction of psychology in the realm of economics and imply seeing economics as an open-system. Indeed, for him,

> "How such beliefs are formed, and which ones are the subject of the theory of 'attribution' in social psychology (see for example Hewstone, 1983). It will be noticed that attribution theory is about second order beliefs, beliefs about beliefs (and other second-order 'propositional attitudes', such as beliefs about desires)." (Bacharach, 1986, p. 180)

Again, for Bacharach (1986, p. 181), this psychological mechanism of attribution is does not exist in economic theory. It is especially the case in game theory because when the game starts players are already endowed with beliefs concerning the other players' preferences or beliefs. As emphasized in the quotation such conception means integrating social psychology in game theory. Asserting the importance of attribution theory sets the entrance of social psychology in Bacharach's thinking. We will also see in the chapter 5 how to offer an alternative explanation to such attribution mechanism to explain players' second order beliefs.

> "There is no theory of attribution in standard economic theory. That is, there is no non-trivial theory, no epistemic behavior relation or set of such relations which determines the values of the agent's attributional variables (his beliefs about other people's desires and beliefs) as a function of other variables, paying due attention to the distinctive features of attributional beliefs. When I speak of the 'distinctive features', I have in mind such things as these: the way in which people understand and predict each other by consulting a socially inculcated catalogues of 'roles' (businessman, Northerner, etc.) or the fact that the beliefs I attribute to you must depend on the degree of rationality that I attribute to you – for example, on whether I believe you capable of revising your beliefs in the wake of new evidence in accordance with Bayes' theorem."(Bacharach, 1986, p. 186)

Bacharach is here criticizing the reluctance of economists and game theorists to progress in their understanding of an "empirical theory of belief", and because of their unwillingness to "incorporate the results of empirical investigations in psychology and elsewhere" (Bacharach, 1986, p. 190). Thus Bacharach recognizes the importance of many determinants in attribution, among which supposing that the other is Bayes rational is only one possibility. Resorting to the other's profession, social category, i.e., to echo Schelling's work, to their 'social roles', offers many other possibilities. Recall, as explained in the previous chapter, that social roles, like social categories and professions involve patterns of behavior, i.e. rules of behavior that orient players' beliefs and therefore allow the attribution of beliefs to the other (by grounding on a community-based reasoning entailing symmetric reasoning and therefore, ultimately, common knowledge of each other's perceptions, beliefs, reasoning, etc.). In such perspective Bacharach's conception of strategic interactions, like Schelling's, stands for open-systems that require an interdisciplinary analysis. One problem for the enhancement of a theory of belief in economic theory is, for Bacharach, its methodological *a priorism* and methodological rationalism.[52] As explained in the chapter 1 beliefs are, in decision theory under uncertainty and in game theory, the representation

52 Bacharach (1986, p. 189) claims that the standard theory of belief in economics "are all deductive theories; they are all aprioristic; they are all rationalistic".

of the choice of a rational agent at the equilibrium. Beliefs are supposed to be rational beliefs and not the subjective apprehension of the agent of the consequence of her choice. There is therefore no account of the problem of the elicitation of the agent's beliefs. There is no explanation of how these agents, who initially form free beliefs, come to revise those beliefs to ultimately hold rational beliefs. Such methodological constraints in the standard theory of belief are incompatible with empirical evidence (ibid, p. 189), and none of the standard accounts of beliefs are rational for Bacharach (1986, p. 193).[53] However,

> "By drawing on stylized empirical facts about the cognitive capacities of economic agents, it may prove possible to build a deductive Theory of Beliefs for economics, with all the fruitfulness of results and all the explanatory power of such theories; a theory which is, moreover, a true (enough) description of the epistemic behavior of agents who are rational (but only as rational as their nature allow). There is nothing impossible about an a posteriori, deductive, rationalistic Theory of Beliefs in economics." (Bacharach, 1986, p. 201)

All of these claims plead for a "radical reconstruction of the whole choice theory" (Bacharach, 1986, p. 183). Such reconstruction would necessitate the integration of results and theories in cognitive sciences, in philosophy, in psychology and social psychology. For Bacharach, in addition to its lack of realism and under-specification of the empirical foundation of individuals' beliefs, game theory faces severe methodological issues regarding its foundations. This is discussed in the next subsection.

2.2. A critical assessment of standard game theory

Bacharach often points out that the lack of interests for foundational issues is recurrent throughout the history of game theory.

> "Not much of the history of game theory is the history of concern over fundamental questions about games. The theory of games has been developed for the most part as a mathematical subject. The study of the existence of mathematical objects meeting certain conditions - for example, of the existence of Nash equilibria and related sorts of equilibrium points - and of methods for computing the values of suchlike mathematical objects, have predominated over the study of the decision-theoretic significance of the conditions themselves. One of the neglected questions at the foundation of game theory is the question of whether different classes of game have valid solution concepts - whether there are recipes which really do succeed in identifying the things that rational players would do in them. There has, it is true, been debate about this. But it has been

53 This is true for beliefs as formalized by subjective Bayesian probabilities for Bayesian revision, for the beliefs supporting Nash equilibrium. For him, none of these accounts of individual beliefs is rational or supported by the process of a rational reasoning (Bacharach, 186, p. 193).

> peripheral; and on the whole it has consisted in trials of plausibility among axioms and has failed to penetrate to deep enough levels of analysis." (Bacharach, 1987, pp. 38-39)[54]

This quotation, mentioning a critic against a (too) mathematical game theory only interested in the existence of a solution, and revealing an under-determined ontological thinking with respect to "how and why" an equilibrium occurs, seems quite similar to Schelling's claims. However, Bacharach also criticizes standard game theory from a formal point of view, especially in its treatment of the solution concept and the preeminence of Nash equilibrium as *the* game theoretic solution concept. He shows that common knowledge of the structure of the game and common instrumental rationality or Bayes rationality are generally insufficient to guarantee the existence of a Nash equilibrium. Surprisingly for an economist trained in mathematics and who has always developed formal models, he is very critical toward a purely mathematical account and a closed-system perspective on decision theory and game theory. For Bacharach, there are serious questions that remain about the solution concept in game theory and the way rational players are led to choose the equilibrium profile of strategy. He then uses epistemic logic – while this approach was quite unusual at that time within game theory – to investigate the standard requirements for the existence of a rational play in games, the existence of a determinate solution and their foundations.

> "Game theory is full of deep puzzles, and there is often disagreement about proposed solutions to them. The puzzlement and disagreement are neither empirical nor mathematical but, rather, concern the meanings of fundamental concepts ('solution', 'rational', 'complete information') and the soundness of certain arguments (that solutions must be Nash equilibria, that rational players defect in Prisoner's Dilemmas, that players should consider what would happen in eventualities which they regard as impossible). Logic appears to be an appropriate tool for game theory both because these conceptual obscurities involve notions such as reasoning, knowledge and counter-factuality which are part of the stock-in-trade of logic, and because it is a prime function of logic to establish the validity or invalidity of disputed arguments." (Bacharach, 1994, p. 17)

In this perspective, he opens the door to the philosophy of logic in game theory, which he uses to question the epistemic foundations of game theory. More precisely, he uses epistemic logic "to broach [an] ambitious task, the task of developing a theory of the "solution" of games" because "having at one's disposal a theory of games does not make it a trivial matter to say whether this or that type of game has a solution; indeed, it is not a trivial matter to say what it is to be a solution of a game." (Bacharach, 1987, p. 17) He sets his theory to describe games without the purpose to solve them contrary to the standard mathematical account. His goal was rather to investigate if the standard requirements imposed in game theory to describe a game are sufficient to provide solutions for any type of games (ibidem). This investigation leads him to conclude that in many cases this account is insufficient: stronger requirements and additional axioms, e.g. with respect to the definition of a rational play, are needed (ibid, pp. 30-31). Although as declared in the quotation at the beginning of the paragraph Bacharach ambitioned to build a "theory of

54 The debates that were peripheral, that Bacharach mentions concern the Refinement program, which thus reveals disappointing on the questioning on strategic rationality, i.e. on why players should play such or such equilibrium.

"the" solution of games" he did not develop it. He however offers a damaging internal critic of standard game theory which fails as a normative theory of strategic interaction because its treatment of solution concept is ill-founded. And such critic is made by using the same axiomatic method as standard game theory to show what would be required for any class of games to guarantee the existence of a solution. However, as we will see in the remainder of this chapter he then assumes a more ambitious task with the VFT and TR theories: not only changing of solution concept but changing the way players are portrayed in games and the way rationality is defined which ultimately challenge game theory more deeply than focusing on such methodological stalemate.

From that prospect Bacharach deconstructs the standard assumptions made supporting the existence of a solution among which are the best reply reasoning and the dominance principle, which together entail, as is commonly supposed, that there exists a unique solution of the games. Those standard assumptions concern the Nash equilibrium, the rule of epistemization and the transparency of reasons. Bacharach states these standard assumptions as follow.

In the standard account, an action is rational or say "satisfactory" if such action is a best reply to the other's action. In this way the players achieve the most-preferred outcome among those that are attainable (ibid, p. 24). The principle of best reply is comparable and generally conflated to the dominance principle. They are "the primitive assumptions about rational decision-making in games" (ibid, p. 27). However, Bacharach (1987, p. 28) considers that "[t]he Best Reply Principle is a necessary condition for an action to be known satisfactory. The "dominance principle" is a sufficient condition." (ibid, p. 29) In order to be a best reply an action must be dominant "in a strong sense", i.e., it must be the unique action whichever the action of the others (ibidem). Players must know this fact and this property of dominance must be "detectable" (ibid, p. 29). Bacharach shows that to be a solution concept dominance must respect these three conditions. (Bacharach, 1987, pp. 28-29) He however also shows that these three conditions are far from being met in many games.

The rule of epistemization (RE) generates "iterated interpersonal knowledge" (ibid, p. 26). It states that everybody knows that everybody knows the characteristics of the games (i.e. the set of available action and preferences), the theorems or axioms of the theory and their logical consequences. If there is in such theory an axiom defining a rational play by implication, everybody is rational, known to be rational, know that she is known to be rational, etc. As a consequence everybody know that the action taken by a player is a best reply to the best reply of everybody else (Bacharach, 1987, p. 26). We understand that such RE in Bacharach's thinking cannot prevail for any kind of players. As discussed above mathematicians can manipulate concepts and frame games in specific ways that are not necessarily accessible to other non-mathematician people. Arguing that non-mathematician players possess the knowledge required by the RE is therefore far-fetched. Besides the strength of knowledge that is required by the RE is also unrealistic when remembering that people are not conceptually omniscient and have limited cognitive abilities.

The transparency of reasons is for Bacharach "the claim that if reason suffices for a player who has certain data to come to a given conclusion - say, as to what he should do - then, if a second player believes the first to have these data and to be rational, reason suffices for him to come to

the conclusion that the first will come to his." He adds "[i]n games, players are aware of each other's data and each other's reason, and these data and this reason are their sole means for coming to conclusions. And so it follows from the Transparency of Reason that, in games, if a player knows that he should do something then the other knows that he knows he should do it" (Bacharach, 1987, pp. 36-37). Such principle is considered in standard game theory as "an essential element in an important argument - influential but problematical - which purports to establish the proposition that "only Nash equilibria can be solutions of games"."(Bacharach, 1987, p. 37) Bacharach finds this statement problematic. It makes players symmetrical, having identical reason and knowledge, knowing for sure what everybody will decide i.e. correctly inferring everybody else's action. A game is thus trivialized in this condition.

With all of these analytical underpinnings, and even with the huge strength of knowledge that the players are supposed to possess, there are however games in which solution-concepts like best reply reasoning and the principle of dominance are ruled out and there are games with a unique Nash equilibrium that Bacharach proves to lack a solution with the standard axiomatic approach (Bacharach, pp. 48-49).

In these cases insuring the existence of a solution would require resorting to an "Existence Postulate" (EP), i.e. the assumption that this type of games effectively possess a solution (ibid, p. 18).[55] The existence postulate can imply the existence of a unique solution, however, again, "there are normal-form types of games for which an Existence Postulate analogous to (EP) is false. This is so for any type which includes normal forms which have no Nash equilibria. Hence the generalization of (EP) to all normal-form types [of games] is certainly false. But if (EP) is false … in general, there seems no particular reason why it should be true in the case of [particular type of] normal-form [games]." (Bacharach, 1987, p. 44) Therefore, what Bacharach's analysis reveals is that the standard game theoretic account in terms of the axiomatic method applied to choices "by itself is likely to prove an inadequate tool with which to tackle the solution question." (Bacharach, 1987, p. 30)

Again we see that what preoccupies Bacharach is to explain how and why a specific solution for a set of games occurs. Imposing the existence of a solution through the EP would not be satisfactory for him as it would neither explain how the players could reach such solution nor why this solution is a rational way to play. He indeed declares "if we can accept the broad Existence Postulate we can exactly specify the actions of rational players … however, that all would not be well, for though we would be able to predict the actions of such players, we would be in the odd position if being unable to explain them!" (Bacharach, 1987, p. 18) Providing a theory of games tied up by an axiom stating the existence of a unique solution as a requirement for rational play and again supposing that rationality is explained by the knowledge of such solution "may not provide an explanation of the actions of the players identified by the solution

55 More specifically, the EP would require to impose the following statements "saying that a particular game "has a solution" implies that (i) for each player of the game there is an action α such that it is in some sense good for him to do α. Perhaps less obviously, it implies that (ii) the player is in a position to come to know that this is true of α. And it implies thirdly, at least for many writers, that (iii) for each player there is exactly one such α." (Bacharach, 1987, p. 35)

concept … With the aid of (EP) we would have learned that in any game with a unique Nash equilibrium, rational players play their parts in this equilibrium. But, as I have shown, we would not know why they do so." (Bacharach, 1987, p. 47)

Considering that "[t]he task of normative game theory is to give and defend a specification of what rational players will do in the games they play." (Bacharach, 1994, p. 7) and considering that Bacharach proved in the 1987 paper (with an axiomatic of choice derived from epistemic logic) that game theory failed in this task, his conclusion is a claim for new principles of "practical reasoning" in games. He argues for yet undiscovered principles (Bacharach, 1987, pp. 48-49): "the current conception of a game may be capable of improvement; fruitful and illuminating theory may well require, for instance, that the notion of rationality in game situations be cultivated in some way that goes beyond what game theorists nowadays mean by "games" or "rationality"."(Bacharach, 1987, p. 32)

One of such investigations would in particular involve players' knowledge and the link between knowledge and practical rationality. Appraising such knowledge means for Bacharach to identify different forms of knowledge and the role they play in individual decision-making (both in decision theory and game theory). He indeed distinguishes occurent from non-occurent forms of knowledge.[56] This opens the door for the introduction of framing in decision-making. Another would be to identify how players resort to social knowledge, to institutional devices bringing rules of behavior, and common reasoning, and how they affect the outcome of the games.

Bacharach claims that "[n]either the exact relation of common knowledge to simple knowledge, nor the possibility for real human agents of achieving it in practice, is yet agreed." (Bacharach, 1997, p. 1) One position often assumed is to consider common knowledge as a purely notational device, a mathematical artifact, which ultimately rules out any eventual questioning on such epistemological matter. We will see in the chapter 5 how to propose a solution concept in games that do not require imposing such common knowledge assumption. Accordingly, instead of strengthening the rationality or the knowledge conditions by which players are exogenously endowed, which as shown its limits and which is grounded on the Nash principle but which is not, as demonstrated by Bacharach, a convincing principle for rational choice, he militates for enlarging players' epistemic condition to empirical, i.e. social, cultural and historical information. "The logical models of knowledge do not take for their subject-matter all aspects of human knowledge. In particular, they are not concerned with its empirical basis, nor with inductive reasoning. (Bacharach, 1994, p. 11)

The condition imposed by the rule of epistemization is too strong for the players (Bacharach, 1987, p. 25). Considering other sources of knowledge and information would be of great value for game theory according to him. Standardly, "[t]he emphasis is on the persons' inferred knowledge – hardly anything is said about how the persons come to possess the non-inferred knowledge they do." (Bacharach, 1985, p. 168) This non-inferred knowledge is the social, cultural and historical knowledge: the knowledge of which rule of behavior, way of interacting, etc. that is

56 See Arena, Larrouy (2016) for more details on those two forms of knowledge and the role they play in individual decision-making.

stored in individuals' background (see Searle 1994, 2005, 2010; Hédoin, 2014, 2015, 2016), so that these mode of interaction become unconscious (see Arena, Larrouy, 2016). As emphasized in the next chapter, the social world can have a strong influence in players' inductive reasoning (see Hédoin, 2014, 2015, 2016).

These considerations explain why Bacharach was finally more inclined to integrate empirical determinants in individual decision-making in strategic contexts, to distinguish players' real knowledge from that of the theorists and with which the theorists endow them. Framing is a methodological ground for that purpose. This is particularly unequivocal in the following quotation

> "In games as traditionally specified the attitude data are limited to preferences over the game's outcomes, plus knowledge of the rules of the game and of the other players' rationality, preferences, and similar knowledge … So we may not use the fact that every real player has the general knowledge her culture gives her, such as knowledge of which arrangements are salient or traditional in that culture and so provide coordination. The limitation to attitude data which omit such knowledge dehumanizes the decision-maker in the opposite direction to the traditional idealization of her powers: instead of exaggerating her resources, it understates them" (Bacharach, Hurley, 199, p. 3).

The world as seen by the theorists should be distinguished from the world as seen by the players. This is exactly the same claim as Schelling's. This can explain why players' practical reasoning may be better than is assumed in standard game theory: players are reasoning when they face a decision problem from their position and from their point of view. We found in these words similar claims to Schelling. And in reality they effectively coordinate and cooperate where standard game theory remains silent. However, in that perspective, coordination and cooperation rely on contextual, social, cultural and historical determinants, which challenge and even contradict some pillars of standard individual rationality in terms of subjective expected utility in games. This calls for considering games and game theory as open-systems. This is one the main message that Bacharach later delivers through the VFT and TR theory.

Bacharach considers that standard game theory "is going so far, abstracting too much, [and] confusing the essences of quite disparate things" (Bacharach, 1976, p.1). His ambition is therefore to follow the opposite path, i.e. to avoid resorting to exogenous preferences, to exogenous knowledge, to the standard representative agent, and to the standard hyperrationality assumption. From that prospect he resorts to an interdisciplinary analysis of individual decision making in both decision and game theories.

3. Bacharach's "Variable Frame Theory" and coordination.

We draw in this section on both the published and unpublished papers of Bacharach on the VFT. They are both important to exhibit Bacharach's main concerns in this theory and the epistemological and methodological concessions he had to make in order to integrate all of the

determinants he considers to be primordial in players' decision making within the formalism of game theory. Indeed, both methodological and conceptual aspects of Bacharach's work on "Variable Frame Theory" (VFT) evolve.

The purpose of this theory is to justify the existence of a determinate solution in coordination games, i.e. to *rationalize* coordination and for that, Bacharach integrates Schelling's focal points in a game theoretical framework (Bacharach, 1991, p. 5; Bacharach, 1993, p. 270; Bacharach and Bernasconi, 1997, p. 2). This implies for Bacharach, that the way players' coordinate on focal points has to be drawn from *rigorous theoretical principles*, i.e, from a *valid theory of games*. In this perspective, Bacharach purports to build a "theory of focal points" (1993, 1997). It is now widely acknowledged that focal points are of a particular interest for standard game theory. They challenge the problem of equilibrium selection within games. Focal points and salience are considered to be one path within the Refinement Program to escape from the problem of indeterminacy (Sugden, 2001, p. 116). Accordingly, game theorists showed a major interest in this concept (Metha, Starmer and Sugden, 1994a,b; Sugden, 1995; Casajus, 1998; Colman, 1997; Janssen, 2001, 2006; Sugden and Zamarrón, 2006). Yet, focal points and salience never became accepted game theoretical principles (Sugden, 2001: 116, Innocenti, 2007; Bacharach, 1991, 1993; Bacharach and Bernasconi, 1997).

Bacharach attempted to fill this gap. For that purpose, he integrated the players' frames within games. And within games, Bacharach defines frames as follows: "a players' frame is, most simply, the set of variables she uses to conceptualize the game" (Bacharach, 1997p. 4). He uses framing in a very specific way, which is quite unusual within the literature on "framing effects". [57] Bacharach mainly attempted to understand the 'natural' process of framing in order to appreciate the influence of frames on individuals' decision-making and 'practical' rationality in everyday life. Framing puts an emphasis on the fact that 'the act' of choosing requires two steps: (i) the conceptualization of the decision problem, and (ii) the evaluation of the different options the agent faces with respect to the way she framed her decision problem (Klaes, 2008, p. 215).

3.1. Framing and gaming.

In order for the VFT to be a valid theory (with respect to Bacharach's definition of a valid theory)[58], its formalization requires three main premises:

---

57 Bacharach's conception of framing is very different from the one of Kahneman and Tversky (1979, 1986) that is commonly used in economics and the 'framing effect literature'. Bacharach retains from Kahneman and Tversky's "prospect theory" (1979) the idea that individuals' beliefs and preferences depend on their subjective descriptions (representations) of the world (Bacharach, 1986: 183). However, Bacharach distances himself from Kahneman and Tversky and the framing effect literature by being interested solely in natural framing, i.e. in the absence of 'manipulations' designed by the theorists to affect individuals' decision making (2001, p. 4).

58 A *valid* Theory of Games must be grounded on a *meta-theory of games* for Bacharach (1987, 1994). More specifically, a *meta-theoretic game theory* "does not regard the game directly, but regard a feature of the game theorists' theory of it, namely the 'logical closure of the information attributed to the players in this theory's

i. The model must specify players' frames, i.e. players' representations of the decision problem they face and "what determines players' frames" (Bacharach and Bernasconi, 1997, p. 5). This requires a description of the process by which frames come to players' minds and an assertion on the structure of players' frames.
ii. According to their frames, players must have a set of alternatives, i.e. subjective strategies – among which they have to choose. Distinguishing subjective strategies and objective strategies means that it is the players *themselves* who establish their own set of strategies according to their appraisal of their decision problem (and which may differ from the strategies considered by the game theorist in the games matrix). In other words, these strategies are subjective in the sense that they are determined by the players' perceptions of the game. For instance, it can imply that players relabel the strategies they face if the way such strategies as labeled by the theorist do not make sense for them (this echoes Schelling's claim).
iii. Some rational principles of equilibrium selection and of the properties of games' equilibrium must be involved (ibidem.)

Understanding how frames influence individuals' decision-making requires for Bacharach to identify: (i) what are the determinants of individuals' frames, (ii) the process by which frames come to individuals' mind, and (iii) and the internal structure of frames (Bacharach and Bernasconi, 1997, p. 5). Bacharach's purpose is to draw a model of individual decision-making permitting a full account of this mechanism.

These premises allow Bacharach to dwell into a game theoretical framework. Nonetheless, the first two premises violate some implicit rules of standard game theory. First, in standard theory of games, the players' conception of their decision problem does not matter and cannot vary. What matters for solving a game is only contained in the payoff matrix – which specifies the combination of the available actions for each player and their 'desirability'. To the contrary, in the VFT, games are no longer self-contained worlds. They are open-systems. It is impossible to build a game without considering its context. For more details on the methodological consequences for game theory to see games as open-systems, or as large worlds, see the chapter 5. The chapter 5 will propose a theory of games in which games are conceived as large worlds, i.e. as open-systems. Second, if players make their own strategies they may no longer act according to the standard rules of the game – those that are defined by the theorist.[59] In VFT players have to build their own rules: "players must derive "rules" – formulate their problem to themselves as a certain game – from some initial apprehension of their situation" (Bacharach, 1993, p. 258).

---

assumptions" (Bacharach, 1994, p. 8). For instance, "it allows one to focus attention on the logical relations between what players know just before they choose and the 'normal form information' with which they are traditionally credited; more specifically on how they know just before they choose depends on what can and cannot be proved from their normal form information" (ibidem). Besides axioms must specify the structure of games, they enclose: "the "set-up" (the available actions and preferences); the players' knowledge about this and about each other; and what it is that makes actions "satisfactory" and how "satisfactoriness" relates to choice" (Bacharach, 1987, p. 17). In addition, such theory must specify what is rational for a player to do, given his information.

59 To quote Hargreaves Heap and Varoufakis (2004 [1995], p. 31), in standard game theory, "individuals know the rules of the game: that is, they know all the possible actions and how the actions combine to yield particular pay-offs for each player".

This statement entails that: (i) who are the players matter and (ii) the way they label their strategy matter. Moreover it can imply a dissymmetry between players which go beyond, and is even more disturbing than, a simple informational dissymmetry – solvable by Bayesian updating. For instance, players can have different strategic spaces. At this point we already understand that Bacharach has to deal with a difficulty. As soon as each player builds herself her own game, Bacharach must define upon what *common basis* players will interact.

Players define the game they are going to play through a two steps process:

i. A first *unconscious phase* which is the 'framing phase': "[t]he way people conceive their options is typically beyond conscious control" (Bacharach, 1991, p. 1)
ii. A second *conscious phase*: the "reasoning phase" (ibid, p. 3), in which principles of rational choices are operative.

Roughly speaking the object of the first stage of the VFT is to "offer a rudimentary descriptive model of the process that brings questions to the minds of the players and the consequences of this process" (Bacharach, 1993, p. 258), whereas the second stage is the game *per se*.

When "normal" agents frame a decision problem, they have families of concepts, which come to mind.[60] A family can be understood as "classifying things along a single dimension" (Bacharach, 1991, p. 3). Concepts are basically characteristics or 'properties' of the objects; for example being green, red, round, etc. In these cases the families to which these properties belong are 'color', 'shape', etc. As Bacharach builds his VFT with matching games, i.e. games in which players have to choose objects and coordinate by choosing the same object, concepts are assimilated to characteristics, i.e. to properties of the objects.[61] There is no semantic in the characteristic associated with the objects. For instance, players could possibly interpret or give different semantics to the colors they perceive. As a consequence, their representations would differ and accordingly, they may not think and deliberate from the same premises.

Here are graphical examples of matching games. These matching games have been used in Bacharach and Stahl's experimental settings in 2000.

60 Bacharach gives variably the labels 'concept', 'attribute' or 'predicate', nonetheless the same semantic remains. I must point out that the label 'attribute' is the usual term to which psychologists refer within the framing literature, whereas the others are specific to Bacharach.

61 The precise definition of a Matching Game is given as follow: it is "a pure coordination game in which there are two players with the same act-set; both get a prize if and only if both choose the same act; and the prize is the same whatever this act may be" (Bacharach and Bernasconi, 1997, p. 2)

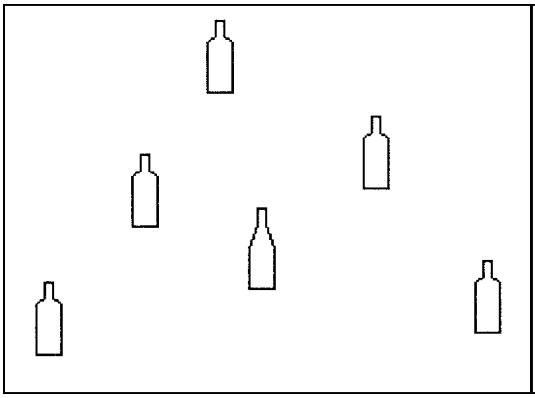


Figure 1

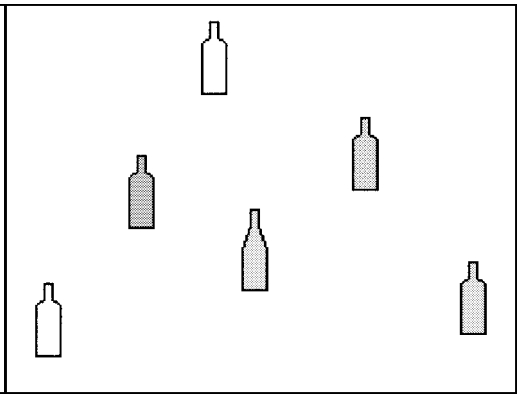


Figure 2

In each case players have to coordinate by choosing the same bottle, if this is the case they win 1€ if not they earn nothing. In the case represented by the figure 1, the frame structure is only composed by the shape family; while in the case represented the figure 2, the frame structure is composed by two families the shape-family and the color-family. The payoff matrices generated by such strategic decision problem will be explained in the next subsection.

In Bacharach's account, the structure of frames respect the following main properties:

i. A family is supposed to be separating, i.e. two families must be distinct. If, for example, *f* and *f'* are two families of concepts (*f* being the color family, an*d f'* being the shape family), *b*, a given concept, cannot belong to both *f* and *f'* (ibid, p. 10).
ii. Every concept belongs to a family (and only one family) (ibidem).
iii. A "generic" family comes to players' mind with a probability of one. It entails that each situation can be described. It is basically equivalent to say that every thing can be described by the concept 'object'. It allows players to at least choose something. As a consequence, under any circumstances, decisions problems are always describable, even if the way by which decision problems are describable always varies (Sugden and Gold, 2006, p. 15)
iv. The set of families of concept, called a "repertoire" (1993, 1997, 2006), represents a "partition" of all the concepts that describe the situation.
v. If one concept of a family comes to players' mind the other concepts belonging to the same family do to (ibidem). That is why Bacharach (1991, 1993, 1997) claims that concepts come to mind into "clusters". He gives the following example: "if I notice that the blocks are square, I am likely to notice that others are triangular" (Bacharach, 1993, p. 259).

Such properties of framing are specific to the case under study, i.e. to matching games. They are indeed very restrictive and do not seem to be operative for many situations of strategic interactions. We will see the consequences of these properties and their limits in the next subsections (see also Janssen, 2001, pp. 123-124). Accordingly, players' repertoire represents their

frames, i.e. the set of families of concepts they handle. Depending on the context some characteristics of the objects and therefore the associated families "instantiate". The likeliness of a characteristic – and its associate family – to come to players' mind is formalized by a probability called the "availability" of the family (Bacharach, 1991, p. 17). This notion of availability is a kind of "propensity" to perceive something and the fact is that, for Bacharach, this propensity can be "interpreted as that of a normal person in the circumstances" (ibidem). Framing is a matter of salience and noticeability: "whether a normal person thinks of a predicate [i.e. a characteristic] in a situation depends on how conspicuous or noticeable is the feature of the situation which the predicate expresses" (ibid, p. 16). Accordingly, "highly conspicuous features are more likely to come to mind" (ibidem). Nonetheless, availability may not only be a matter of context and environment, but may depend on players: "frames may vary … across players" (Bacharach, 1997, p. 4); and more precisely on players' culture: "[the availability function] is given by the culture to which the players belong" (Bacharach, 1993, p. 267).

Then, players' set of action are defined from a set of options; "an option of a player is a feasible action for her described as she herself describes it. It follows that a player's options must be descriptions of possible actions which only use attributes [i.e. predicates] in her frame" (Bacharach and Bernasconi, 1997, p. 6). Again we see Bacharach's particularity. In standard game theory there is no need to label a strategy in such a way, since the 'desirability' of each outcome is sufficient for players to make their decisions. More specifically, Bacharach and Bernasconi assert that "a player has one option for each attribute belonging to a family" (ibid, p. 7). From each available concept, the description of an act follows (Bacharach, 1991, p. 16). In a matching game, this conceptualization allows only two types of actions: (i) "choose" the unique object which possesses the property *b*, for instance, or (ii) "pick at random" one of the objects when none possesses a unique characteristic.

Players' strategies, i.e. the set of actions that appear in the game matrix, are only subsets of their available options. They depend on the others' sets of options and, in turn, their strategies, defined in the same way, i.e. defined within their representations. Indeed, in each game matrix, as it is framed by the players, only a subset of their available options remains. Players' strategies are defined by taking into account the possible options and thereafter strategies of the others. Therefore players eliminate the options they consider to be irrelevant according to the interaction they are going to have with the others. Bacharach however does not explicitly underline such distinction between the set of options and the ultimate set of strategies.[62] Besides, in his own definition of matching games, he specifies that the set of players' strategies are symmetrical.

An ambiguity therefore appears. Bacharach specifies that at the end of the first unconscious phase, the game is determined: "this phase determines a game in which [the players] have the options they conceive" (1991, p. 2). We can extrapolate that the process which determines the game still belongs to the first unconscious phase. In addition, in Bacharach's VFT, deliberation is about the game itself, i.e. the moment in which players solve the game and rationally choose a

62 However, Bacharach does not have a clear distinction of these two concepts. This first ambiguity raises another one, more noticeable in the 1991 paper. There is no clear account in Bacharach's VFT, on when the first unconscious phase ends.

strategy. Yet, when players think of the options that may eventually be shared, they seem to deliberate. Such ambiguity, which appears in the 1991 paper, tends however to be lessened in subsequent papers (see Bacharach, 1993; 1997). Bacharach asserts in these other papers that players' options are the final strategies. As a consequence, the shortcoming to which we refer become less obvious.

In fact, this problem may be considered along with the fact that "actions need to be individuated by the theorist in some way that is independent of agents' descriptions" (ibid, p. 5); and so it is for players' strategies. We understand that in one way or another, the theorist must 'objectivize' the game or at least some of the fundamental aspects of the game to ensure a symmetrical set of strategies. Yet, this is question begging: how players themselves can come to agree, without communicating and considering that they can have different perceptions, on symmetrical action-spaces?

3.2. The "status" of the game: what is a payoff matrix?

Incorporating the players' subjective representations of their decisional problem raises the question of how they can come to a common game? In order to coordinate players have to agree on the game they are playing. Bacharach must indeed justify in some way or another a common structure within which players will be able to play. As stated by Scazzieri (2008, p. 187),

> "a critical feature of rationality assumptions in economic theory is their association with the way in which reasoning and decisions by heterogeneous and independent individuals (or collective agents) may be made consistent with one another within a congruent structure"

Because players' representation of a common decision problem may be potentially dissimilar, according to their frames, Bacharach needs to guarantee a "congruent structure". To understand how possibly different subjective representations can lead to a matrix shared by every player, we must first understand the status of a game in VFT.

Analytically, the matrix of games matrix in VFT corresponds to the joint probability for players to choose the same object.

I refer to the example of the "Bottles Game" described above (cf. figure 1), to highlight how the strategic decision problem faced by the player leads to a matrix.[63] Recall that the players face different bottles and have to choose one of them. Coordination is successful if they choose the same bottle.

K is the number of bottles, and presently $K = 6$. The shape of the bottles is either 'hock' or 'claret', the frame partition is accordingly threefold: {hock-shaped, claret-shaped, empty-

63 Such example of matrix comes from Bacharach and Stahl (2001, p. 222)

shaped[64]}. Three options are therefore at the disposal of the players: (i) 'choosing the hock bottle' (h), (ii) 'choosing a claret bottle (c), or (iii) 'choosing any bottle' (b) when the shape-frame is empty and neither the hock-shape nor the claret-shape is perceived.[65]

This decision problem with two players induces the following matrix structure:

		Player 1		
		h	c	b
Player 2	h	1	0	1/K
	c	0	1/(K – 1)	1/K
	b	1/K	1/K	1/K

With the 'option-induced payoff function' the matrix then become:

		Player 1		
		h	c	b
Player 2	h	1	0	1/6
	c	0	1/5	1/6
	b	1/6	1/6	1/6

We understand at this point that Bacharach, as the game theorist, assesses the payoff matrix that is built from the "Bottles game" objectively. If, for instance, one of the players under scrutiny does not perceive the difference between the hock and claret bottles, she cannot come herself and independently to this payoff matrix because she would not possess as options (h) and (c) – i.e. 'choose the hock bottle' and 'choose a claret bottle'. Such claim will be understood more precisely with what follow.[66] Bacharach (1993, p. 256) emphasizes that standard non-cooperative game theory "misspecifies the payoffs matrices". This is due to the fact that, for him,

> "[a]ccording to the way that normal agents tend to 'cut-up the world' as they come to terms with a problem, this problem gives rise to payoff-matrices that differ widely in terms of the number of options and the pattern of payoffs: the payoff structure generated by a problem is concept-dependent" (Bacharach, 1991, p. 3).

64 In this case the player does not perceive the shape of the bottle.
65 The empty frame is for Bacharach a subframe of the shape frame.
66 The possible equilibrium according to this payoff matrix will be detailed in the next subsection.

This statement implies that the purpose of the VFT is a respecification of the standard coordination games. This 'respecification', attempts to transform the standard coordination games in which there are multiple equilibria in games in which *only one* Pareto optimal equilibrium remains. In the example described above, one Pareto optimal equilibrium emerges, but this is explained by the specific structure of the frames that the players handle. In other games the problem of multiple equilibria may however remain (see Janssen 2001, pp. 123-24). In Bacharach's view, a payoff matrix is the mathematical translation of the strategic problem that the players face. He intends to begin with real strategic decision problems, as they can be perceived by the players, then attempts to translate such appraisal in game theoretic terms and ultimately applies a solution concept (in order to propose a "theory of games"). In the example above, as one of the profile of strategies (h;h) expresses the fact that players have a higher probability to choose the same bottle, i.e., the hock bottle, it becomes by construction a Pareto optimal equilibrium. But this is not because the profile of strategies (h;h) is Pareto optimal that this will be the solution of the game. Some frames will be more salient than others for these normal players and accordingly, the option associated with the salient frame will be more likely to be chosen by the players. The joint probability of the underlined strategy increases. This is why it becomes a Pareto optimal equilibrium. Nevertheless, as already mentioned, it is the theorist who necessarily accomplishes this process. It may seem in contradiction with the fact that a game is an abstraction process carried out by the players themselves. Besides, it still does not explain how players come to a common respecification or whether they perceive the same joint probability. So how do players recognize the game from the moment that it is no longer an objective game but a framed game in which each player has the options she herself conceives? This difficulty induces several epistemological and methodological changes between the published and unpublished papers. Such difficulty will also be discussed in the chapter 5. We will in particular resort to mindshaping theory to explain how players can independently come to a common apprehension of their decision problem, i.e. to common inductive inferences.

In his first contribution on the VFT in 1991, Bacharach argues that among players, and especially 'normal' players, there is a tendency to perceive the situation in the same way, and to understand this fact. This assertion is explained by the combination of different elements. First, players' frames are structured in the same way, i.e. their frames tend to be constituted by common sets of families. Second, "the propensity [to perceive such or such thing] is interpreted as that of a normal person in the circumstances" (ibid, p. 17). As every player is considered as a normal player, those propensities tend to be similar. This statement implies that each player ultimately tends to possess the same repertoire because each player is normal. Third, knowledge in the VFT is an "occurent knowledge", i.e. it occurs by the observation of the context. This only piece of knowledge that players possess is such occurent knowledge, i.e. the one that comes from the observation of the context of the decision problem. This methodological device allows Bacharach to erase the possibility of different individual epistemic structures, and it explains how players come to commonly know the game matrix.[67]

67 Those postulates echo Bacharach' work, when he investigates the sources of common knowledge in 1997. And for him, what he calls "the shared environment approach" could explain the genesis of common knowledge. He

Some facts are considered to be N-evident for Bacharach as soon as they are observable:

> "One condition for p to be N-evident in S [in a given situation S] is that any normal [agent] thinks of p in S; another, that in S any normal [agent] who thinks of p is able to confirm that p (e.g. through observational evidence available to her normal sense) – that is, p is 'N-confirmed' in S" (ibid, p. 16).

Basically, entailing that normal agents think of p means that they are aware of p because p is a proposition that has come to their mind by the framing process described above. Then, he claims that

> "if (in a situation S) p is N-evident then, if p, a normal knows that a normal knows that … that p. Now suppose that (in S) p is 'N-evident occ F' [i.e. N-evident conditioned by the occurrence of the family F to which the proposition p belongs], in the sense that if p is true then any normal player knows occ F that p. Then we get an analogous bonus of layers of iterated knowledge: it can be shown that if p is true then a normal knows occ F that a normal knows occ F that … that p. If it is also an axiom that normal recognize each other in S, this strengthens to: if P1 [player 1] and P2 [player 2] are normal and p is N-evident occ F then, if p, each knows occ F that … that p; that is, if p is N-evident occ F and true, there is mutual knowledge occ F between P1 and P2" (ibid, pp. 17-18).

That is why, *in fine*, the representation of the situation is not only common among all of the players who face this situation (because their frames are equivalent) but it is mutual knowledge too (ibid, p. 3). Now, because players know they can act on their frames, if a concept, say *k*, instantiates, players know they can act on *k* and accordingly possibly choose the object possessing the characteristic *k*. As a consequence, because each player has equivalent frames and knows she can act on each concept of her frames: each player possesses the same set of act-descriptions, and this fact is mutual knowledge. In this manner, Bacharach (ibid, p. 3) asserts that he "provide[s] an endogenization of mutual knowledge between the players of the game they are playing, described in a certain way".[68] By asserting that each player is a normal player, plus the fact that normal players are like-minded, and by assuming that only occurent knowledge is relevant for the game he describes, i.e. that only observation is a source of information,

---

mentions Gilbert (1981) and Clark and Marshall (1981) to support this statement. He gives the example of shared environment and explain the epistemic consequences of such environment in the following quotation: "the carafe situation is a shared environment situation in which *f* is the presence of the carafe, since it seems that both you and I will recognize the (seeing the carafe) situation as one which both of us will recognize. But this still leaves the question: what enables us to recognize it as this? An appealing answer was first given by Lewis (1969) and Schiffer (1972), based on the notion of 'normality' (Bacharach 1992 gives an axiomatic version). A normal human will not only see the carafe, but will also see the normality of the other co-present normal; lastly, normality has the reflexive property that it is part of being normal to know the perceptual and epistemic capacities of normal people." (Bacharach, 1997, p. 4)

68 Some other characteristics proper to the rules of the games are also assumed to be of mutual knowledge. These characteristics are the following: (i) the "admissibility condition", i.e. if there is only one object which possesses a given characteristic a player must choose that object; (ii) the "preference condition" i.e. players attach utility 1 to the fact that they choose the same object or 0 otherwise; (iii) the number of objects which possess a given characteristic, say, *h*; (iv) the fact that a player can pick a non-unique object at random; and (v) the fact that each player is rational in the sense in which Bacharach defines rationality in this paper (i.e. *via* the different principles of rational choice that I will explain in the following subsection).

Bacharach justifies that the players tend to perceive the same thing in the situation they face and know that. Therefore there is ultimately common knowledge of the common appraisal of the strategic decision problem they face, i.e. of the game.

To postulate that everyone perceives the same thing from the same situation, and mutually know this fact, is very restrictive. It is probably one of the most controversial assumptions made by Bacharach. It may be plausible in very specific contexts. For instance it can be true if players are closed enough, in the sense that they regularly face the situation together and interact. Knowing each other quite well implies that, in a way, individuals know how the other perceives the situation and thinks. In this case this assumption can make sense. It is true for community-based reasoning which induces common inductive inferences with respect to a common institutional context for instance (see Hédoin, 2014, 2015, 2016; and the previous chapter). Such statement can be linked to the hypothesis of mindshaping that will be extensively discussed in the chapters 4 and 5. The mindshaping hypothesis also entails that the social world induces in specific contexts and with respect to specific social situations, a degree of homogeneization of people's mind. However in other cases it is even more overconfident than standard non-cooperative game theory. Within an interpretation of standard game theory, a payoff matrix can be considered as pre-existing before the game, and can be considered as an objective material upon which players have to interact. It is even more restrictive in Bacharach's model to presuppose that each player, as a normal agent, comes to build herself, and independently, a game which is common to everyone. Therefore, we are no longer in a strategic context. Moreover, in 1991, Bacharach never refers to beliefs, as if beliefs were not necessary in his model. We could almost claim that the uncertainty is null.

In Bacharach's second paper on VFT in 1993, an important methodological change occurs. Bacharach no longer makes reference to 'shared' perceptions or to occurent knowledge. He describes a game as a decision problem involving a potentially high level of uncertainty. Games become "open universe problems" (to refer to Binmore's [1987] term). The level of uncertainty is indeed deeper that in Bayesian game theory for Bacharach. He states:

> "Quite generally, a given question may or may not occur to an agent facing a decision problem; she may or may not think of it. This phenomenon gives rises to variations in agents' belief systems of a more radical sort than those to be found in the familiar Bayesian model. In the Bayesian model the different belief states of a person (produced in her by this or that evidence) merely redistribute the weights she attaches to the propositions (or events) of some fixed set (the sigma field of some fixed probability space). Here by contrast, it is the set of propositions about which she has beliefs at all – her belief- space – that varies." (Bacharach, 1993, p. 257)

Hence, the role of beliefs becomes predominant. Bacharach asserts

> "[t]his model gives us probabilities for players' having various belief-spaces. It allows us, too, to say something about the beliefs which a player with one belief-space has about the belief-spaces of others; and this is vital if we are to have a theory of encounters between such players that's worth calling a theory of game between them" (ibid, p. 258)

In this 1993 paper, players have to form beliefs about their available choices – according to their frames, others' frames and accordingly others' subsequent choices and acts.

To understand such postulates, I briefly detail some of the notational and formal aspects of the model:

A Variable Universe game is described by a 5-tuple: $\langle \mathcal{A}, U, \mathcal{F}, V, A \rangle$ with A the act-description set of the players, U their utility function, F the set of families of concepts, V the availability function and A the coverage function. We assume two players *P1* and *P2*.

Those elements are described as follow:

- U is defined on a pair of act description on A such as $U(a_1,a_2)$ is the payoff of *P1* and *P2* when *P1* choose a_1 while *P2* chooses a_2
- R describes the set of players' repertoires *r*. A repertoire *r* corresponds to the set of families of concepts within F that a player handles. *F* is a given family of F in *r*. And if *r'* and *r* are two repertoires such as $r' \subseteq r$, *r'* is a 'subrepertoire' of *r*.
- The availability function V which determines if a given family of concept *F* or a given repertoire *r* instantiate, i.e. if it occurs in the players' mind is defined by $V(r) = \prod_{F \in r}[V(F)] \prod_{F \in \mathcal{F} \setminus r}[1 - V(F)]$ (1)
- The coverage function A is a function from R to subsets of A. $A(r)$ therefore describes the act-description space that a player has when her repertoire is *r*. And *a* is given action within A.
- B_i is the belief space of a player *Pi*. And B_i belongs to $A(r)$.

Bacharach claims that "If player has repertoire r, then for each a in $A(r)$ she believes that she can *a* … and for each *a* not on $A(r)$ she has no belief as to whether she can *a*" (ibid, p. 260).

It means that she cannot believe that she is able to handle an action which is not based on her repertoire (ibid, p. 260). Therefore, if *Pi* has for repertoire *r*: $(a_1,a_2) \rightarrow U(a_1,a_2)$ belongs to B_i iff $a_1, a_2 \in A(r)$.

In addition, each player has a subjective probability for every sub-repertoire of her own repertoire. There is not, in this label of subjective probability, any reference to Savage or Bayes. Those probabilities are subjective in the sense that they are derived, in first instance, from the players' perceptions. Players evaluate from their own perspective, what they assume to be the perceptions of the other players and thus in turn the states space of their beliefs. Players' beliefs are defined by a set of subjective probability functions $\{\pi_r | r \in \mathbf{R}\}$. Bacharach (1993, p. 261) assumes that the subjective probability functions for each player are the same. The subjective probability $\pi_r(r')$ is therefore interpreted as "the subjective probability that a player having repertoire *r* assigns to the other player's having *r'* " (ibid, p. 261). A player cannot attribute to others the same repertoire as her own: "*Pi* implicitly assumes that *Pj*'s repertoire ($j \neq i$) is some

sub-repertoire of hers; that is, if she has *r*, she implicitly assigns probability 0 to his having occur to him any family not in *r*" (ibidem).

This subjective probability is defined as follow: $\pi_r(r') = \prod_{F \in r'}[V(F)] \prod_{F \in r \setminus r'}[1 - V(F)]$ iff $r' \subseteq r$ which implies from (1) that $\pi_r(r') = V(r') / \prod_{F \in \mathcal{F} \setminus r}[1 - V(F)]$.

Hence, because players' perceptions are not necessarily shared, again it may be impossible to derive a common game from players' subjective representations of the decision problem they face. Nothing ensures that players themselves can independently come to perceive the same thing and build a common game. This is indeed the major limit faced by Bacharach's attempt to enrich standard non-cooperative theory by players' frames. The present risk is that no common game can be specified. An intersubjective dimension must exist to guarantee the existence of a commonly understood game. Each player herself computes her own game, in Bacharach's approach, making assumptions about the other players' frames and choices. Therefore, even the game is subjective. Resorting to a social dimension bringing in some respect common understanding and common reasoning among players become determinant to overcome the introduction of players' subjective perceptions. In standard game theory such intersubjective dimension is provided by the hypothesis of common knowledge (see Livet and Schmidt 2014). In this manner, in addition to the knowledge of the rules of the game, each player can ultimately know each other's decision. But with the introduction of players' perceptions, games become open-universe, they are no longer self-contained worlds in which players decide according to fixed rules as determined by Bayes rationality (see Binmore, 1993, 2009; and for more details see chapter 5). To guarantee a certain degree of congruence among players' perceptions and reasoning when players' perceptions of their decision problem are taken into account and games are considered as real world strategic decision problems, focal points, institutional devices, and mindshaping (as bringing homogenization in players' mind) become the only possibility to bring an intersubjective dimensions and *in fine* coordination.

Subsequently, to build a game Bacharach makes a restrictive statement: players must have the same set of act-description of action. Although this postulate justifies the structure of the game, and makes sense because players believe that others' set of act-descriptions is necessarily a subset of their own (and accordingly their own set of strategies is symmetrical vis-à-vis the set of strategies they ascribe to others); it does not justify the fact that two or more players agree on the game they are going to play. That is why Bacharach assumes that players must have the same set of subjective probability towards the others' repertoire. A convergence among players' perceptions is required. Which means that their repertoires are *de facto* equivalent like their strategy space. As each player has her own evaluation of the game she is to play and of the others' perception beliefs and strategies, each player can assess a different game. Such reduction that Bacharach makes can be grounded on a common understanding of the game which is de facto assumed in standard game theory in which the only objective characteristic of the game matter, the players know the rules of the game and act within the rule of the game and possess the same rationality which is common knowledge or at least of common belief. In Bacharach's methodology, none of these hypotheses prevail. Common understanding of the game should be based on common inductive inferences, which are equivalent to community-based reasoning as

explained in the previous chapter. It thus relies as will be also explained in the chapter 4 on mindshaping, to social device such as institutions, conventions that shape the mind of the members of a common community to perceive the same thing and act in the same way. Bacharach however does not ascribe to this justification even if he implicitly endorses such a perspective when resorting to social psychology and detailing his conception of focal point that will be detailed in the section 3.3).

From a formal point of view, Bacharach is therefore constrained to make *ad hoc* assumptions, to set a game. Allowing as many conceptions or perceptions of the situation as there are possible conceptual representations, when considering as he claims, that "different possible descriptions may be strategically inequivalent" (Bacharach, 1997, p. 5), endangers the possibility to build a game. There is indeed a lack of justification from cognitive sciences to ensure the homogenization that Bacharach postulates to guarantee the existence of a solution for the games (see the next chapter and the chapter 5 for one possible justification). Thus, Bacharach has to postulate that *in fine*, the probability of occurrence of a family is independent from the players (1993, 1997). More precisely, these probabilities are independent both across families and across players. Subsequently, frames are only explained by the context even if he recognizes that it is highly simplified. In other words, as the theorist, Bacharach endows players with a set of frames *a priori* according to the decision problem under scrutiny. The existence of a solution depends on these simplifications. We argue that these restrictions could be avoided by grounding the integration of players' mental states, the explanation of the elicitation of players' beliefs and of their reasoning process on another dimension of cognitive psychology, the theory of mind. Adding a theory of mind and in particular the simulation theory in game theory as will be explained in the chapter 4 and demonstrated in the chapter 5, in addition to the inclusion of players' frames of the game, allows to bypass these difficulty that undermine Bacharach's project. Bacharach indeed makes very restrictive compromises to some of his epistemological claims which can seem contradictory with his very first objective even though he refers to Schelling's concept of focal point to justify common perceptions, common beliefs and common act descriptions spaces for instance. As will be exposed in the next subsection common act description spaces can be related to community-based reasoning. However, Bacharach's account of focal point can be in some respect dubious, and can question its social dimension which in turn questions the possibility to interpret the common understanding and common reasoning aspects of his theory as a community-based reasoning. As already emphasized Bacharach had an ambitious task enriching standard game theory with psychological, sociological and philosophical dimensions. He nonetheless has to compose with the framework of game theory and retains some of its features, which explain some of his simplifying assumptions. Though, it is important to note that Bacharach sees himself as a social scientist and not a mathematician (personal discussion with Sugden) who opens game theory to interdisciplinarity and who considerably changed the way players and their rationality are understood in games, notably by acknowledging that theorists' and players' apprehension of games and of rationality differ. In this perspective the models of games he offers in the VFT are objective descriptions based on the players' subjective description of their decision problem. This latter point already provides a significant difference with standard game theory.

3.3. Games' solution and focal points: what principle of equilibrium selection?

As soon as the framed game is defined, or shall we say, the 'objectivized framed-game', players choose strategically their actions between the two kinds of actions previously specified (i.e. "choose one object" or "pick at random").[69] Bacharach (in Bacharach and Bernasconi, 1997, p. 4) claims, "strategies are chosen in a way which is rational in a perfectly familiar game theoretical sense; however, the game that gets played is determined by non-rational (though not irrational) features of the players; these are the players' "frames"."

As in standard game theory, the solution of the games must be an equilibrium, and more specifically a pair of subjective best reply, i.e. a pair of "decision functions" that entails mutual best reply reasoning (Bacharach, 1993; Bacharach, Bernasconi, 1997). This principle is similar with standard non-cooperative game theory. Accordingly, what does Bacharach qualifies as a solution involves exactly the same principles as standard non-cooperative game theory. In the VFT, each player indeed maximizes her subjective expected utility (see, e.g. Bacharach, 1993).

Formally: $EU_r(a,\rho) := \sum_{r' \in \mathbf{R}_r^+} \left[\pi_r(r') U(a, \rho(r'))\right]$, i.e, $a \in \arg\max_{a \in A(r)} EU(a,\rho)$ t'es sur du

It means that if the player *Pi* has the repertoire *r* her conditional subjective expected utility is $EU_r(a,\rho)$ when she chooses *a* and the other *Pj* follows ρ (the meaning of the latter element: ρ is going to be explained).

The theory must '*indicate*' to the players the rational action to choose within *A(r)*, when they handle the repertoire *r*. There is, in this perspective, an "indicator function" σ from the repertoires to the act-description sets, i.e. $\sigma(r) \in A(r)$. σ 'indicates' the action *a* in the repertoire *r* if $\sigma(r) = a$. Now the function ρ is an "r-indicator" from $\mathbf{R}_r^+$ to A, and it indicates for the set of the subrepertoires (having a positive probability) handled by the players when their repertoire is *r* and defined by $\mathbf{R}_r^+ = \left\{ r' \middle| r \in \mathbf{R}^+, r' \subseteq r \right\}$, the action to choose according to those subrepertoires. Accordingly, if ρ is an r-indicator and the player *Pi* follows ρ, for every $r' \subseteq r$ she chooses $\rho(r')$ when she has for subrepertoire *r'*.

The rational play is specified by these indicators, according to the belief-spaces that the players have, which are again, contained within their repertoires, i.e. within their frames. Each rational player follows these indicators, and knows that each other player as a player rational will follow the same indicators. This is stated by the following property:

---

69 In this respect, Janssen (2001) is critical of Bacharach's approach which ambitions by respecifying games to solve the problem of coordination (thanks to the principle of coordination) but in some games players are compelled to choose at random, which is for him "choosing a 'non-solution'." (Janssen, 2001, p. 124)

> "If *Pi* has *r*, and knows that *Pj* follows the *r*-indicator ρ, then *Pi* chooses a best reply modulo *r* against ρ." (Bacharach, 1993, p. 266)

Such statement determines the "Variable Universe Equilibrium" which is in this perspective, for Bacharach, a pair of r-indicator (ρ,ρ), i.e. a pair of best reply modulo *r* against ρ and therefore a symmetric pair (σ,σ).

The connection between the solution of the game and the equilibrium derives from the "general" game theoretic argument that "being an equilibrium is a necessary condition for being a solution" and the requirement for that is "a determinate theory of rational play" (Bacharach, 1993, p. 264).[70] The solution of the game is accordingly a "variable universe equilibrium" (ibidem)

Bacharach, advocates for four principles of "rational decision". The first ones are the following:

i. "A principle of coordination" or, in other words, a principle of "payoff dominance" (Harsanyi, Selten, 1988), meaning that if an equilibirum is Pareto optimal, and if "players have sufficiently strong mutual knowledge with the other of this fact" (Bacharach, 1991, p. 3), it is rational for them to choose the corresponding pair of option and only this one. In a game with multiple equilibria, if players are rational, they play their "part" "in the unique Pareto-optimal equilibrium if there is one" (Bacharach, 1993, p. 257).
ii. "A principle of insufficient reason" entailing that if options are perfectly symmetrical (i.e. if none of the characteristics, concepts or families, allow players to differentiate them), i.e. if they ultimately lead to the same pay-off, it is not rational – in other words, there is no "sufficient reason" – to choose one of them. For instance, "if two options *a* and *a'* are alike in all relevant respects, a solution concept should not picked out one rather than the other" (Bacharach and Bernasconi, 1997, p. 11). In matching games these relevant aspects for Bacharach are families and frequencies (that is, the probability for families to instantiate).

The two others principles explaining and justifying the two principles of equilibrium selection above, and considered to be "general feature of VFT solutions" (ibid, p. 10), are: (i) the "principle of rarity" and (ii) the "principle of availability" (Bacharach, 1993; Bacharach and Bernasconi, 1997). The former means that a player prefers to choose a rare object. By doing so, and against several identical objects, she maximizes her payoff: "Ceteris paribus, players prefer to pick an object which is rarer" (Bacharach, Bernasconi, 1997, p. 10). The latter implies that, in some circumstances, choosing a rare object can be too risky, so a player prefers a more available object: "Ceteris paribus, a player is more inclined to pick an attribute which is more available" (ibid, p. 11). Bacharach and Bernasconi (ibidem) qualify games in which an "availability preference" competes with a "rarity preference" as "tradeoff games".

---

70 See Bacharach (1987) for this connection between the equilibrium, the solution and the principle of rational play that has to be found in the theory of games for such theory to be worth calling a theory of rationality in interaction (see also von Neumann and Morgenstern, 1944; Luce and Raiffa, 1957).

To understand a tradeoff game I quote an example given by Bacharach (1993, p. 269):

> "[In a matching game] a grain-aware player should mark the wavy block if the probability that the other is grain-aware exceeds the relative rarity of red blocks, and should mark the red one if it falls short of it … [but] it may be that her partner is grain-unaware. In that case he doesn't have marking the wavy blocks on his list of options, and there is no reason to think that he will mark it ... Indeed, picking a red will appear to him the best option to coordinate on. So it is riskier to mark the wavy block than to pick a red, because of the relative availability of the grain-pattern concepts"

As emphasized in this quotation, the probability that a player perceives the grain of the object may be too low compared to the probability that she perceives the color. In this case, the player chooses the more available object, i.e. she picks at random one of the red ones. By this way, she maximizes her subjective expected utility.

However, trade-off games may exhibit failures of coordination. The trade-off problem entails that in some cases players do not successfully coordinate. In that circumstance, Bacharach do not succeed in rationalizing the "payoff dominance principle". This is according to me, why from a methodological point of view, Bacharach later develops the TR theory drawing on some of his underpinning statements within the VFT, and in particular the framing phenomenon. We will explain in detail why in the next section.

The principles of equilibrium selection and in particular the principles of rarity and availability, emanate from Bacharach's conception of salience and focal points. Those principles correspond to the game theoretic translation of the psychological phenomenon of framing. To "outline a rigorous theory of salience" (Bacharach and Bernasconi, 1997, p. 2), Bacharach (1993, p. 256) purports "to solve the unsolved problem of how to give a game-theoretic rationale for choosing the salient". For him, "a satisfactory theory should show whether these choices [i.e. choosing the salient options] are indeed rational and, if so, why and what their rationality has to do with salience" (Bacharach, 1991, p. 5). Accordingly, he claims,

> "I will show how the framework of MGs [i.e. Matching Games] makes possible precise definitions of salience concepts – even if it is not easy to say which is best. I shall show, secondly, that we can say precisely what the connection is, in MGs, between choosing the salient (in a defined sense) and choosing rationally" (ibid, p. 34) .

In Bacharach's account of a theory of focal point, being a salient option should not be "reason-giving": "in order to explain why rational players choose salient options, we need never invoke the salience of an option" (ibidem). To the contrary, it should be the result of a rational choice. That is why the VFT "derives focal-point play from an explicit model of salience and from well-defined rationality postulates, and it also makes predictions where the traditional theory is silent" (Bacharach and Bernasconi, 1997, p. 37). From that prospect Bacharach distances himself from Schelling's conception of focal point. For the latter it is because an option is a focal point that it is rational for the players to choose it. Salience is reason giving in this case. Such difference is according to me mainly explained by the status of the games in Bacharach's VFT. They are indeed, as already emphasized, the mathematical translation of players' appraisal of real world strategic decision problem. Such mathematical translation encompasses principle of rational play

that belong to the game theoretical tool-kit let say. Recall that in Bacharach's conception of game theory, a solution must be strictly defined; it is the core of the axiomatic approach of game theory. So how can "a determinate theory of rational play known by a rational player indicates to the player the right action to choose, that is, brings such a player to believe that a certain action is the right one" (Bacharach, 1993, p. 264) be compatible with games that are the output of a subjective and unconscious process?

To define salience, Bacharach draws on Lewis' definition (1968), i.e. salient options possess two characteristics: (i) 'conspicuousness' or 'noticeability' and (ii) 'uniqueness'.[71] According to Bacharach (1993, p. 270) these two dimensions must be distinguished: "[conspicuousness and uniqueness] are logically independent, although they frequently appear in amalgam in discussion of salience". He does not assert that they are totally independent but he insists on the fact that they involve different mechanisms in individuals' process of reasoning. In addition, because the definition of salience is "theory-specific" for Bacharach, he has to make the link with framing.

In his account of framing in the VFT,

> "a frame is said to be salient if it has a strong tendency to be operative; some frames are more salient than others; and the salience of a given frame depends on the context. The salience of a frame depends on the salience (similarly defined) of its constituent concepts, and in particular of any constituent classifiers: for example, the colour-shape frame is salient if both the colour and the shape classifiers are" (Bacharach, 2001a, p. 5).

Besides, Bacharach affirms "frames can have more or less power or potency to influence decision" (ibid, pp. 5-6). The last quotation is explained by the fact that in VFT, players must think of salience strategically: "the essential depth assumption of VFT is that players think about salience strategically: *i* asks herself how likely it is that *j* has noticed what she has and so has the same options as her" (Bacharach, 1997, p. 39). Indeed, as Bacharach emphasized, in "[t]he VFT assumptions that saliencies are shared and that complete-frame players accurately perceive this are inessential simplifications … Saliencies may not be shared … and whether or not they are, players may assess them with bias" (Bacharach and Bernasconi, 1997, pp. 38-39).

The distinction of the two dimensions of salience – i.e. conspicuousness and uniqueness – raises changing explanations in Bacharach's different contributions to the VFT.

In 1991 Bacharach claims that the two dimensions of salience do not belong to the same phases in the players' reasoning process. That is why "to explain how salience brings about choice it is therefore necessary to decompose it and to explain the operation of the separate components in these two quite different mechanisms" (Bacharach, 1991, p. 34). Conspicuousness relies on the unconscious phase whereas uniqueness results from the conscious phase. More precisely, a characteristic of the objects under scrutiny in the matching games, is conspicuous if the predicate denoting this characteristic is 'highly available' (ibid, p. 34). This is what Bacharach calls the "salience-1" and by definition "salience-1" belongs to the unconscious phase. "Salience-2"

71 In the VFT, conspicuousness is more specifically formalized by the availability function (see also Janssen, 2001, p. 126).

implies that such characteristic describing the objects is the only one such that "for some highly available family *F* [the player] knows that it's the odd one out in terms of *F*" (ibid, p. 36). In other words, for Bacharach, 'salience-2' "raises the expected utility of the pair in which both choose the salient option, and so give a game-theoretic reason for choosing it (expressed in the principle of coordination)" (ibid, p. 35). It implies that *Pi* thinks of the *b* that can be salient for the other players. A *b* is salient-2 if *Pi* knows that *b* is also salient for *Pj*. Therefore 'salience-2' belongs to the reasoning phase.

However, Bacharach no longer dissociates the two dimensions of salience by the unconscious and the conscious phases in the subsequent versions of the VFT. After 1991, salience relies entirely on the availability of frames and is especially linked to the properties or attributes of the objects. And these two dimensions may be linked to different objects. If this is the case, Bacharach identifies a "trade-off" game (Bacharach, 1993, p. 270). In this type of games, a "trade-off theorem" allows players to make a rational choice, thanks to the "principle of availability" and "principle of rarity" mentioned above. As a consequence Bacharach distinguishes a "primary salience" and a "secondary salience".[72] These terms have however the same sense as salience-1 and salience-2. For him, the former is "the traditional notion" (Bacharach and Bernasconi, 1997 p. 38) – i.e. noticeability and conspicuousness of a property; whereas the latter entails that an option is of "secondary salience for a player if she thinks it has primary salience for her co-player" (ibidem). Although Bacharach gives two dissimilar justifications for the difference he makes between conspicuousness and uniqueness, the underlying idea remains: one dimension of salience is strategic (which merely means that one dimension relies on the unconscious phase and the other is explained by the deliberative phase, i.e. the strategic phase).

Two possible interpretations of focal points and salience could compete in the VFT: (i) a "naturalistic" and (ii) a "community-based". The former implies that salience ensues from "objective and natural properties of some entities (events, strategies, outcomes)" (Hédoin, 2012, p. 2; referring to Lewis' naturalist interpretation of salience). In this account salience is an objective fact, directly observable in the environment, and above all, understandable by everyone in the same way[73]. By contrast, social institutions explain salience in the community-based interpretation. Individuals belonging to the same community inherit the awareness of some institutional facts from this community. Furthermore, they understand and interpret these institutional facts in the same way, and know that, but they must be able to recognize each other as member of the same community (ibid, p. 16). The two competing accounts of salience in VFT are stressed by a methodological change occurring between 1991 and 1993 and onwards).

When Scazzieri refers to Bacharach's account of framing and salience, he emphasizes that "framing shows the *interplay between structural principles and evolutionary (historical) principles*" (Scazzieri,

72 Bacharach refers to Metha, Starmer and Sugden (1994a,b) who propose this terminology.

73 Referring to Postema (2008; pp. 43-44), Hédoin (2012, pp. 18-19) claims that naturalistically, "the salience of an entity […] is taken to be a brute, objective and explicit fact. Salience is then a characteristic constitutive of an entity which can be recognized "straitforwardly" by agents"

2008, p. 197; my emphasis).[74] While these "natural associations" between frames and context are partly explained by history, e.g. culture, it means that they are socially based; and accordingly, we can interpret salience as "community-based". If players' subjective frames are partly explained by precedents, players can ground their reasoning on what they learned about coordination devices inherited from this culture, and accordingly the way others generally perceive the coordination context they face. Since players think of salience strategically, it means that they think about what is salient for others; they have to refer to institutional dimensions of the culture in which the other is embedded; hence they use the knowledge their culture gives them. Indeed, players cannot think of others' perceptions if they are not from their own background (Bacharach, 1991, 1993, 2001; Bacharach and Bernasconi, 1997), which implies, in a way, that each player's perception must be embedded in a common conceptual framework, i.e. in a common cultural legacy. Moreover, all of the VFT models incorporate language. Players must share a common language to play a game. These facts reinforce the suggestion that salience may be "community-based".

However when salience depends on the properties of the objects only (i.e, from the 1993 paper on the VFT onwards), such postulate contradicts the "community-based" account. A focal point in the community-based account is meant to generate symmetric reasoning among players. Every player expects that everyone will conform to focal points, and that this statement is of common knowledge (Hédoin, 2012, p. 2). Asserting that saliencies may not be shared is contradictory with this account. Players are not necessarily symmetric reasoners because they may not have the same subjective representations.

In summary, the numerous variations of Bacharach's VFT illustrate his attempt to make his theory both a "theory of rational play in games" and a more realistic one, without endangering its validity. His ambition is to emphasize how framing is a promising framework for game theory. Indeed, he argues that VFT belongs to "a much larger research program for game theory: the study of the perception by players of the problem that confronts them and the condition in which this issue in games of varying form" (Bacharach, 1991,p. 38). It is a promising framework in particular from a methodological point of view, to *rationalize* the payoff dominance principle, as a mathematical translation of its psychological counterpart: the salience of frames. However, what is underlined in this section is that Bacharach's VFT sheds lights on the limits of explaining strategic interactions with subjective and then intersubjective and social dimensions within a standard game theoretical framework. Somehow, like in standard game theory, objective structure or objective determinants in players' strategic interaction must be postulated. That is why Bacharach is entangled in two different viewpoints. He takes on the position of both a player and of a theorist at the same time. It seems that he looks at frames and subsequently, framed games, from both angles. This echoes the important distinction pointed out by Schumpeter (1942) that prevails in the social science between the 'observer' and the 'observed'. We suggest on the contrary and this will be the epistemological stance adopted in the model of games proposed in the chapter 5 that a theory of choice in games explaining how a player should rationally choose,

74 This account suggests that a meta-theory of framing in Bacharach's VFT is required to explain why frames are structured in one way or another.

given the nature of the game and of the other players should lay on three premises: (i) determining the problem faced by each player, (ii) defining a mechanism consistent with the rationality of the player giving her a reason to play a specific strategy, and (iii) assessing the final outcome, from our perspective as theorists.

What is interesting to note on Bacharach's VFT is the fact that he attempts to enrich standard game theory, to augment it. The methodological difficulties he encounters translate such attempt, i.e. to broaden game theory and make its frontiers moving and at the same time staying within. However his contribution leads to numerous and very fruitful questioning with respect to these limits that game theory as a general theory imposes. In particular, it exhibits the difficulties of making games and game theory as open systems instead of considering them as self-contained worlds as they both are standardly. Games become "open universe games" (Binmore, 1993, 2009). Open universe games are games in which the representation of the players have been integrated so that the possible states of world handled by the players and describing the games are potentially unlimited and thus do not correspond to a close state space in which the rule of Bayes and in which subjective expected utility theory can be applied. Another form of decision theory under this form of uncertainty must be applied. This is the difficulty to which Bacharach is confronted and this is why he made the restrictive compromise we emphasized in this section to close the state space of player's beliefs and strategies. The methodological consequence of an open universe problem will be explained in detail in the chapter 5 and we will present in this chapter the methodological ground on which we rely to define a principle of strategic rational choice within this frame without the difficulties that Bacharach met.

He also goes beyond Bayesian game theory by allowing non-conceptual omniscience and assuming that players may not know that they don't know which has for consequence that there is no common-knowledge about the uncertainty within the game (i.e. common knowledge of priors). In this perspective it authorizes different complex structures of beliefs and games are thus no longer self-contained worlds.

4. Bacharach's theory of "Team Reasoning": a theory of cooperation or of coordination?

In order to show how and why it is rational to coordinate analytically, i.e. to rationalize the payoff dominance principle, Bacharach constructs a different mode of reasoning than the standard one: The Team Reasoning (henceforth TR), (Bacharach, 1995, 1997, 1999, 2006). To rationalize the payoff dominance principle, the nature of choices changes. More specifically, players become "profile directed reasoners" instead of "means-end reasoners", i.e. best-reply reasoners. They are supposed to cooperate for the team, which in other words, means to coordinate each other's decision and action in order to reach the most satisfying issue for the team as a whole. Hence, Bacharach's main aim is to justify why players switch from one mode of reasoning to another.

In his very early thinking on the VFT, Bacharach (1991) explicitly argues that the payoff-dominance principle should belong to the analytical tool kit of game theory for solving the

problems of indeterminacy. According to him, it offers a principle of equilibrium selection in cases of multiplicity of Nash equilibria. The VFT is a 'respecification theory' that attempts to translate a coordination game into another type of game ensuring the existence of a unique Pareto optimal Nash equilibrium. Recall that such Pareto Optimal Nash equilibrium is the theoretic translation of the option – the profile of strategies – that the players perceive as the most salient, i.e. the focal point. However, even with such respecification, some coordination games still raise a difficulty when there is a trade-off problem for instance. In this case, players can fail to coordinate. I argue that this is one of the main theoretic explanations of Bacharach's switch of interest from the VFT to the TR. TR indeed rationalize coordination on focal points (Bradsley et al., 2010; Butler, 2012; Colman et al. 2014; Faillo et al., 2017; Bardsley and Ule, 2017; Pulford et al., 2017).

As emphasized by Sugden and Zamarrón (2006, p. 615)

> "A number of theorists have tried to assimilate focal points to game theory in a way which represents the players as reasoning together … theorists propose models in which players solve problems of equilibrium selection by choosing the equilibrium (if there is one) which is uniquely best for both of them. Some theorists treat this principle of equilibrium selection (often called payoff dominance) as a primitive. Others treat it as an implication of a distinctive mode of rational choice (team reasoning or we-reasoning) in which individuals "identify" with a group (in this case, the group comprising the two players) and find the combination of actions which maximises that group's shared objective; each individual then chooses her part of this combination." (Sugden, Zamarrón, 2006, p. 615).

Colman (2014) also supports the same claim, for him TR is a solution for the payoff dominance principle.

In order to solve this payoff dominance problem, Bacharach had to explain why in such coordination games players decide to cooperate in order to reach the Pareto optimal equilibrium and for that, why they decide to act as team members. Such reasoning process emanate from the players themselves and not from an outside observer or a leader which independently and unilaterally impose to the players to act as a team and dictate the action to accomplish for each member of the team (Gold, 2018, p. 344). Bacharach attempts to justify how the problem of coordination that players face lead them to see the decision problem as a collective decision problem which involves a degree of interdependency such that they have to solve it together as a team and not as separate and independent individuals. Like Schelling, Bacharach thus asks the question of the degree of interdependence that is required in coordination when games are seen as real strategic interactions, i.e. as open systems and no longer as self-contained worlds in which players must act according to the fixed rules of game theory and acting according to their knowledge of how the model functions. TR conveys the idea that players can acknowledge that solving a game implies a degree of interdependence such that it must be done together, i.e. that coordination must require, to a certain extent, cooperation. In this perspective, we see the similarity with Schelling. TR is a mean to entail the convergence of players' intentions, decisions and actions.

Bacharach's TR theory also shows again how he is inclined to an open system account of game theory as he makes the link with philosophy, social ontology and social psychology. For instance, thinking on the relation among individuals and the collective is a central topic of social ontology. The link between the collective and the individuals that form this collective, how it organizes the interaction among this collective, the capacity to coordinate and cooperate and how it influences the players' mental states such as their preferences, intentions, objectives, etc, are topics of social ontology. We find here a strong echo to Schelling's concern in his reorientation of game theory. And this marks the strong need for Bacharach that the other social sciences impact on the understanding of the resolution process of a game and on the understanding of players' reasoning in games. We will see that the TR adds in game theory another dimension of human cognition than the VFT. When both the VFT and the TR are seen as a coherent unity, as a broad theoretical program encompassing specific models (each requiring specific simplifying assumptions), they ultimately embrace players' cognition in its complexity. They ultimately offer methodological solutions to the problem of coordination in game theory. However as with the VFT, Bacharach faces numerous methodological difficulties that are the price to pay to enrich game theory from an interdisciplinary perspective. He does not only import concepts from other social sciences without challenging game theory, in the fashion of psychological game theory for instance. On the contrary, he formalizes players' process of reasoning, integrate players' mental states by relying on many dimensions of individual decision making as studied in other social sciences (Stirling and Tummolini, 2018, p. 471). Gold (2018, p. 344) indeed declares that Bacharach "regarded team reasoning as a "mental activity", which is a causal process that determines, or partially determines, choice." Therefore, as players' mental states and reasoning process are integrated in games, these games become open-systems (Binmore, 2009).

4.1. Drawing boards and evolutions

The different published and unpublished papers on Bacharach's conceptualization of the TR theory exhibit several methodological changes. He first gives an explanation of cooperation that seems quite counterintuitive in a standard game theoretical framework, and again, like in the VFT he progressively makes conceptual and methodological changes. It emphasizes the constraints imposed by standard non-cooperative game theory regarding Bacharach's account of TR. Some contradictions appear between an attempt to formalize games as open-systems and standard game theory which entails a closed-system which operates according to fixed rules. One of these contradictions is for instance that Bacharach constantly tries to endogenize the probability that players, in a given interactional context, adopt the TR, and then eventually cooperate. This endogenization problem is a matter of framing.

Regardless of the model of TR proposed by Bacharach, as soon as a player team reasons she is "guided by the team's objective" (Bacharach, 1999, p. 118). Team reasoning is a "profile directed reasoning". The team reasoners first identify a profile of strategy from a collective point of view, which define 'group choice function' and then they play their part in this collective profile of strategy, they play the strategy available for them in this group decision function (Bacharach,

1995, p. 3). In other words they identify the profile of strategies that would be optimal for them as a whole. TR entails the pursuit of collective interests. The 'group choice function', which is a collective 'profile of actions', stems from a collective utility function. It is the sum of individual utilities and a Paretian ranking (Bacharach, 1999, p. 120) of the "utils" – or payoffs – (i.e. numeric values) in the objective game matrix. Bacharach's TR account therefore necessitates that players' interest must be perfectly aligned. And the team objective derives from individual payoffs: collective preferences are deduced from individual preferences in some way (see Lecouteux, 2018). Team reasoning always requires two phases – or in Bacharach's words, two "activities": an "evaluation activity" and a "selection activity" (Bacharach, 1997, p. 12). Hence, a player who team reasons first computes the best profile of actions for the team – regarding the collective utility function – then chooses the "component" of this profile "under her control" (Bacharach, 1999, p. 120), and finally performs its component.

A pervasive characteristic of team reasoning in Bacharach's models is the fact that team reasoners, to some extent, remain "strategic thinkers"; they have to form beliefs on other players' possible behaviors. They do not cooperate systematically; it depends on the probability that the others cooperate, i.e that the others also team reasons. TR does not impose unconditional behaviors. In order for players to team reason, and accordingly to accept its validity (from a practical perspective), they must have confidence that the other players also team reason.

Although Bacharach is consistent throughout his work on the consequences for players' reasoning to adopt TR, he has made slight modifications to explain the conditions enhancing TR, between 1995 and the later models (1997, 1999, 2006). The paper "Cooperating without communicating" in 1995, draws the underpinning statements of TR, and some intuitions of the underlying ideas which will be clearly developed in his later work, even if cooperation does not rely on framing. It is my opinion that his 1995 paper is particularly relevant because it underlines how some of the conceptual and methodological choices, which are problematic in the model proposed in 1995 – considering Bacharach's objective to formalize realistic and practical principles of reasoning – will be pervasive throughout Bacharach's models of TR. Moreover this paper clearly exhibits, even more than the others, how Bacharach's account of TR seems in some relevant aspects hardly compatible with a standard game theoretic framework. We shall stress that this problem of compatibility is mainly explained by the fact that Bacharach's wants after all to justify and eventually explain that realistic players, in some strategic contexts, naturally tend to team reason. And to propose a realistic collective mode of reasoning, Bacharach purports to draw the TR on framing after 1995. Subsequently, his TR adheres to the same logic as VFT, i.e. showing how players' subjective representations induce various possible games, and then, possibly, various modes of reasoning: either individualistic reasoning or team reasoning. However, in TR, players' frames concern the way they perceive themselves rather than the context *stricto sensu* (even if they are linked). Nevertheless, like in VFT, subjective representations of players may induce a re-specification of standard games. Depending on the way players perceive themselves, they re-specify the initial or objective matrix (i.e. as defined by the theorist). Again it means that players' perceptions and more specifically players' appraisal of their strategic decision problem may be different from the ones of the theorist. When the theorist sees a game between two or more independent decision units, players see a coordination problem in which the unit of agency is a team: the collective formed by every players.

In the 1995 model, Bacharach formalizes two types of players: (i) the "fellow member reasoners", i.e. team reasoners or cooperative reasoners, and (ii) the "best reply reasoners", i.e. standard individualistic players. This ability to reason cooperatively is explained by "natural types". Someone who is a "type T" is a "fellow member reasoner" and has a natural tendency to cooperate with other "type T" players, and someone who is "type T"' is a natural best reply reasoner. Nonetheless, a fellow member reasoner cooperates if she is 'sufficiently' sure that her co-player reasons in the same way and cooperates (Bacharach, 1995, p. 1). Therefore, players must have the ability to recognize each other. They must be "good enough at recognizing each other" (ibid, p. 2). Such need for a recognitional capacity seems hardly compatible with a static Bayesian game theoretic framework in which the model is grounded. A 'fellow member reasoner' has the ability to switch from cooperative to standard choices depending of the type of player she faces (ibid, p. 8). When a fellow member reasoner faces a player who adopts a BR-reasoning he is "rationally forced to play a best reply to her expectations" (ibidem). In this model, like in Bayesian game theory, Nature moves before the game, and twice. The first move induces that players observe their own type (they can be of type T or T'). The second moves reveals to players the others' "demeanour", i.e. their behavior. There is an objective joint probability distribution over players' types and players' "demeanour" (ibid, p. 12). Then players are able to choose strategically to cooperate or not according to their beliefs providing this joint probability distribution. The formal framework in which this model is embedded is static. As with epistemic game theory the game is settled before the fun goes. For a detail discussion on the methodological difficulties of epistemic game theory and Bayesian game theory see the chapter 1.

Bacharach does not seem satisfied by this standard conception of games and especially by two of its central assumptions. On the one hand, he disagrees with static and preexisting types justifying either a cooperative or a selfish nature, whatever the circumstance under scrutiny, i.e. whatever the structure of the strategic decision problem that players face. On the other hand, he is also dissatisfied with the prior beliefs imposed by the objective distribution probability (which thus implies common priors). Recall that Bacharach attempted to go beyond Bayesian game theory in the VFT (see the previous section). This epistemological issue partially explains that in 1997 Bacharach no longer resorts to "natural types" to explain cooperation but to frames. In fact, the link between VFT and TR is explicitly made, and well developed, in this 1997 unpublished paper and a "variable frame model" explains "We-reasoning", i.e. TR (Bacharach, 1997, p. 2). Bacharach however still refers to the concept of "mutual recognition", i.e. "the psychological capacity of persons with a certain disposition – e.g. to cooperate in some sense – to recognize each other" (ibid, pp. 2-3).

In the 1997 model, instead of applying frames to context description, frames concern the way players perceive themselves and these different perceptions affect their respective reasoning. Bacharach presents three types of frames: (i) "I frames", (ii) "W frames", and (iii) "S frames". I and W frames are qualified by Bacharach as "basic" frames entailing a "simplistic" way of reasoning (ibid, p. 2). They imply "I/he" concepts and "we" concepts respectively, and induce reasoning of the sort "what shall I do?" and "what shall we do?" (ibid, p. 5). Players in "I frames" are individualistic and instrumental reasoners, i.e. they use a "BR-reasoning", whereas players in "we frames" are profile-directed reasoners. When players are in "we frames", they respecify the objective game, i.e. the initial matrix, in a "group payoff matrix" which is basically a formal

representation of the collective "utils" (ibid, p. 11). The fact that "we frames" induce simplistic modes of reasoning raises a problem in certain types of games involving either a conflict of interest or, like in the prisoners' dilemma, mutual defection as the only possible equilibrium. Indeed, a player in a basic frame and presently in W frame cannot think that her co-players can reason in another way than her.[75] If such player in W frame cooperates in the prisoner's dilemma for instance, while the other plays the dominant strategy, i.e. defection, the payoff both for herself and for the team will be the worst.

> "if the objective of the W player is to achieve the group best, she does not necessarily achieve her objective by playing A [e.g. cooperate]. … if her coplayer chooses B [e.g. defect], she will do better not only for herself but also for P [the team composed by the two players] by choosing B too. But the limitations of her frame – her blinkeredness – preclude such considerations" (ibid, p. 13)

To the contrary in 'S frames' (the "superordinate" frames) players conceive the situation both in "I/he" and "we" perspectives. The S frame encompasses the two simple frames. In this case, players can handle the two "associated patterns of reasoning" (ibid, p. 14). Bacharach claims that when an agent is in an "S frame" "she thinks about the possibility, and consequences, for her choice, of her co-player's in each of the subordinate frames", i.e. either in W frame or in I frame (ibidem). Accordingly, a player in an "S frame" can evaluate both groups and individual payoffs (ibid, p. 15). Reasoning within "S frames" allows players to evaluate the respective risk of reasoning individualistically or collectively, they can handle the "two cautious evaluations: personal evaluations and group evaluations" (ibid, p. 16).[76] S frames therefore conveys the idea that players can dissociate what would be their personal interest from the collective interest and that in some circumstances, they can see a conflict of interest between the personal and the collective. It portrays players that are able of reflexivity.

To set out the idea behind the S frames, Bacharach proposes the following schema:

75 In fact, Bacharach asserts "the agent's frame 'blinker' her: it is too narrow to enable her to see a highly relevant possibility. This possibility is that her coplayer is in the second simple frame" (ibid, p. 14).

76 This contribution echoes Bacharach's later work on the level-k theory (see Bacharach and Stahl, 2000).

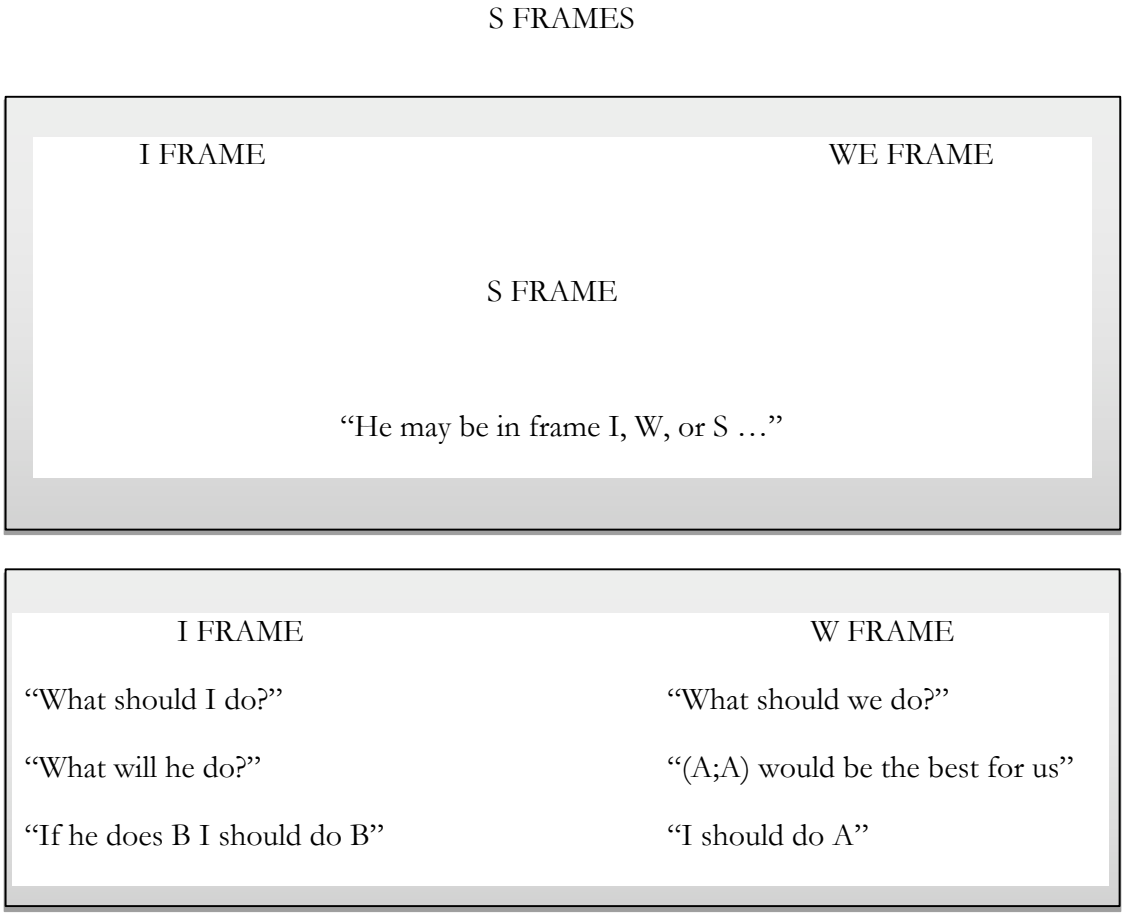


Figure 2: The Superordinate Frame (Bacharach, 1997, 14)

Formally, Bacharach assumes a probability distribution over each frame in the set of players. Besides, Bacharach asserts that "a player in S [frames] has correct probabilities for her co-player's frame" (ibid, p. 15). She knows the probability distribution and knows what the other players in each of these cases can play so her utility is traditionally a subjective (and individual) expected utility.

Nevertheless, for Bacharach this model does not fully explain why it is rational to team reason. Indeed, in "S frames" TR remains partly an individual and instrumental reasoning:

> "the present model fails in an important task, namely to show the possibility of rational we-reasoning. For it might be thought that when the agent is in the sophisticated frames, S, and looking at things from the group perspective, she is not we-reasoning, but simply guided as an individual by 'group altruism'. She works out, on the hypothesis that her coplayer is in a certain frame and also will take a corresponding choice, what act of her own is best for the group interest. This sounds like ordinary instrumental I-reasoning with group regarding goals" (ibid, p. 25).

The limit Bacharach identifies is explained by the fact that to fully adopt the TR, players must undergo an "agency transformation." It is this "agency transformation" which leads to the TR (in Gold and Sugden, 2006, p. 90), and to the respecification of games. As he later asserts

> "having the group interest at heart does not … suffice to explain … cooperation … something more happens … This something is 'agency transformation'. The key to my explanation is that agency transformation involves not only a transformation of payoff but also a transformation of reasoning" (ibid, p. 90).

When such agency transformation operates, the "unit of agency" in games is no longer individual players but the set of players as a team. Quite obviously, players in "S frames" do not undergo this agency transformation. Besides, frames are supposed to be "non integrable". Bacharach (1997, p. 25) emphasizes "the psychological impossibility of simultaneously seeing the problem from the 'we' and 'I/he' point of view". This statement contradicts with the concept of "S frames" since they both entail "I" and "we" concepts in the players' mind. Hence this concept is dismissed in the 1999 model.

In addition to the deletion of "S frames", the 1999 model makes the relationship between players and the collective more complex. Multiple teams can compete and players can belong to different teams. However, each player is active in only one of the team to which she belongs. The underlying idea is that with an important number of individuals interacting some of these individuals may fail to cooperate. It is an unreliable context. From the viewpoint of a player, because some other players may fail to be active in her team (as they belong and act for other teams), in order to maximize the collective utility of her team, she has to take into account who is indeed active in her team. She has to consider the "co-members' unreliability" (Bacharach, 1999, p. 119). In this way, team reasoning becomes "circumspect" (ibid, p. 118).

A "signal" specifies for each player in which team she will be operative. More precisely, each player receives a signal telling her: (i) her "participation state" (i.e. the team in which she will be operative) and (ii) the probability distribution over the others' "participation state".[77] The set of signals plus the set of others' states is given by an objective joint probability distribution (ibid, p. 122). According to her signal, a player "'finds herself' participating in a certain team: her participation is not a choice, nor otherwise endogenized" (ibidem). And finally a function, labeled "a protocol", ensures the link between players' signal and their available and corresponding actions (ibidem). Basically, a protocol is a decision function. A team reasoner has to determine a "protocol" for her team, i.e. the set of individual actions (or the set of individual protocols) leading to the highest collective expected utility, and to do that, she needs to know what the other players will do, i.e. if they lapse to be active in her team, or not (if they lapse they will be active in another team) (ibidem). Consequently, an equilibrium in this model is a profile of protocols, which is optimal for the teams (ibid, p. 124).

---

77 Bacharach explains his choice to refer to a "state" and not a "type". He explicitly recognizes that "type" may be a confusing term in the following quotation: "[w]e can think of *i* as being one of two types in the sense of incomplete information game theory: participating, in which case she team reasons for *M* [the team], or lapsing. But since "type" may misleadingly suggest a permanent trait, I shall speak not of i's "type" but for her "state" " (ibid, p. 121). Such position echoes the limits that were identified in his 1995 paper, i.e. to resort to natural trait.

Following this idea of unreliability, Bacharach (in Gold and Sugden, 2006) finally identifies three types of contexts of team reasoning, with respect to the available information that the players have.

i. In the "Simple coordination context", the entire set of agents in the game adopts the team reasoning. If everyone in the set team reasons the outcome is Pareto optimal – the best profile in terms of the team utility function is reached (Bacharach, in Gold and Sugden, 2006, pp. 122-123). A condition is required, all the players must know the payoff structure (ibid, p. 123), and must be sure that everybody team reason.

The problem of failure leads to two other contexts inducing a "restricted team reasoning" and a "circumspect team reasoning".

ii. In the "restricted coordination context" several agents fail to team reason. In this case only a subset of agents team reason, the others are considered as the "remainder". This time, the team reasoners have to do "as best as they can between them, doing without the non-team reasoners" (ibid, p. 127). Three conditions about the information structure are required: the players must know the structure of payoffs, the team reasoners must know the actions of the remainder, and they must know how many players constitute the remainder.

The epistemic requirements of this context explains why Bacharach asserts that "the restricted team reasoning still has two serious limitations: first, for it to be feasible the reasoners must know who is and who isn't a team reasoner, while in practice there is very often great uncertainty about who is and is not governed by a choice mechanism" (ibid, p. 130); and second, as specified above, the choices of the remainder must be known (ibid, p. 131).

iii. The "unreliable coordination" context is a generalization of the previous one and is defined exactly as in the 1999 paper. There is a probability of failing and a determined profile of actions for those who failed: the "default choice" (ibid, p. 131). In this context, team reasoning and more peculiarly the "circumspect team reasoning" is a valid way of reasoning only if all the members of the team know the payoff structure, the probability of failing and the default choice for those who failed (ibid, p. 133).

4.2. Is cooperation naturally or "interactionally" based? How can multiple selves be conciliated?

A tradeoff between standard reasoning and team reasoning exists in every models of TR proposed by Bacharach. We suggest that this struggle stems from two intertwined aspects. First, players possess a predisposition in a way or another to cooperate, like if it was in their 'background'. This implies that even if the games' outcome relies on strategic interactions, cooperation is weakly explained by these interactions. In 1995 cooperation relies on natural types. In the other models, it is explained by cognitive and psychological predispositions. Some of the time, i.e. depending on the game, collective cognitive representations take hold over

individualistic ones. These presuppositions imply that Bacharach has to justify why players (or some of the players) who are predetermined to cooperate, do not. Second, players can have dual or multiple selves (depending on the models) and Bacharach has some problems to reconcile players' competing selves. This tradeoff between (i) the two modes of reasoning, and (ii) players' competing selves, raises some difficulties and, again, there are different methodologies to bypass them.

In that perspective, in the 1995 model, two elements must be underscored. On one hand the use of the term "natural types" suggests that players' capacity to be cooperative is an intrinsic capacity, a quasi-genetic characteristic. Players of type T are, we could say, predetermined to cooperate. Bacharach (2006, chapter 3) believes that natural selection favor cooperative individuals. For him without cooperation our specie would have been unable to survive. The mindshaping hypothesis developed in the next chapter and in the chapter 5 is grounded on a similar claim. Bacharach's social philosophy goes further by supposing, like Schelling, that the more a society is structured through institutions for instance, as coordination devices, the more individuals can successfully interact, i.e. coordinate and cooperate. On the other hand, players do not cooperate invariably. To justify this, Bacharach (1995, p. 17) confesses: "the type of reasoning I have been modeling may become detached or abstracted from the specific natural types which were its seeding ground". Two consequences follow. Firstly, Bacharach models an "outward sign", minimizing the capacity of types T to recognize the other types T, to justify that some of the time types T do not cooperate with other types T. Secondly, when types T encounter best-reply reasoners, they must play against their natural tendency, since they are forced to adopt a standard individualistic reasoning. It is in such ways that players are "abstracted" from their natural types. For Bacharach the problem of this model can be resolved by the integration of framing:

> "[a]n alternative development would make T membership a variable element in players' frames in the sense of variable frame theory (Bacharach, in Binmore et al., 1993): that is a player might or might not think about the game in terms of whether she and her coplayers belong to T. In the case in which T is the player set, we may put this by saying that a player may or may not think in 'we' terms about how to play the game. The more inclined a player is to 'we' thinking, and the more inclined she takes coplayers to be, the more will fellow-member reasoning be favoured" (ibidem).

To bypass the problem which Bacharach identifies, the purpose of the 1997 model is to justify via a "variable frame model" when players cooperate and when they fail to cooperate, i.e. when they team reason or remain individualistic reasoners. In this model, it is the structure of games which enhances the frames associated with the two modes of reasoning. Hence the game may enhance pre-existing collective cognitive representations. Again, as the individuals who survive natural selection are cooperative individuals, a stock of cooperative repertoire and cooperative actions are transmitted to each individual's offspring. We thus all inherit a repertoire of cooperative concepts and cooperative behaviors (Lempert, 2018, p. 435). If players conceive themselves as members of a team they are, in some ways, naturally led to cooperate. Nevertheless, another problem appears. Games can induce a tradeoff between cooperation and defection. In fact, even if coordination games show scope for cooperation (or coordination) in order to reach the Pareto optimal outcome, the riskiness of cooperation may induce

individualistic behavior.[78] In this perspective, the "agency transformation" allowing players to team reason and then to cooperate, may not be enhanced. Moreover, this tradeoff is internal for the players in S frames, when the evaluation of cooperation in a I or a We perspective do not lead to the same conclusions.

Subsequently, Bacharach refines the conditions enhancing we frames in the later models (1999, 2006). Referring to social psychology, Bacharach defines the conditions and the mechanism which trigger the we frames. In particular, several experiments in the theories of self-identity and self-categorization (in social psychology) point out that "common fate" or "common interest" and "interdependence" (among others), induce group-identification(Bacharach in Gold and Sugden, 2006, p. 82). More precisely, he argues that

> ""interdependence' and 'common interests', are of particular interest for our inquiry because they connect group identity to characteristics of decisions problems". They "may help explain why a group of people locked into a game might tend to group identify. They have a common interest in Pareto-optimal outcome and, usually, they had no hand to in choosing the payoff matrix. I shall call the tendency of this combination of factors to produce group identification in the player set the common problem mechanism" (ibid, p. 133).

Formally, the game theoretical characteristics in sight in this "common problem mechanism" are the following:

6. The Common interest: "P1 [Player 1] and P2 [Player 2] have a common interest in s* over s [s* and s being two different outcomes] if they mutually know that both prefer s* to s" (ibid, p. 83). This concerns the payoff dominant outcome.[79]
7. The Copower: "among all the feasible outcomes there is one – call it s* – both P1 and P2 rank highest, and that it can only be brought about by each acting in a particular way. This is the pattern in a 'common-interest' game" (ibid, p. 84),
8. The interdependence: "the interdependent hypothesis concerns situations in which s* is not assured by individualistic decision-making, and P1 and P2 perceive that it is not" (ibid, p. 85).

As for Schelling, interdependence brings situations in which people understand that they have to come together to a solution of the decision problem they face. When the consistency of choices is not assumed with the standard and mathematical conception of the equilibrium as an hypothesis of the model, the acknowledgement of this interdependence is of prime importance for the players to recognize that strategic reasoning must induce a process of convergence among perceptions, intentions and beliefs. A state of consistence must be reached and it requires the

78 In the same perspective, see Tan and Zizzo (Zizzo and Tan [2007, 2011], Tan and Zizzo [2008],) on the link between games harmony of interest and the risk of defection and Smerilli (2012) on the probability of vacillation according the games.

79 The importance of common interest for collective raises numerous debates in philosophy with respect to the nature of a collective (see Gold and Sugden 2007; Gold, 2018). Common interest is the pillar of Margaret Gilbert's account of plural subject in philosophy. For Gilbert (1989), when people have a common interest they cease to act individually and become a plural subject acting of concert to perform an action.

acknowledgement of the strong interdependence of their choices and behaviors in which players as accurately emphasized by Schelling first, must adapt each other. They must accommodate each other. They acknowledge that they must collectively promote the best possible outcome. The emphasis that Bacharach puts on this interdependence shows how he sees the scope for coordination as a collective enterprise in strategic decision problems, in the same spirit as Schelling's. Accordingly, group identification is produced if "[f]or some S, S*, the players have common interest in, and copower for, S* over S, and sol(G) [i.e. the set of solutions in the game G] contains outcomes in S"(ibidem). However, in addition, players must have common knowledge of these facts, i.e. of the "common interest", "copower", and "interdependence" characteristics (ibid, p. 83). Although, once more, players may not perceive these characteristics, they may not perceive this "common problem mechanism". Thus, "we frames" are not necessarily salient and thus not activated.

The difficulties in the different and successive justifications of cooperation must be considered alongside with the players' competing selves. In 1995, Bacharach hardly reconciles the fact that "types T", which are naturally cooperative reasoners, may act as best-reply reasoners. As stressed above, Bacharach needs make an *ad hoc* assumption contradictory to his conceptual choices. In the subsequent models onward, as soon as TR relies on framing, an incompatibility between the players' two selves (i.e. the individualistic or the collective self) is exhibited. Indeed, Bacharach highlights and advocates that frames are "non-integrable". To justify that, Bacharach refers to the vase-faces or the duck-rabbit illusions to show that two perceptions, for instance in the former example, either the vase or the two faces, are incompatible. We cannot handle the two perceptions at the same time.

If the I and We frames are "non-integrable" it means that there is no link between players' different selves, either they are a member of a group, or they are an individual. In the presentation of the 1997 model Bacharach (1997, p. 2) emphasizes: "[t]he 'we' and 'I' ways of thinking about a problem are rivals". The concept of "S frames" is an attempt to reconcile these two competing selves. Nevertheless, players in S frames cooperate in their own interest and we see an incompatibility between cooperation as a profile directed reasoning induced by a "collective profile function" and evaluating the risk to cooperate from an individual perspective. Players in "S frames" must have an individual utility function. They do not follow the team profile of action. They must evaluate for themselves the action to take, so they refer to a proper subjective utility function.[80] Hence even if they may choose to cooperate – as they can in "we frames" – they remain individualistic reasoners. The tradeoff between team reasoning and traditional reasoning when a player is in a "S frame" is resolved by a principle of "reason dominance".[81] Bacharach explicitly recognizes that this concept is a limitation for his model: "[i]n the absence of a natural way to resolve conflicts, we are limited once again, in making predictions

80 For Bacharach, players is "S frames" have basically altruist preferences (ibid, p. 25), i.e. a specific from of individual preferences.

81 Bacharach (1997, p. 17) defines the concept of reason dominance as follow: "z *strongly reason-dominates* z' if it is better than z' in term of both group evaluation and personal evaluation in S. If z strongly reason-dominates z' [z and z' are two possible actions, e.g. cooperate or defect respectively], then in cases in which personal and group evaluation are comparable, z defeats z' on balance *whatever* the relative weights. z *weakly reason-dominates* z' if it is better than z' in terms of some evaluation, and not worse than z' in either".

from our theory, to the use of dominance criteria" (Bacharach, 1997, p. 17). So again, we see the incompatibility between collective and individual selves. Bacharach is aware of this issue, as he recognizes that he fails to demonstrate that cooperation, i.e. we reasoning, is rational.

This methodological ambiguity disappears in both the 1999 paper and the 2006 book. In addition, the 1999 model offers an original account of the players: they can have multiple selves with respect to the fact that they can belong to several teams (Davis, 2011, p. 118). Nevertheless, once again, only one self can is activated at a given point in time, since players can be active in only one of the teams to which they belong. When a player has to make a decision and when this decision is acting for one of the team to which she belongs, she fully endorses the collective self, inherited from the team. Thus, again, there is the same idea of "non-integrability", this time, between multiple selves and no longer between dual selves. Players do not have the reflexive ability to choose between competing aspects of their identity, contrary to Schelling (see Orléan, 2004). It means that to cooperate individuals must fully follow one of their collective selves; being guided by an individual and a personal self would bring defection or failure of coordination.

Another question linked to the competing selves, is how far can Bacharach go in the tradeoff between an individual's interest and the collective interest? In other words, how far can individuals go in their renouncement of their individual identity when they belong to a team and act for that team?[82] Except in his 1997 paper, the question of the competition between individual and collective interests is never explicitly posed. Bacharach argues that individual and collective interests may not be concomitant. This postulation is clear when he asserts that "the goal of the group in this framing of self need not agree with the person's goals under her individual framing of herself, but perceived agreement of individual goals among a set of individuals favours framing as members of a group with this common goal" (Bacharach, in Gold and Sugden, 2006, p. 90). The competition between "I frames" and "we frames", or between individualistic reasoning and team reasoning, is obviously the proof that if individual and collective interests compete, team reasoning may not be activated. In addition, when players belong to and act for a team, Bacharach implicitly presupposes a kind of convergence within the members' representations and in turn among their actions. In fact, Bacharach is much more insistent in this book about the "common ranking" condition. Team reasoning is induced by condition in which this common ranking condition prevails: "the basic class of situations in which the possibility of team reasoning arises is that there is a set of agents who have alternative action options, and a common ranking by the agents of the profiles made up by these alternatives. I call situations of this kind coordination contexts" (ibid, p. 122). Coordination contexts are therefore situations in which players are interdependent and have common interest and co-power on the issue. Coordination contexts are situations in which players have common understanding and make common inductive inferences (as in community-based reasoning). Subsequently, on the one hand the players' individual interest and the collective interest must be concordant, and on the other hand players' interest must be convergent to activate TR. Besides, Bacharach's models rely on a

82 This questioning is at the center of Gilbert's account of collective entity (e.g. 2003). For Gilbert, when individuals accept to enter in a collective action, they become entirely committed to the collective purpose and then become collective entity. They cease to be fully-fledged individuals.

principle of aggregation. The collective utility function is simply the addition of the individuals' utilities and Bacharach admits that "there are well known difficulties in aggregating values in general" (Bacharach, 1997, p. 17). Bacharach is aware that this mechanism is quite controversial to grasp the problem of what is a collective. In fact, it implies that players' individual interests are perfectly concordant to the collective ones.

In summary, in the TR theory cooperation seems to rely on a natural propensity possessed by the players, inherited by natural selection, who are likely to cooperate with individuals who belong to the same group as them. However, as embedded in the realm of strategic interactions and in a game theoretic context, the integration of framing in the TR makes cooperation much more context-dependent. Some game situations tend to enhance TR while some others, on the contrary prevent the TR. Nevertheless, it is not through interactions that, at a certain point in time, players decide to cooperate or to act collectively as a team. They do not agree on a mode of interaction. Again as embedded on a static game theoretic framework, Bacharach struggles to integrate players' capacity of reflexivity which would explain how players conciliate their different selves. This statement contradicts Bacharach's attempt to endogenize the probability that a player, in a given context, perceives herself as a member of a team and no longer as an individual. Moreover, the endogenization problem exemplifies the tradeoff between the players' individual and collective selves. Besides, coordination contexts which trigger TR seem indeed quite restrictive: players interests must be perfectly aligned.

4.3. Salience and the "endogenization problem"

As stated above, the incompatibility between the players' different "agency" must be considered alongside with the salience of the frames and the endogenization problem. If collective identity is more salient than individual identity, team reasoning and consequently cooperation takes hold over individualistic reasoning and a potential defection. This problem of endogenization arises with the inclusion of frames in team reasoning. As soon as frames matter, Bacharach's purpose is indeed to endogenize the probability that in a given context the "we frames" get the upper hand on the "I frames". This endogenization problem explains the progressive refinements of the conditions that trigger "we frames", and the issuing concept of "common problem mechanism". As Bacharach claims (in Gold, Sugden and 2006, p. 81),

> "I am interested in the effects of spontaneous framing in games. Group identification is a framing phenomenon. So I am interested in the role of spontaneous group identification in decision-making".

Besides, he affirms:

> "which of my collective personae is activated depends on the current 'accessibility' of the categories to which I belong … In human interactions, the accessibility of categories is a special case of the notion of availability of frames at the heart of the variable frame theory

> of games. The process in which categories are activated is context-dependent and jumpy" (ibid, p. 74)

Consequently, the problem of endogenization and salience are tightly linked. To avoid this problem of endogenization, Bacharach needs to explain how and when collective identities may be more salient, i.e. in which kinds of games or interactional contexts.

In 1997, Bacharach proposes two models: (i) one without players in "S frames" and (ii) one with players "S frames". In the former if players are in "we frames", it is a priori postulated, and not triggered by the game. Besides, players in "we frames" are "blinkered" by their frames. They cannot think of the others' unreliability – that is induced by the relative risk of cooperation. This relative risk does not compete with the enhancement of "we frames". Players are already cognitively predisposed to cooperate. Accordingly, framing is not a matter of games and salience. In the latter, the respective probabilities of being in I, W or S frames are given by an objective probability distribution over the set of players, whereas Bacharach (1997, p. 6) argues that "collective identity effects are … endogenous to [games]" which means that "the personal and social levels of categorization have varying different salience in different interactions" (ibid, p. 23). These probabilities should therefore be determined by the games' characteristics, and particularly by what Bacharach calls "the 'gain from cooperation'" (ibid, p. 24). The probability that "we frames" come to the players' mind increases with this "gain from cooperation", and relates to what Bacharach calls "the harmony of interest in game" (ibidem). Such difficulties enhanced by the definition of players' beliefs by prior probabilities will be dismissed in the theory of games proposed in the chapter 5. We will rely on a specific cognitive mechanism called simulation to define players' beliefs with respect to the other players' choice without resorting to these priors. If games display these two dimensions, i.e. "gain for cooperation" and "harmony of interest", for Bacharach, they are games of "common interest" (ibidem). The same arguments about the games' characteristics activating we frames are developed in 1999, and once again Bacharach points out the endogenization problem. He identifies some paths to resolve the problem by arguing that

> "it is plausible that ω [the probability that "we frames" instantiate] may be an increasing function of certain quantitative features of the payoff structure such as "scope for co-operation" and "harmony of interest"... To endogenize ω … one must show that the payoffs and other constitutive features of the basic game make collective identity salient or otherwise tend to induce team-thinking" (Bacharach, 1999, p. 144).

So here again Bacharach is not able to bypass the limits he encounters in the previous papers, even by changing some methodological and formal aspects of his model. Even with a precise identification of the required game characteristics, in his posthumous book, the problem of endogenization still prevails.[83] Indeed, due to the incompleteness of the players' frames, players

83 Tan and Zizzo (2008) build their model around the concept of "harmony of interests" and "disharmony of interests". In a comparative way, they dissociate games by showing that some games like the Hi-Lo games show a "harmony of interest" while others, like the prisoners' dilemma show at the opposite a "disharmony of interests". However, contrary to Bacharach, they succeed in endogenizing the probability that a given player in a given game fails to team reason.

may still not perceive the salience of the "common problem mechanism". Group-identification as a result may not be prompted (in Gold and Sugden, 2006, p. 87). Bacharach explains that the characteristics underlying the "common problem mechanism" are more noticeable in some games than in others, as is group-identification (ibidem). Bacharach gives the following example:

> "In prisoners dilemma, players might see only, or most powerfully, the feature of common interest and reciprocal dependence which lie in the payoffs of the main diagonal. But they might see the problem in other ways. For example, someone might be struck by the thought that her coplayer is in a position to double cross her by playing D [i.e. Defection] in the expectation that she will play C [i.e. Cooperation]. This perception might inhibit group identification" (ibid, p. 86).

In fact, Bacharach argues:

> "There is common interest in Stag Hunts, Battles of the Sexes, bargaining games and even Prisoner's Dilemmas. Indeed, in any interaction modelable as a 'mixed motive' game there is an element of common interest. Moreover, in most of the landmark cases, including the prisoner's dilemma, the common interest is of the kind that creates strong interdependence, and so … creates pressure for group identification. And given group identification, we should expect team reasoning. But for the theory of endogenous team reasoning there are two differences between the Hi-Lo case and these other cases of strong interdependence. First outside Hi-Los there are counterpressures towards individual self-identification and so I-framing of the problem" (ibid, p. 144).

The second difference is about the "unanimity condition". Saying that there is a unanimity condition implies that players "have the same value on every profile in the profile set" (ibid, p. 88).

We would like to point out nowthe three types of contexts identified in the 2006 book -- (i) the simple coordination context, (ii) the restricted coordination context and (iii) the unreliable coordination context -- could induce an ambiguity on what can affect the salience of the "we frames". If we interpret the three different coordination contexts as different informational contexts, it could challenge the interpretation of salience with respect to the objective structure of game matrices – i.e. the characteristic described above and involved in the 'common problem mechanism'. Consequently, two facts could potentially affect the enhancement of the players' collective identity: (i) the types of games, and (ii) the risk of cooperation according to the epistemic structure of games. Subsequently, it seems that the context of coordination is also a matter of information and knowledge and not only of unreliability or riskiness *per se*. The different salience of frames would accordingly found two explanations: (i) the type of games, or eventually (ii) the lack of information regarding the other, and in that perspective, for instance, the size of the set of players. Nevertheless, as the distribution of cooperative and non-cooperative players is already postulated before the game, as symbolized by an objective probability distribution, it remains problematic to justify that it is the uncertainty regarding the available information, which affects salience.

Finally, we would like to highlight that the problem of endogenization, contrary to aggregation, is not specific to game theory but to Bacharach himself and his methodological choices. If the

aggregation problem must belong to standard game theory and neoclassical economics (in their approach of collective entities and collective states), the endogenization problem at play in Bacharach's theory of TR is totally specific to the integration of framing in a game theoretical apparatus. This issue can be explained by the incompatibility of Bacharach's attempt with standard non-cooperative game theory. In fact, as the structure of the game must be known prior to the game in order for players to have beliefs about the others' type and corresponding behavior, there is an impossibility to endogenize the salience of frames. We therefore witness once more the difficulties that Bacharach faces to conciliate an open-system vision of games with a game theoretic framework which needs to operate according to 'fixed rules' (Giocoli, 2003), i.e. the mathematical rules, and necessitates closed-systems in order for Bayes rationality to be operative (Binmore, 2009, pp.134-135)

> '[t]he models constructed by game theorists are small worlds almost by definition. So we can use Bayesian decision theory without fear of being haunted by Savage's ghost telling us that it is ridiculous to use his theory in a large world. Bayesian decision theory only ceases to be appropriate when attempts are made to include the minds of the players in the model to be considered' (Binmore, 2009, pp. 134-135)

Bacharach opens the door maybe to a challenging research agenda in the integration of players' mental states, of players' reasoning process in games, which bring strong interdisciplinarity and makes games as open-systems. He anticipates the integration of philosophy on the nature of collective entities, on the role of institutions in players reasoning and more generally on the way players appraise a game and subjectively define the game they are to play. He also anticipates the integration of cognitive sciences on the issue of players' mental states. Finally, he uses experimental methods in game theory. Bacharach thus opens the doors to an enriched research agenda pursuing Schelling's program but also adding some new questions about the compatibility of the mathematical formalism of game theory and a conception of games as open systems.

Bacharach is often mentioned in the reflexive and critical literature on game theory. His assessment of the payoff dominance principle (which is in first instance a mere theoretical and methodological problem) entails a deeper reflection on coordination and on the integration of players' mental states and reasoning process as a means to explain how and why players coordinate, i.e. how and why a solution may exist. In his thinking, the payoff dominance principle is simply the mathematical translation by the theorist of the option that is perceived as a salient option, i.e. as a focal point for every player in the game. The TR and the VFT offers complementary justifications relying on a common underlying cognitive mechanism: framing; that can however impinge differently on players' reasoning.

We suggest that, like in VFT, Bacharach faces numerous complications because of the integration of the players' perceptions of the game and of themselves in the standard game theoretic formalism. Despite the difficulties that Bacharach faces in his TR theory, we would like to highlight in the next section that it does not mean that his program fails, quite the contrary. We emphasize how, through both VFT and TR, he changes the standard conception of players and their rationality and how he substantially enhances standard non-cooperative game theory.

5. A rational reconstruction of VFT and TR's enrichments of standard non-cooperative game theory: a new conception of players and their rationality.

In this section we highlight how Bacharach with the VFT and TR theory progressively enriches the standard conception of players and their rationality. In this manner VFT and TR must be considered as a continuation participating in a common purpose. Bacharach has always purported to justify why it is rational for players to coordinate and cooperate. He attempts to offer a game theoretical explanation of empirical regularities demonstrating that players are more inclined to coordinate and cooperate than to defect. The VFT and the TR are a means to justify coordination on salient options; both purport to offer a more realistic account of interdependent decisions in strategic contexts. We intend to show how this purpose, translated in his work, entails a different conception of players compared to standard game theory and then a different account of strategic rationality in games.

5.1. Which conception of players?

Bacharach regularly refers to 'normal' agents (1991, 1993, 1997). We are going to emphasize that this conception of players is mostly based on three dimensions. Even with various models of VFT and TR, different conceptual and methodological choices, and different formalizations, we always find three concerns: (i) he tries to avoid to endow players with unlimited reasoning depth, (ii) he integrates psychological dimensions in the way players frame their decision problem, and (iii) he tries to justify how players' identity matters in decision problems, even for strategic decisions.

First of all, Bacharach's conception of 'normal' players entails limited cognitive abilities. This is only one aspect of the psychological dimension that Bacharach integrates. This limit relies on two distinct but related aspects: the depth of players' beliefs and the epistemic requirements of rational choices and their conceptual limitations.

As Bacharach (1992, p. 247) already claims, "there are limits about the depth of the beliefs people can attain." Consequently game theorists should not explain games solution using common aligned beliefs and common knowledge. For instance, to the question "[w]hat difference does common knowledge make?" he answers: "in many cases, none; ck [common knowledge] itself is not crucial – all that matters is whether there is iterated knowledge of sufficient degree" (Bacharach, 1992, p. 10). This echoes Schelling's work in which mutual knowledge is sufficient for coordination. Accordingly, in his skeletal development of both the VFT and TR (in 1991 and 1995 respectively), one of his major concerns is the strength of players' beliefs. They are both grounded on a "sufficiently strong degree of belief" to avoid the unrealistic common knowledge requirement (Bacharach, 1991, p. 30; 1995, p. 9), and again it is due to the players' limited cognitive abilities. In summary, Bacharach generally avoids adhering to the common knowledge

hypothesis throughout his models of VFT and TR. Even in 2006 (in Gold and Sugden), when Bacharach states his "common problem mechanism" on common knowledge, he tries to assess team reasoning in contexts characterized by a lack of information (we must refer to the three types of coordination identified). For instance, common interest is a psychological phenomenon that may be sufficient to bring coordination in the TR (provided that the context may not be of unreliability) without the need for common knowledge. Interdependence can break the standard and closed system view of standard or Bayesian game theory in which common knowledge is sometimes necessary.

Another aspect of players' limited cognitive capacities is explained by their psychology, and more precisely by their perceptions. In this perspective Bacharach relies on many forms of psychologies and frameworks in psychology (from cognitive to social psychology, from theoretical to experimental psychology). In his early works Bacharach does not mention psychological assumptions about players' reasoning so he only refers to knowledge, but as soon as he begins to work on framing, players' cognitive limits are first and foremost due to the incompleteness of their perceptions. Indeed, Bacharach argues, "our conceptual endowments are limited, that we as normals are not 'conceptually omniscient'" (1991, p. 29). It means and implies that players' frames are incomplete. For instance, Bacharach asserts that there are universal frames encompassing all the possible representations of a given problem but players, as 'normal' individuals, cannot have access to these universal frames. The methodological consequence is that the universe, i.e. the states of the world that the players deem possible, cannot be complete, so that Bayesian rationality cannot be operative. Players cannot perceive all of the relevant ways to represent and describe a decision problem (1991, 1993, 2001, 2006). Thus, as Bacharach refers to the incompleteness of framing – or in other words to incompleteness of representations, by direct link it means that the players' information is incomplete. Besides, since framing is anterior to reasoning, there are necessarily limits about the depth of players' reasoning. Players' awareness of all their own possible available strategies and, by implication, of others' strategies are necessarily incomplete. Hence players' beliefs are necessarily limited and incomplete.

I would however remind that Bacharach's integration of framing in his theories portrays players quite differently from framing in the literature on this topic, and especially from Kahneman and Tversky (1979, 1986) to whom Bacharach refers. At the time of Bacharach's contribution to Decision Theory and Game Theory through his Variable Frame Theory (VFT) (1990, 1991, 1993, 1997, 1999, 2000, 2001, 2006), the work of Kahneman and Tversky (1979, 1986), and the 'framing effect literature', started to influence the work of economists in decision and game theories. However, framing does not involve irrational players for Bacharach as is generally assumed. Framing does not imply biases from perfect rationality and optimal choice. Contrary to Kahneman and Tversky, perfect rationality as defined by the RCT or Bayesian rationality are not a benchmark for Bacharach for identifying how framing precludes from optimal choices. It is quite the contrary; frames allow Bacharach to nicely shape what could be understood as an idiosyncratic individual rationality. By arguing that deliberation, i.e. the phase during which rationality is operative is distinct from the unconscious phase, he shows that framing does not entail irrationality.

In most of his work knowledge is an "occurrent knowledge", i.e. conditioned by players' frames. Because frames are necessarily incomplete, so is players' knowledge. Bacharach (e.g. in 1991,

1993) regularly insists on the fact that, due to frames, the incompleteness of players' information (or knowledge) is deeper than in Harsanyi's (1967-8) work. For instance Bacharach (1993, p. 257) claims:

> "[q]uite generally, a given question may or may not occur to an agent facing a decision problem; she may or may not think of it. This phenomenon give rise to variations in agent's beliefs systems of a more radical sort than those to be found in the familiar Bayesian model".

An analytical consequence is in particular that the axiom of logical omniscience disappears. This axiom states that players know that they know something, and know that they don't know something. On the contrary every model that is grounded on framing imply in Bacharach's account that players cannot know that they don't know (e.e. see Bacharach, 2001b). Those two postulates are however essential in incomplete information game theory in order for players to assess their subjective beliefs when facing new information.

In addition, Bacharach presupposes that culture induces biases of representations. He refers to an empirical example: the Eskimos have numerous ways to describe the snow and the color white which we are not aware of as Europeans – or more generally as non-Eskimoss (Bacharach, 1993, p. 259). In this perspective, individuals and subsequently players, are in a way culturally determined and accordingly cognitively determined. This cultural determinism may induce again limited cognitive capacities.

Bacharach differentiates two forms of knowledge: (i) occurent and (ii) non-occurent (see Bacharach 1991, p. 15). The former entails that "a necessary condition for someone to know that p is that the question of p should come to her mind or occur to her; that she should think of p". This knowledge is deliberative but nevertheless constrained by framing. Individual frames determine individual inference and epistemic conditions (see Scazzieri, 2011). The later is tacit, i.e. understood in term of rule following behaviors, "but nevertheless action-guiding". Tacit knowledge is expressed in terms of unconscious rule following. In other words, tacit knowledge corresponds to stored and generalized knowledge enabling individuals, by congruence between similar contexts to follow patterns of behaviors (Scazzieri, 2008, 2011). This paves the way to the acknowledgment that community-based reasoning matter in decision and game theories. We can assume that in Bacharach's framework, tacit knowledge stems from individuals' "everyday experience" which then, by generalization, forms an "everyday theory". As we have seen above, Bacharach views agents as non-omniscient and cognitively bounded. In an uncertain world, unconscious rule following behavior therefore plays an essential role. Understanding how occurent knowledge progressively generalizes to a non-occurent knowledge, i.e. how short-term and long-term memory structures individual cognitive frames, requires resorting to psychological investigations. We will detail the psychological theories to which Bacharach refers to develop his account of framing to adapt such account to strategic reasoning in the next subsection.

Furthermore, frames give access to a part of players' identity according to Scazzieri (2008). For instance, he emphasizes that in Bacharach's account:

> "framing would primarily be associated with the cognitive and linguistic ability to grasp specific problem situations through the activation of a particular set of 'naturally

connected' features. It is reasonable to assume that frames would often be associated with relations among attributes, and that such relations will often be of the causal type. In short, framing would generally rely upon pre-existing cognitive structures (the different subsets of naturally connected features), but only specific (contingent) circumstances could turn a virtual frame into an effective one" (Scazzieri, 2008, pp. 196-97).

Accordingly frames rely on both the players' culture and personal experience. This implies that games can no longer be closed-systems in Bacharach's view. As for Schelling, the social world, cultural devices, and history, impinge on players' decision making. In Bacharach's theories, players are not discarded from what comprises their specificity, their differences, contrary to standard non-cooperative game theory in which it is assumed that rationality sweep away any difference (being of preferences or information (cf. the principle of symmetry in Nash (1950) or in Harsanyi (1956); see Innocenti, 2005)

The TR, by relying on framing too, sheds lights on another aspect of players' identity. When players TR, i.e. when they are in we frames, they express their "group identity". From the viewpoint of a player, seeing herself as a member of a group is a process of "affiliation" which is "a psychological process in which a person who does think about a certain group, defined by some shared property, comes to thinks about it as 'us"' (Bacharach, 1997, p. 2). In this perspective the individual perceive her own properties or characteristics as those of the group, as shared with the other members of the group (Bacharach in Gold and Sugden, 2006, p. 73). When a player "identifies herself with the group; her self-conception is as a component part of the group. This is reflected in her language. She thinks and speaks of the group not as "them and me" but as "us"" (Bacharach, 1999, p. 134). Being a member of a group and acting for this group implies in her mind a 'cognitive extension' of her own interest, meaning that "the group's goals define her basic preferences" (Bacharach, 1997, p. 16). Accordingly we frames and TR express players' "collective self-conception".

Bacharach emphasizes that a "personhood is to some extent constituted by group membership … personhood is the resultant, to the extent that it is so constituted, of a set of group identities; more exactly, the person is defined by the intersection of her group identities" (in Gold and Sugden, 2006, pp. 88-9). As a consequence "team reasoning is a basic decision-making proclivity and mankind" (ibid, p. 121). Moreover, Bacharach grounds his conception of the connection between collective and individual agency on psychologists within self-categorization theory and more specifically on Tajfel and Turner (1985) who assert: "the sense of group identity precedes, developmentally, the sense of personal identity".

In addition, the culture in which individuals are embedded, the collective to which they belong endows them with 'social' and tacit knowledge. Bacharach (in Bacharach and Hurley, 1991, p. 3) believes in the existence of a "cultural common sense", i.e. "the fact that every real [individual] has the general knowledge her cultures gives her, such as knowledge of which arrangements are salient or traditional in that culture". This "cultural common sense" may allow individuals to coordinate or at least can orient them toward an effective decision in an uncertain world populated by cognitively bounded individuals and in which the cognitive load of computation is too heavy (see also Aoki, 2001).

In the previous section we underlined that Bacharach's models of TR depict players as dual and multiple selves. We insisted on the recurrent tradeoff between individualistic reasoning and team reasoning, and the incompatibility between them. The difficulty we emphasized was primarily theoretic. This was the price to pay to propose an enrichment of standard game theory which presuppose accepting some of its rules. We do not mean that Bacharach's conception of players, as individuals, may not be consistent, or that Bacharach fails to demonstrate what a player is. I merely suggested that such a conception of a player raises theoretic issues. We indeed agree with Davis' (2011, p. 118) point of view when he asserts that in Bacharach's TR "having multiple selves underlies being a single individual". This assumption is even more reinforced by the previous quotations. According to him, "Bacharach sees individuals' multiple selves as their supra-personal social identities." (ibid, p. 119).[84] In this perspective Bacharach embeds his account of game thery in sociology, in history and biology. Recall first that what constitutes the history of a particular community explains its institutions and the way its people interact. Community members thus inherit this institutional knowledge and possess a repertoire of coordinating and cooperative behaviors, which constitute their collective identities. Recall secondly that evolution in the long run favoring collective which cooperate and coordinate has for consequence that we all inherit cooperative and coordinative repertoires of actions. Again all of this pleads for an open system account of games and of game theory.

To sum up, Bacharach portrays his players as heterogeneous individuals, with realistic cognitive capacities, who are socially skilled, constituted by multiple collective identities, and who are not deprived of their personal history and experience of sociality.

5.2. On what 'psychologies' Bacharach draws to portray the players in his games?

Bacharach draws on different subfields of psychology in his theories. For instance, in view of the two forms of knowledge Bacharach identifies -- the occurent and non-occurent -- Bacharach draws upon both social and cognitive psychology, and on cognitive science more generally, to understand how this non-occurent knowledge is built and stored, and how it can affect individual decision-making.

For Bacharach, "the study of rationality has much to gain by triangulation from different disciplines" (in Bacharach and Hurley, 1991, p. 4). This includes, indeed, cognitive psychology and social psychology. Understanding how players' representations matter in their beliefs and mode of reasoning goes beyond the scope of economics. Recall that for Bacharach the role of

84 Bacharach's challenge is accordingly for Davis "the fusions of agency" (2011, p. 119). Bacharach's modus operandi is the following for Davis, he "employ[s] one kind of relational conception of the individual – one in which people are single individuals in virtue of how their interaction with others makes them more than collection of multiple selves" (ibidem). This is true if we consider the fact that only one mode of reasoning can be enhanced at a time.

individuals' beliefs is central to the analysis of interactions. In this perspective he uses and crosses various contributions from cognitive psychology and social psychology. Alongside the role of framing, a large part of psychology that matter for Bacharach is linked to the role of sociality, and of social and strategic interactions, in human cognition. Besides, because frames are structured, different levels of this structure require different mechanisms of activation (Scazzieri, 2008). Bacharach therefore draws on different frameworks in psychology to identify what determine the structure of frames and the different levels of activation that prevail in individuals' recognitional capacities.

A first element of framing that matters for Bacharach is 'entification' since concepts come in bundles and are classified in families (Bacharach, 1991, 1993, 2001, 2006). In this perspective, he refers to Gestalt psychology (e.g. Wertheimer, 1923; Campbell, 1958) which focuses on visual perceptions. In this approach, 'entification' is associated with the following characteristics: "'contiguity', 'common fate' (moving in parallel over time), 'good figure' (forming a recognizable pattern) and 'similarity'…'closeness and impermeability" (Bacharach, 2006, pp. 70-71). One of the previous characteristics, "similarity", leads to a second trend of research that Bacharach uses: Post Gestalt psychology (e.g. Campbell, 1958; Tajfel, 1969; Tversky, 1977; Rosch, 1978). This approach defines similarity as the "criterion of grouphood" i.e. "the meta-contrast principle [which] explains categorization, that is, the cognitive activity of dividing a domain of items" (ibid, p. 71).

On the one hand, by referring to Gestalt psychology Bacharach offers a first element in the structuration of frames. This first element explains the horizontal classification of 'things' through families of concepts. On the other hand, similarity is a determinant characteristic to form generalizations in context of uncertainty and to draw generalizations from recurrent practical situations and experiences. Accordingly, similarity between situations, commodities, etc. allows individuals to form their everyday theory and in turn to appreciate their response despite the uncertainty surrounding their decisions (Scazzieri, 2011). This criterion can therefore refer to a second element in Bacharach's account of the structure of frames: the vertical classification. It explains why some concepts, some frames, acquire progressively higher level of generalization; it explains why some concepts become deeply anchored in individuals' cognitive structure. Specific situations involve the instantiation of some concepts or frames. The more those specific – or quite similar – situations occur the more the preexisting associated network of concepts will instantiate, and the more this network will be therefore deeply anchored. And in turn, the more the behavioral response will be unconscious. This phenomenon must be considered alongside the role of direct social perception that will be exposed in the next chapter. Direct social perception entails that in such circumstances, i.e. when behavior becomes unconscious, recognizing in interaction one of such type of behaviors induces the understanding of the other automatically. It does not necessitate thinking of mental states such as the preferences, the intentions or the beliefs that lead to the behavior in question. This has important consequences for strategic interactions which in many cases resort to situation in which such kind of behavior prevails. This will be discussed and taken into account in the model proposed in the chapter 5.

In this perspective Bacharach's account of framing is close to the post – behavioral – Skinnerian (1985) relational frame theory. This psychological theory is centered on the relationship between networks of stimuli – including visual perceptions, smells, noises emotions etc. – their meaning

and associated semantics for people, and their behavioral responses. This theory postulates that the different networks of stimuli will be stored in long term memory and any context involving one of the stimuli will provoke the instantiation of the whole network as we previously showed for Bacharach's VFT. Indeed, as already pointed out, and emphasized by Scazzieri (2008, pp. 196-97), in the VFT, framing is associated "the activation of a particular set of 'naturally connected' features … framing … generally rely upon pre-existing cognitive structures".

For scholars in the post Skinnerian account of relational frame theory "the act of relational framing is thought of as a process, an ongoing way of responding to stimuli as they are presented. People frame events relationally in the moment as an active process that is a function of their extensive learning history and stimulation in the present environment." (Blackledge, 2003, p. 429)

Besides, the recognitional capacities mentioned by Bacharach (1986) are conceived as in the relational frame theory, i.e.

> "With a vast amount of training, using multiple relations across many stimuli, words come to share the functions of a wide variety of experiences and events. At first, this occurs through direct training, and along formal stimulus dimensions. After repeated experiences of doing so across multiple exemplars, we learn to bring relational responding to bear on non-formal, or arbitrary, relations between stimuli. Once we do so, our verbally constructed worlds become increasingly complex as we derive more and more relations between virtually every stimulus we discriminate." (Blackledge, 2003, p. 427)

Drawing on Post Gestalt Psychology, Bacharach aims also to show why individuals' identities matter in their decision-making. He states that "psychological, cultural and social" similarities can provide the basis for an individual to entify a group of individuals and in turn activate her sense of group identity (ibidem). As we shall see in the next section, this feature of human cognition plays an important role since it allows individuals to interact, communicate, and coordinate. However this statement does not mean that individuals' heterogeneity vanishes like Bacharach's subjectivism.

> "Personhood is the resultant, to the extent that it is so constituted, of a set of group identities; more exactly, the person is defined by the intersection of her group identities. But it is only to some extent, since there are plenty of person-defining features which do not correspond to group memberships." (ibid, pp. 88-89)

Bacharach refers to the theory of self-categorization in social psychology in order to understand how each individual constitutes her own identity through interactions between her multiple group identities.[85] Self-categorization theory postulates that self-identification is a matter of framing. This means that salience in specific contexts explains which of the self-identities are activated (ibidem). Bacharach relies on Bruner (1957) and Gurin and Markus (1988) to postulate that

85 There exists a strong link between this theory in social psychology and what will be identified in the chapter 4 as the direct social perception literature and more generally the literature of mindreading and mindshaping.

> "Which of my collective personae is activated depends on the current accessibility' of the categories to which I belong the relative accessibility of a category depends upon many things, which include the perceiver's current expectations, tasks and purposes. In human interactions, the accessibility of categories is a special case of the notion of availability of frames at the heart of the variable frame theory of games." (ibid, p. 89)

Individuals' collective identities shape their decision-making since for Bacharach they affect individuals' goals (ibid, p. 75). Therefore, Bacharach tries to define the conditions that tend to enhance individuals' 'group identification', which "in turn produces certain judgments, attitudes and behaviour" (ibid, p. 76). The circumstances identified by social psychologists and the consequences of group-identification on the members' mental states are numerous, as are the references in Bacharach (2006). Among other things, membership induces the "internalization of group norms" (ibid, p. 80). It thus ultimately produces community-based reasoning. In people's mind, these norms determine reliable principles of coordination. Individuals are partly determined by their social environments for Bacharach in the sense that their collective identities may influence, in some circumstances, their mental states. However for Bacharach, this statement is compatible with subjectivism. Cognitive frames are specific to each individual. The structure of frames – or of individual conceptual repertoire – primarily relies on each personal experience. The way individuals filter salient characteristics of the situations they face remain subjective – by being primarily determined by their personal experience (Scazzieri, 2008). Nevertheless, collective identities allow heterogeneous agents to rely on social conventions or institutions in specific contexts. Collective identities and community belonging play the role of mindshaping (that will be exposed in the next chapter), i.e. some perceptions, beliefs, modes of reasoning and inductive inferences tend to be shared among the members of these communities, when confronted to identifiable and known institutional contexts for instance. And this ultimately explains coordination.

Besides individuals' group identities are only one of the self-identities that matter for Bacharach. He quotes Brewer and Gardner (1996) who identify three self-identities: (i) "personal", (ii) "relational" and (iii) "collective" (ibid, p. 74). An individual's personal identity is linked "to aspects of her representation of herself that differentiate herself from others" (ibidem), while her relational identity concerns her "self-conception in terms of relationships with other individuals with whom she interacts" (ibidem). In Bacharach's conception of sociality, individuals' relational identity affects their preferences, goals, mode of reasoning, etc. This is explained by the fact that when interacting, the issue of individuals' decision or behavior depends on the people with whom they interact, their own perceptions, beliefs and goal. Individuals have to guess in what dispositions are these others – i.e. what are their mental states – to make their own decision. This contention strongly echoes Schelling's vision of strategic interdependence. This matter relies on social cognition and more specifically on the attribution theory within the theory of mind (see the contributions of Fiske and Taylor, 1991; Hewstone, 1983; Schneider, 1995; all quoted in Bacharach, 1989, pp. 180-81). Attribution Theory involves individuals' second order beliefs, i.e. their beliefs about the beliefs of the other individuals with whom they interact (ibid, p. 181). Besides the researches on stereotypes and attribution in social cognition that Bacharach uses, "have revealed a remarkable degree of consensus in people's understanding of their social environment, and … shown that the same basic cognitive processes underlie people's predictions

and explanations of their own behaviour and that of others" (ibidem). This is grounded on community membership and on what has been identified as community-based reasoning on the one hand and one the mechanisms of mindshaping on the other hand.

In conclusion we see how numerous are the types of psychology impinging on Bacharach's appraisal of individual decision-making in strategic context which entails coordination and cooperation.

5.3. A different conception of strategic rationality: challenging the individualism postulate.

Bacharach's attempt was first and foremost to formalize more realistic players (Gold and Sugden, 2006) by integrating social and collective determinants in their strategic rationality. Players' choices are not entirely explained by internal consistency of choices but by the context, by cultural determinants, and partly by the others with whom they are interacting, their culture, and so on. From this perspective, Bacharach shares the same vision as Schelling but tries to conciliate this vision with an enriched formal game theoretical framework. For instance, players' preferences regarding games' solutions depend on others' representations or frames. Or, the TR is grounded on the assumption that rationality is collective: players have a collective objective necessitating strong interdependence and knowledge of this fact. Players no longer act for themselves as individuals but for the group. In Sugden' words (2005, p. 183): "[t]he rationality of each individual's action derives from the rationality of the joint action of the team".

In some relevant aspects both the VFT and TR can be considered as "revisionary theories", they are not mere "bounded rationality theories" (Gold Sugden, 2006; Sugden, 2001). As Bacharach and Bernasconi claims (1997, p. 39), in the VFT for instance, "though players are conceptually bounded by their frames, they are not bounded in their strategic thinking within those frames". To the contrary, both VFT and TR offer a *revision of individual rationality* in strategic contexts. Players' choices are no longer strictly individualistic. As emphasized for Schelling's contribution (see section 4 of the chapter 2), Bacharach's contribution while staying consistent with a form a methodological individualism does not agree with the strong form of methodological individualism as defined by Boudon (2004) and imposed by standard rationality – i.e. instrumental and Bayesian – in classical and epistemic game theory (again see the section 4 of the chapter 2 and the chapter 1 for more details on strategic rationality in each case). Collective determinants in players' strategic indeed intervene and impact the way they appraise their decision problem. Both the VFT's and TR's conceptual and methodological characteristics and the profound modifications of the standard conception of players they induce, have necessary consequences in terms of rationality. However, some aspects of rationality in both of them remain quite familiar within a game theoretical framework.

We would like to draw attention on the fact that by using VFT and TR Bacharach challenges the confusion between two distinct aspects of individual rationality which became confounded and

assimilated in decision theory: (i) consistency of choices and (ii) maximization. In the VFT players explicitly maximize their subjective expected utility. Besides,

> "Bacharach interprets the payoffs of a game as specifying what the players want to achieve as individuals (or what counts as success for them as individuals). He assumes that payoffs can be treated as utility indices in the sense of expected utility theory so that, in situations of uncertainty, a player's success is measured by the expected value of her payoff" (Gold, Sugden, in Gold and Sugden, 2006, p. 8).

This raises a difficulty and a possible contradiction with the attempt to integrate players' mental states. EUT is merely a representation of choice. Accordingly preferences and utility cannot be considered as mental variables. We agree that in Bacharach's work players attempt to maximize their expected utility and that they are rational in the sense that they adopt the best reply according to their beliefs but this is very different from Bayes rationality and the EUT. We will adopt the same position as Bacharach in the model of games proposed in the chapter 5.

What matters for Bacharach, in game theory, is to formalize "practical" modes of reasoning, i.e. "what it is that makes actions "satisfactory" and how "satisfactoriness" relates to choice" (Bacharach, 1987, p. 17). As for Schelling, satisfactoriness means coordination. As coordination whether it be in the VFT or TR implies reaching the highest payoff, satisfactoriness relies on coordination. For him, a mode of reasoning must indeed be "valid": "a mode of reasoning in games is valid if it is 'success-promoting': given any game of some very broad class, yields only choices which tend to produce success, as measured by game payoffs" (Gold, Sugden, in Gold and Sugden, 2006, p. 8). However such validity must be compatible with a game theoretic framework in Bacharach's work. This explains why he proposes principles of rational choice that must be followed to rationalize coordination and cooperation in game theoretic terms. In VFT, Bacharach indeed postulates principles of rational choices (cf. 2.3) that any player must follow to be rational. Players must respect these principles in any strategic context, i.e. in any type of games. Regardless of the subjective representations they hold, rational reasoning (according to Bacharach) relies on the principles of rational choices and the equilibrium selection which Bacharach edicts. This is the only way to reach the highest payoffs. In the TR, if players team reason they endorse the objective of their belonging team which is to reach the highest collective payoff. Collective rationality is accordingly "success-promoting" too. From the same perspective as VFT, team reasoners have a determinate mode of reasoning, regardless of the coordination context (i.e. respectively to the amount of information) or the type of games (i.e. in prisoners dilemma, or Hi-Lo game, for instance).

Stating that rationality should merely imply coordination and cooperation contradicts with the standard game theoretic account which leads either to an indeterminacy problem or to mutual defection in the PD and not to cooperation. In fact, for numerous reasons as detailed below, standard individual rationality and the corresponding postulate of individualism are challenged.

In order to understand the changing aspects of players' rationality both in VFT and TR, we link them with Bacharach's early thoughts about the standard conception of rationality in game theory. He identifies, with Hurley, in the introduction of the book *Foundations of Decision Theory* in 1991, what he considers to be the misspecifications of decision theory and game theory. By doing

so, they propose few ways to revise this standard account of rationality. Bacharach and Hurley investigate five dimensions of decision making, of which three are particularly relevant to understand rationality in VFT and TR: (i) "the structure of attitudes", (ii) "humanity" and (iii) "individualism". I intend to highlight how these early thoughts translate in his accounts of VFT and TR.

The structure of attitudes questions "the restrictions on the attitudes of the decision-maker – her beliefs and preferences, subjective probabilities and utilities – that rationality imposes" (ibid, p. 1). They argue that the problem is that "[p]references are often sensitive not just to the intrinsic characteristics of the good but also, ... to the context in which the intrinsically characterized good is located" (Bacharach and Hurley, 1991, p. 12).

In VFT, players' preferences are not "context-independent", since they primarily rely on players' subjective representation which are context-dependent. The outcome that a player ranks as the highest is determined by her frame, i.e. by her perceptions. Her choice is dependent from the context of her decision-making. Players' choices do not respect the standard axioms of "consistency of choices", their choices would not be identical in different contexts even evolving the same options or choices. Besides the way a player labels her options, her strategies, has an influence. As a consequence, the standard axiom of extensionality is violated. The principle of extensionality, states that the descriptions of individuals' options do not change their preferences (Bacharach, 2001a, p. 2).[86] For rational choice theory, extensionality implies that framing concerns irrational players. Naturally, this is not Bacharach's point of view.

In addition, players' preferences depend on their beliefs about others' subjective representation. Again, the options a player ranks higher than others depend on her beliefs about others' perceptions. In standard game theory preferences are totally distinct from players' beliefs about others' choices or beliefs (Hargreaves Heap and Varoufakis (2004 [1995]). In the TR, it is the collective profile of actions of the team which is reason giving. As emphasized above, the rationality of players derives from the rationality of the team (which again means reaching the highest collective payoffs). Players commit to the team (Hédoin, 2018), which could in principle lead them to act against their own preferences. If they team reason, they do not follow their individual preferences but the actions required by the collective profile of actions, i.e. by the team.

The "humanity" of a decision-maker encompasses "her boundedness and her culture" (Bacharach and Hurley, 1991, p. 2). Bacharach and Hurley assert that, even if we suppose an idealized rational player (endowed with unlimited cognitive capacities and perfect knowledge), it may not be sufficient to lead to the determination of solutions in games: "even an extreme idealization of inferential powers fails to deliver the goods in the shape of determinate solutions to games" (ibid, pp. 2-3). As a consequence, "if we are to avoid the conclusion that rational action in games is impossible, it may be necessary to move away from the conception of rationality of traditional decision theory toward a more naturalistic conception in which "ought"

86 How and why framing entails a violation of the standard conception of the individual decision theory, and especially the expected utility theory, is well argued in Bacharach (2001a).

implies "can"" (ibid, p. 3). Taking into account the players' humanity is one of the possible solutions.

The two characteristics at stakes in players' humanity translate in both the VFT and TR theory. Both in the VFT and TR theory, players are conceptually bounded: their frames are incomplete. Bacharach proposes principles of rational play that take into account the perceptions that the players handled. Those principles are established according to players' cognitive abilities and limitations – as players' perceptions are necessarily incomplete due to players' non-omniscience. Rationality therefore relies on what the players *can* do.

Furthermore and as developed above, players are culturally determined. In the VFT, their representations are influenced by cognitive predispositions inherited by their experience and their culture. In the VFT, Bacharach explicitly assumes that "the rational solutions of some games depend on the culture of the players" (Bacharach, 1993, p. 271). He explicitly integrates the impact of players' culture in their utility function through the availability function (which symbolizes the salience of the different frames). The variations of values of this availability function translate the effects of culture on individuals' perceptions (ibidem). Players' frames are therefore constrained by their culture. Different cultural backgrounds induce different structures of frames and this fact is taken into account in Bacharach's formalism.

In addition, in an environment characterized by a multileveled society, with multiple types of interactions between heterogeneous agents – or groups –, conventions or institutions drive toward convergent patterns of behavior. A common social background stored in cultural knowledge supplements the non-omniscience of individuals and the bounds of their cognitive abilities in decision-making (see Bacharach and Hurley, 1991; Schmidt and Livet, 2014). When the cost of computation is too high, or the uncertainty surrounding who are the others with whom someone interacts, she tends to rely on her knowledge of conventions or social rules in order to decide and to act (Schmidt and Livet, 2014). When contexts of interaction are familiar and some institutional devices – bringing coordination – are known, players can even rely on what Bacharach identifies as non-occurent knowledge, which corresponds to what will be presented in the next chapter as direct social perceptions. However such form of community-based reasoning requires, as emphasized by Hédoin (2014, 2016), that players recognize belonging to the same community in which this institutional heritage prevails.

According to Bacharach, a pervasive trait of humanity is the ability of individuals to coordinate (and eventually cooperate). For him, real individuals are perfectly able to coordinate when they rely on institutional device inherited from a cultural or social community. In fact, he mentions that "human framing propensities stand behind the well-known ability of people to solve coordination problems by exploiting 'focal points'" (2001, p. 7). Focal points are devices of coordination. They enable heterogeneous people to recognize consistent and convergent pattern of behavior; i.e. to identify "congruent structures" (Scazzieri, 2008, p. 187). In this perspective, focal points intervene in individuals' decision making like institutions for Aoki (2001). Focal points act as mindshaping devices and ultimately entail community-based reasoning.

In a similar manner, Aoki (2001, pp. 3-4) "conceptualize[s] an institution as a salient, common component of the players' subjective game models—that is, as shared beliefs about the structure

of the game that they actually play." Individuals progressively 'select' successful rules of behavior, i.e. allowing coordination, through the recurrence of interactions and the repeated use of coordination devices – like focal points. Focal points, as repeated coordination devices, belong to individuals' cultural knowledge, i.e. to their social and generalized conceptual repertoire. Due to the ability of individuals to classify by congruence and similarity, focal points allow information filtering and individual selection of behavioral rules (see Schelling, 1960; Scazzieri, 2008; Aoki, 2001).

Therefore, in some circumstances, the rationale of agents' decisions relies on their non-occurent knowledge, i.e. on the rule-following form of individual knowledge.

Players' cultural/social background is also of great importance in order for them to have beliefs towards the others perceptions and beliefs. Recall that community-based reasoning relies on common inductive inferences which means having the same perceptions, beliefs and then adopting the same action. The ability of agents to engage in interaction and to coordinate relies on a common background. Coordination relies in the sharing of some representations stored in a given community (see the mindshaping hypothesis in the next chapter). Again this implies that players' rational play according to their beliefs of others is socially determined.

> "[W]ithin a same culture the parameters of a conceptual scheme – its membership, its clustering, and the readiness to mind of the clusters in a given situation – are essentially shared. Furthermore, they are shared in this strong sense: not only does everyone have certain conceptual competences by virtue of belonging to the community, but every member knows that every member has them" (ibid, p. 259).

Referring to empirical results (in Metha, Starmer and Sugden, 1994a,b; Bacharach and Bernasconi, 1997; Bacharach and Stahl, 2000), Bacharach argues that within a common culture, e.g. in a particular community language, individuals tend to represent situations in the same way (Bacharach, 2001, p. 9). In other words, individual frames have a propensity to be shared. Individuals are inclined to partly hold common structural frames (i.e. common conceptual repertoires, common classifiers concepts, etc.). Again, this echoes mindshaping and community-based reasoning which again entails common perceptions, beliefs and inductive inferences (Hédoin, 2014, 2016). The partially sharing of frames will in turn enable individual agents to form beliefs about others' perceptions and beliefs (Bacharach 1986, 1990). This propensity of shared perceptions within a common culture (or community) induces a tendency for mutual consistent beliefs about each other's perceptions. That is, beliefs about each member's frames (and availability functions) tend also to be shared (Gold and Sugden, 2006, p. 8).[87] In the VFT, because players belong a common community language (Bacharach, 1991) they can attribute to the others the same set – or eventually a subset – of their own frames and beliefs – since they are defined within the players' frames (Bacharach, 1993).

Therefore, players can ground their reasoning on the way the others generally perceive the coordination context they face and on what they have learned about coordination devices

87 See Metha et al. (1994a,b) and Bacharach, Bernasconi (1997) for experimental confirmations of this statement.

inherited from this culture. Since players think of salience strategically, it means that they think about what is salient for others. Players' capacity of empathy is therefore at the basis of Bacharach's methodology to explain players' capacity to solve games. In this perspective, they use the knowledge their culture gives them. Indeed, players cannot think of others' perceptions if they are not of the same background (Bacharach, 1991, 1993, 2001; Bacharach, Bernasconi, 1997), which implies that each player's perception must be embedded in a common conceptual framework, i.e. in a common cultural legacy.

In a given (language or cultural) community, the tendency of sharing of frames and in this perspective of individuals' beliefs, can justify the hypothesis of "transparency of deliberation" and focal points in particular orient toward a "transparency of deliberation". Even if individuals' perceptions are subjective, from shared perceptions stemming from socialization, they can draw common inferences, when referring to focal points. In Aoki's words a focal point – as an institution – act as "a guiding symbolic system [that] becomes consistent with, and reconfirmed by, [individuals] experiences. It then serves as their summary representation of equilibrium incorporated into agents' stable beliefs" (Aoki, 2001, p. 19).

Finally, Bacharach and Hurley, clearly stress that the "individualism" postulate, mainly in game theory, is question begging. They claim that

> "a number of questions arise about the relationship between individual rationality and game-theoretic rationality. We usually think of individual decision theory as in a sense prior to game theory, but as several points … we are led to question whether games may be embedded within supposedly individual decision problems. Moreover, there may be doubts about indeterminacy of individual rationality of the traditional sort in the interactive setting of games." (ibidem)

Such indeterminacy refers to the incapacity of game theorists to provide determinate solutions for some types of games (like the coordination games). The VFT and TR theory precisely offer the methodological ground to escape such indeterminacy problems, and for that, they draw on a non-individualistic form of reasoning. Again, both the VFT and TR bypass the presupposition that strategic rationality should fit with individual decision theory, and accordingly with individualism.

As already emphasized the determinants that intervene in players' rational play are no longer intrinsic to themselves but determined by the others and by social and cultural determinants. In VFT, players' rationality includes a sense of 'otherness', what it is rational to do for a player intrinsically depends on the others. In a way, in the VFT, players integrate in their reasoning that solving a game is a collective purpose. Players' rationality is no longer merely individualistic.

Then, the challenge of the individualistic form of reasoning is particularly obvious in TR. Bacharach (1995, p. 1) explicitly stresses that "cooperative reasoning … must differ from the type of reasoning standardly assumed in game theory". Moreover, Bacharach (1997, p. 13) asserts that "We-play is rational in the external sense of being efficacious or functional". Team reasoning possesses its own requirements in order to be valid: "cooperative reasoning is *sui generis* and not derived from the standard (individualistic, instrumental) type of reasoning" (ibid, p. 25). In the

TR theory, the "unit of agency" is a team and no longer individuals. Players' mode of reasoning entirely relies on the team's rationality.

I would like to mention a last element that explains players' capacity to coordinate and cooperate: natural selection. In such perspective, rationality cannot be discarded from it evolutionary dimension.

For Bacharach, natural selection runs both at the individual and collective levels. In long period "natural" selection operates at a group level. This selection is in favor of the groups which perform well, i.e. structured and organized groups in which the members coordinate and cooperate – through cultural coordination devices (see Bacharach, in Gold and Sugden, 2006, chap. 3). In his work, Bacharach puts also a great emphasis on self-perception or self-categorization as a fundamental process in human cognition. He argues that the way individuals perceive themselves (and others) determines their capacity to coordinate and cooperate (ibid, p. 70). As we have seen, for Bacharach individuals' group identities matter in their decision-making. Individuals are constituted by multiple group identities which in turn enhance collaborative and cooperative behavior. This paves the way for the introduction of Bacharach's evolutionary argument. He postulates "that group identification is the fundamental evolved proximate mechanism for collaboration in man" (ibid, p. 112). He thinks that

> "[M]an has evolved to have a repertoire of cooperative behaviours geared to different types of situations with scope for cooperation … Call the hypothesis that people are endowed with such a range of cooperative, situation-dependent behaviours the cooperative repertoire hypothesis." (ibidem)

Each individual inherits a cooperative conceptual repertoire, which explains why individuals are able to successfully coordinate and cooperate. Finally, if selection works at a collective level, groups whose members are effective at cooperation have a better chance to perform well and survive in the long term (see Bacharach, in Gold and Sugden, 2006).

To sum up, our view is that players' rationality in the VFT and TR is highly different from all the others trend of research investigating the limits of individual decision theory in game theory. As already mentioned, players' preferences (or their reasons for acting) depend on the context, their perceptions, their beliefs, others' perceptions, their culture, and their self-identities. In other words, the determinants of choice are no longer intrinsic to players. Because players' subjective representations depend on time and place (as for Schelling, 1980 [1960]): "frames may vary both across players and from occasion to occasion" (Bacharach and Bernasconi, 1997, p. 4) and because "[t]he agent's description matters strategically" (ibid, p. 5), the outcome of rational plays is necessarily specific, contextual and socially grounded. We recognize in that claim Schelling's conception of strategic interdependence. Both the conceptual and methodological implications of Bacharach's purpose to appraise players' decision process in the game theoretic approach of players and interpersonal strategic interactions render the players' rationality more complex than in any other behavioral model. And for that purpose he draws on many other approaches belonging more broadly to cognitive sciences.

6. Conclusion

Throughout this chapter we have insisted on the difficulties faced by Bacharach in regard to his conceptual and methodological choices both in VFT and TR. For instance, in VFT two of his major challenges are to draw a game as an objective coordination device and to justify the players' capacity to coordinate. Therefore, he repeatedly resorts to *ad hoc* assumptions. In the same manner, his theory of TR enhances troubles mainly materialized into the endogenization problem, and again to justify cooperation he resorts to *ad hoc* assumptions. In this perspective Bacharach's achievements do not seem to equal his ambitions. I suggested that this is mainly explained by the fact that Bacharach wants to incorporate subjective determinants in games and players' reasoning, as well as determinants stemming from outside the games and which do not belong to a game theoretical framework.[88] This points out that some ways to enrich game theory are not easily compatible with its research program and formalism. As a close and coherent system, blurring some of the frontiers of game theory may endanger the models proposed to challenge it. Some paths, and precisely the path taken by Bacharach are in some significant aspects hardly concealable with the core of standard non-cooperative game theory. However as, the chapters 4 and 5 will show, Bacharach opens the way to the formal integration of players' mental states and in particular to their beliefs understood as real mental states and not mathematical artefacts that are used to describe a rational choice at the equilibrium as in epistemic game theory. In this perspective, he attempts to offer some answers to the questions of how and why an equilibrium occurs as the players' reasoning process is integrated as an explanation of their capacity to coordinate. Pursuing the contribution of Bacharach we will add to his formalism a justification of the way players' attribute to the other players some perceptions and beliefs, drawing on the theory of mind; a theoretical justification that was lacking in the VFT and that can explain why he resorted to *ad hoc* assumptions.

We would like to highlight Bacharach's merit in asking worthwhile questions and in underlining the limits of game theory which are rarely discussed and emphasized in such a way. From that prospect we see the similarity with Schelling; except that Bacharach also challenges from a theoretical and formal point of view standard and epistemic game theory. He formally extends standard game theory showing at the same time the porosity of some its frontiers and the impermeability of some others. Like Schelling, Bacharach is against a purely mathematical conception of game theory in which games are self-contained worlds and in which the equilibrium is an assumption of the theory more than a result so that the resolution process of a game is not investigated. The interest of Bacharach is to make compatible an open vision of game theory with a mathematical formalism which nonetheless, as already underlined, bypassed the purely mathematical conception of games entailing the *a priori* postulate of the existence of a solution as the game describes the choice of the players at the equilibrium.

88 Schmidt and Livet (2004, chapter 2) indeed argue that it is the problem of intersubjectivity more than subjectivity that was predominant for economics.

Bacharach incorporates social and cultural determinants in a very different way from the current new behavioral economics. Like Schelling, he questions the intersubjective dimension required for players to form beliefs about other's frames, beliefs and then choices with the integration of not only subjective and personal determinant but social and collective determinants, that can ultimately induce the convergence of players' beliefs and choices, i.e. induce the state of consistency required for the existence of a solution. For instance, social and "other regarding" preferences (Rabin, 1993; Levine, 1998; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Fehr and Schmidt, 2006; for a selection of the most representative models) incorporates new arguments in players' utility functions, without challenging the content of individual standard rationality. Besides, this literature cannot justify coordination (see also Lecouteux, 2018). The standard conception of economic individuals remains. Players still maximize a pre-determined utility function and continue to have the same objective, namely the maximization of their individualistic utility function. The only change is that social preferences "internalize sociality" (Davis, 2011, p. 89) and that players have "social tastes" (ibid, p. 85). It however poorly enhances the resolution of coordination problems in game theory.

Bacharach acknowledges that he could have incorporated other social and cultural determinants in the players' choice to justify coordination and cooperation, in a similar fashion as is done with "other regarding preferences". Nevertheless, it would not have enhanced the debates at stakes in both VFT and TR. Besides, Sugden (1993) – among others – stresses that 'altruism' and accordingly altruist preferences (a subcategory of the other regarding preferences (see Sobel, 2005) may be inconsistent with standard instrumental rationality (he especially refers to Andreoni [1990]) and may not be sufficient to explain cooperation.

Bacharach seeks to explain individual rationality by "reason within nature": "reason is a part of nature, so adopting the rational approach does not mean sacrificing the insights of the evolutionary perspective. The method of evolutionary psychology allows us to explain choices by 'reason within nature'" (Gold and Sugden, 2006, p. 7). In fact, "Human beings are endowed with capacities for reasoning that are the product of evolutionary selection; they then apply those capacities to whatever specific decisions problems they face" (ibid, p. 9). However, quite obviously, his perspective is highly different to the evolutionary game theory.

Moreover, even if Bacharach did not stimulate a huge trend of research, some of his intuitions are valuable. The question of framing is rarely discussed by game theorists although it is considered as a major research agenda by allowing games to become open-systems (Binmore, 2009). This explains how Bacharach plays an important role for this thesis. We will show in chapter 5 that the integration of players' mental states has for consequence that the games can no longer be self-contained worlds but need to be open-systems, i.e. 'variable-universe games' (in Bacharach's terminology). A new intersubjective dimension for explaining the players' capacity of attributing to the others' some beliefs and choices is required. This intersubjective dimension is provided, as already mentioned, by the simulation theory. Bacharach nevertheless initiated a few amount of contributions attempting to give a game theoretic content to frames and to explain focal points via framing (e.g. Sugden, 1995; Casajus, 1998, 2000; Colman, 1997; Janssen, 2001, 2006). His work participates to a largest thought on collective entities, still rare in economics and which offers a dialogue with philosophers and social ontology. As in Schelling's view of strategic interdependence, both VFT and TR emphasize that players have a sense of otherness and that

being successful, i.e. coordinating, is a collective enterprise. In a way, in VFT, players integrate in their reasoning that resolving a game is a collective purpose. The 'other' is materialized in players' reasoning since the 'other' has perceptions; the 'other' is no longer a mere probabilistic event or, shall we say, an objective event like in Bayesian game theory (Lesourne et al., 2006, p. 69). Some contributions enhance Bacharach's introduction of framing in game theory (e.g. Gold and List 2004). Papers have deepened Bacharach's intuition with respect to the switch from individual to collective reasoning according the characteristics of the game and the perceived strong interdependence of decision and action (Zizzo and Tan 2007, 2009; Tan and Zizzo, 2008; Smerilli, 2012). In TR, a collective purpose can be reason giving, teams count as a "unit of agency" for a game. Indeed, Bacharach purports to justify that agents can rely on collective entity, and can naturally consider the collective as the relevant "unit of agency" to resolve decision problems. Contributions on TR continue to be proposed. Different experiments have been conducted after Bacharach's work to test the predictive power of the TR (Bardsley and Ule, 2017; Faillo, Smerilli and Sugden, 2017; Pulford et al., 2017; Colman et al., 2014; Bardsley et al., 2010; Colman et al., 2008; Butler, 2012). And a recent special issue on TR has been published in 2018 in the *Revue d'Economie Politique* for instance gathering 8 papers from economics and philosophy.

Finally, Bacharach claims that the theory of TR crosses the frontiers of non-cooperative game theory. It can necessitate the combination of non-cooperative game theory, cooperative game theory and multi-agent systems (2001b). From this perspective, it is our opinion that Bacharach offers a new line of research for modeling collective decisions and collective behavior in cognitive economics, in economics of networks and in multi-agent systems. This is very similar to Schelling's work on this point. Some of the conceptual and formal aspects of both Bacharach's VFT and TR seem more suitable within these frameworks than in standard non-cooperative game theory. This is especially the case regarding his endogenization problem. What is certain is that both the VFT and TR are critical of the standard and closed-system mathematical frontiers of game theory. Psychology, history, biology, philosophy and more generally cognitive sciences, influence Bacharach's work, which is grounded on an open-system methodology. His theories are open-systems. He attempts to conciliate formalism which such epistemology which is according to social ontologists not incompatible (Chick and Dow, 2005; Chick, 2004; and Mearman 2003).

Chapter 4

A new frame for intersubjectivity in game theory: the insights of the Theories of Mind and Simulation

1. Introduction

Discussion around intersubjectivity (when it exists) is related to two dimensions: (i) strategic rationality in game theory with respect to the way players are able to from expectations on other players' choices and (ii) empathy which is absorbed by the literature on interpersonal comparisons of utility and social preferences. This chapter on the one hand will show that empathy is not confined to an ethical dimension as is nowadays often assumed but is broader and encompasses a cognitive or epistemic dimension which is of particular interest for game theory. This chapter will also show that empathy does not necessarily involve other-regarding feelings and behaviors; empathy also refers to a psychological and cognitive mechanism allowing understanding of someone else – by eventually sharing some feelings, some mental states (intentions or beliefs) or some cognitive schemes, ways of reasoning, etc. For psychologists, there are indeed both a cognitive and an affective dimension in empathy (Eisenberg and Strayer, 1987). In its cognitive dimension, which is of particular interest for us, empathy means intending to understand the other – e.g. her feelings, dispositions, and characteristics like preferences or needs, etc.

As the problem of coordination became less central in economics during the 20th century, discussion on empathy and intersubjectivity in economics disappeared even though they had "a long history" (Kirman and Teschl, 2010, p. 304.) Intersubjectivity in economics is indeed rarely discussed despite being a pillar to human interaction and coordination (e.g. see Binmore, 1990, 1993, 1994; Schmidt and Livet, 2004; Singer and Fehr, 2005; Kirman and Teschl, 2010). This is one of the consequences of the denial of psychological and subjective determinants in individual decision-making and the escape from psychology proposed by Pareto and Samuelson that culminate in Savage's contribution to microeconomics (Giocoli, 2003). The type of intersubjectivity supposed in game theory through common knowledge of rationality, whilst highly problematic (see chapter 1), is unanimously assumed and rarely stressed. Thus, on the other hand this chapter will present the type of intersubjectivity that can be offered by the

epistemic side of empathy which is called in cognitive sciences: mindreading. The type of intersubjectivity assumed in mindreading is compatible, as we will show in this chapter, with a strategic form of reasoning embedded in methodological individualism (understood in a broad sense). The kind of intersubjectivity assumed, while compatible with some form of collective knowledge induced by the experience of sociality, by recurrent interactions, etc. will have as a basis the individual, her own apprehension of the situation she faces and of the other. Individual subjectivity is thus compatible in our approach with some form of knowledge about the other.

Intersubjectivity should be of particular interest in economics and even more in game theory, as a prerequisite of the capacity of individuals to anticipate each other choice. Indeed as Kirman and Teschl (2010, p. 303) put it

> "One might well ask why considerations of empathy disappeared for so long from the economics literature. One answer is that as economic theory developed and became formalized in the twentieth century, almost all of the emphasis was put on the idea of anonymous individuals satisfying specific axioms of rationality and interacting only through the market. In such a view, there was no place for the idea that individuals might want, or need, to put themselves in the place of others. However, with the development of game theory, such an idea became central. Here, the idea is that individuals interact directly and consciously with each other."

Though intersubjectivity is of particular interest for game theory, as Kirman and Teschl remind us, it is rarely stressed in the reflexive literature on game theory. Such lack of interest in intersubjectivity is quite surprising considering that the very fundamentals of strategic interaction requires an intersubjective dimension in order for players to be able to form beliefs on other players' beliefs and choices. The CKR as the only pillar of intersubjectivity, i.e. as the only methodological device allowing players to anticipate each other's choice – i.e. to form beliefs regarding each other's choice – is, as is well recognized, more than problematic. It is besides enable to bypass the indeterminacy problem. It is well known that game theory faced for several decades major challenges with respect to coordination problems. The need to find alternatives for the required intersubjective dimension in games is therefore of urgency. We suggest in this chapter consideration of the Theories of Mind (ToM) as one of such alternative and in particular Simulation Theory (ST). From this perspective, the last chapter will show how to incorporate ST in games. The aim of the thesis is to show that some possibilities exist to counter the various difficulties that game theory faces with respect to coordination problems and more specifically to explain how and why players may coordinate in games, as has been discussed in the first chapter. The second part of the thesis shows the ways opened respectively by Schelling and Bacharach. From that perspective, this chapter purports to provide an explanation of a new intersubjective dimension that can replace the standard common knowledge rationality (CKR) or the common beliefs in common (Bayesian) rationality – which supposedly leads to well specified prior beliefs – assumptions. As argued in the first chapter both CKR and Bayesian rationality led to deep and pervading stalemates. This chapter will provide a continuation of Schelling and Bacharach's work, mainly grounding their respective intuitions on a contemporary theory within the cognitive sciences, the Theory of Mind, and will explain why it may be important to integrate this theory in a formal game theoretic framework. Such a formal framework will be proposed in the last chapter of the thesis.

Contrary to the standard view of empathy in economics, empathy will be appraised in this chapter and the next one in its epistemic dimension only, i.e. as a means for being acquainted with other's perceptions, beliefs, desires or intentions, i.e. the mental states leading someone to adopt a specific act: e.g. cooperating or defecting in a prisoner dilemma. More specifically in this chapter we will refer to the term empathy as the cognitive mechanism consisting in putting oneself in the others' shoes which ultimately ensures the intersubjective dimension needed in games to form beliefs toward others' beliefs and action. Empathy will be approached through the Theories of Mind and we will in particular defend Simulation Theory within this framework as a means to account for this intersubjective dimension in game theory.

The history of the links between economics and empathy, as a frame for intersubjectivity is very limited. We can indeed mention Fontaine (1997, 2000), who offers, to my knowledge, the only attempt to propose a systematic overview of the relationship between economics and empathy; Singer and Fehr (2005) who focus on the ethical (and altruistic) dimension of empathy rather than the epistemic dimension, and Kirman and Teschl (2010) who more focus on empathy within behavioral economics and game theory. The amalgamation of empathy and altruism is mainly explained by the rational choice theory paradigm (Sugden, 2002; Kirman and Teschl, 2010). Within this framework any intersubjective consideration must be integrated in the individual preferences or utility functions, i.e. translated in terms of altruistic preferences or other-regarding preferences. The latter allows rationalizing of non-selfish or altruistic behaviors in the economic sphere. However, such formalism ultimately leaves no room for empathy understood as a form of epistemic device allowing acquaintance with other's mental states and therefore as a means to form beliefs on their choices (for a comparative claim see Kirman and Teschl, 2010, p. 303). This explains why we will ultimately depart from the standard rational choice paradigm in the last chapter in order to integrate a new frame for intersubjectivity with the theory of mind. Integrating players' mental states in game theory and, thanks to the theory of mind providing an explanation of the way players may tacitly understand each other, which again, ultimately explains how they form beliefs towards each other beliefs and action, must require accepting some deviations from standard game theoretic rationality and in particular Bayesian rationality in game theory. Nevertheless, empathy in its epistemic dimension has received an upsurge of interest in psychology in recent years, thanks to the development of neuroscience, and to recent studies using imaging which have led to new and enlightening findings (see Zahavi, 2014; Singer and Fehr, 2005; and for an early review on neuroimaging studies see Gallagher and Frith, 2003). Paradoxically this is not the case in epistemic game theory which this upsurge of interest has not penetrated.

It is on the cognitive, mindreading or mentalizing dimension that the Theory of Mind (ToM) focuses. The ToM which is nowadays at the center of cognitive sciences – and encompasses philosophical, psychological, social psychological and neuroscientific approaches – investigates the capacity of mental-states attribution to self and others (Goldman and Shanton, 2012, p. 1). These mental states concern individuals' perceptions, feelings, emotions, desires, intentions, preferences and beliefs. The existence of a ToM is therefore a prerequisite in human capacity to interact. Understanding the others' perspective, and therefore attributing to her beliefs and intentions for instance, is necessary to come to a prediction of her choice and behavior. Attributing mental states to the other in order to understand her behavior can in turn allow

prediction of her behavior – providing there is sufficient information on the decision context triggering the potential utterance or behavior (Churland, 1991; Goldman, 2012). Both these capacities are at the very foundation of strategic interactions. For instance, in repeated games, a player interprets the past behaviors of her co-players by using the mental states such as the intentions or beliefs that triggered that past behavior to predict their future choices. According to such beliefs toward the others' choice this player will be able to play her best reply and maximize her expected payoff. Recall that in epistemic game theory there is no room for such mental states, as player's beliefs are, as already heavily underlined, merely mathematical or notational artifacts used to represent the choice of the players at the equilibrium, once the game is solved. In one-shot games, a player must attribute to her co-player(s) intentions and beliefs to predict her/their choices and come to a decision. Understanding how and why an equilibrium occurs in a one-shot game requires understanding of this strategic reasoning and therefore of this mechanism of attribution. For instance, a player can assume that her co-player has the intention to cooperate and believes that she is going to cooperate too. The player will therefore come to the conclusion that her co-player will cooperate and that her best reply will be cooperation. However, if the player assumes that her co-player intends to cooperate but believes that she is going to defect, it will be in her interest to defect and not to cooperate, contrary to her intention. In that case the player believes that her co-player will defect. It is thus in her interest to defect too. Indeed mindreading can "enable people to predict others' behavior and, therefore, help them meet their individual goals." (Singer and Fehr, 2005, p. 343)

In the ToM, empathy is therefore merely a mechanism of mental states attribution. Mental-states attribution has been variously labeled since the 20th century as 'folk psychology', 'theory of mind', 'mind reading', 'mentalizing' or 'empathizing'. In the history of the ToM, whatever the labeling of this mental-state attribution phenomenon the focus is on the intersubjective dimension entailed in empathetic identification, and its role in individual interactions. Here intersubjectivity is understood merely as a mode of interpersonal communication and understanding, either explicitly through language and dialogue, or tacitly through pure reasoning or by behavior and the ability of individuals through empathy to give meaning to and eventually understand individual behavior.

In the history of the ToM the three main accounts of mindreading (i) the Theory-Theory (TT), (ii) the Rationality Theory (RT) and (iii) the Simulation Theory (ST) successively dominate the landscape of the field. The explanation of mindreading was first grounded on a folk or naïve psychological theory of mental states (the TT). Then the rationality postulate was added stating that the ordinary man is a rationalizer so that his mental states are grasped by this rationality postulate (RT). And finally these inferential approaches have been challenged by the very primitive heuristic of putting oneself in the other's shoes to try to be acquainted with her mental states (ST). Goldman is one of the defenders of the ST who brings to our view one of its most accomplished and comprehensive versions. We will subsequently mainly rely on his contribution hereafter and defend his account as particularly convenient philosophically and methodologically, for economic theory and more particularly for game theory.

Traditionally the philosophy of mind had dealt with two intertwined aspects of mentalizing; the metaphysical part and the epistemological part (Goldman, 2006). While the former addressed questions relying on the nature and the essence of minds and of mental states, the latter handled

questioning such as: how can people have access and know both their own mental states and that of others? i.e. how can people approach mind functioning? What is the nature of their knowledge of their own and other minds? These epistemological roots received the most attention within the literature. The main developments of the ToM followed this epistemological investigation. As Goldman (2006, p. 5) put it "[i]n terms of the philosophical literature, mindreading is naturally understood as a descendant of these epistemological problems." Therefore even the recent and central achievements of the ToM, whatever the approaches (TT, RT or ST), are oriented towards this epistemological investigation.

The contributions of the ToM more specifically attempt to provide answers to four main questions: (i) how do people mindread – i.e. attribute mental states to – themselves and the others, (ii) how this capacity is acquired and developed, (iii) what are the contents of the mental states (i.e. the beliefs, preferences, intentions, etc.) attributed, and (iv) what are the cognitive mechanisms involved in mentalizing (see Goldman, 2006, p. 21). According to the importance and answers given for each of these questions, some variation arises within the ToM literature, and the different approaches mentioned above can be identified. Recall that mental states refers to psychological notions, such as individual beliefs, preferences, intentions, perceptions, etc. They refer to psychological notions instead of the standard account of beliefs or preferences in classical and epistemic game theory.

Bringing into economics a discussion of the most recent developments of the ToM that we will present and detail in this chapter will clarify the mechanisms involved in empathetic identification and will clarify why empathy is an important determinant in individual decision-making in interaction. This will ultimately allow us to show how insightful it can be to bring into economics and game theory the developments of the ToM and to explain why we claim that the role of empathy has been wrongly underestimated. Our interest is first and foremost in the concept of empathy, and in particular in its cognitive assessment as a way to bypass the methodological difficulties emphasized in the chapter 1 regarding Bayesian rationality in game theory. We propose to show that empathy as developed by a particular instance of the ToM, and translated into the realm of game theory, offers a methodological solution for bypassing the main issue of recent non-cooperative game theory: it has the incapacity to account for the players' cognitive processes in strategic interaction and for the mechanism of convergence required for the existence of a solution in games. In other words, for explaining how and why players coordinate. Therefore, discussing the insights of the ToM can show that far from being a controversial determinant that eventually, in very specific contexts, intervenes in the individual decision making, empathy ultimately characterizes the ability of the decision maker to perceive her world as encompassing other decision makers and as a consequence, to take into account these others in her decision making.

To set the extent of the contributions that can be brought to game theory by the ToM we will detail its different approaches: the TT, the RT, and the ST. Among these approaches some debates exist with respect to the cognitive and/or social mechanisms involved in empathetic identification, and the way it is executed and acquired during infantile development (Shanton and Goldman, 2010, p. 1; Goldman, 2006). The hot topic of mentalizing, which is a 'second order mental activity' and therefore (initially) unobservable, first falls within the metaphysical and epistemological branches of Philosophy of Mind. The mentalizing issue has then progressively

been tackled by Psychology and in particular Developmental Psychology, which introduced experimental methodology and more recently by Neuroscience, which at last makes – some of – the cognitive processes of mentalizing observable. Neuroscience helped to develop the approach that we will refer to in this chapter and integrate in chapter 5. By providing data on individuals' interactive cognitive processes, Neuroscience confirmed the shift from the first approach of mentalizing, the TT, to the third approach, the ST, that was happening in the end of the 20th century. Neuroscience finally offers new empirical perspectives for empathetic identification and intersubjectivity in economics and game theory. These help to develop the type of intersubjectivity that we will integrate in a model of games in chapter 5 by providing data on mindreading understood in its epistemic dimension, i.e. the pertinent dimension for strategic reasoning in games.

In this chapter we will assess each of these approaches, assuming that the ToM is still quite unknown in economics and game theory. Very few contributions in economics discuss the ToM or employ any of its insights. Among these contributions can be mentioned for instance Morton (2012), Devaine, Hollard and Daunizeau (2014), Schmidt and Livet (2014), Guala (2016; 2018), Hédoin and Larrouy (2016) Larrouy and Lecouteux (2017) Lecouteux (2018a,b). As Singer and Fehr (2005, p. 340) declare, even though game theory necessitates the individual capacity of anticipation regarding each other's choice, which "require people to be able to view the game from the other players' perspectives (i.e., to understand others' motives and beliefs) … [e]conomists still know little about what enables people to put themselves into others' shoes and how this ability interacts with their own preferences and beliefs." Fehr and Singer's assertion is still true more than 10 years later; economists still know very little of the capacity of individuals to be acquainted with the others' mental states. The theories of mind offer such knowledge and detailing the different approaches helps to underscore to what extent such knowledge is provided by them. The exegesis of the whole ToM framework helps to underscore its main achievements and its issues, but foremost, the new contribution and enhancements that it can induce for game theory. Even if each of the above mentioned approaches can be relevant for game theory, and even if the TT and the RT approaches as we will see seem to be much more familiar and compatible for a game theoretic framework, we will detail the ST more extensively than the others to explain why we draw on the ST account of mindreading in the rest of the thesis and why we choose to integrate the ST in our analysis and formalization of coordination in game theory (cf. Chapter 5).

From the same perspective, Guala (2018, p. 356) defends the value of ST for coordination by escaping to the infinite regress problem imposed by the symmetry position and the CKR. Besides, the value of the process of tacit communication and tacit acquaintance with others' mental states involved in ST emphasized early on by the father of the formalization of CK in game theory, Lewis. He declares

> "We may achieve coordination by acting on our concordant expectations about each other's actions. And we may acquire those expectations, or correct or corroborate whatever expectations we already have, by putting ourselves in the other fellow's shoes, to the best of our ability. If I know what you believe about the matters of fact that determine the likely effects of your alternative actions, and if I know your preferences among possible outcomes and I know that you possess a modicum of practical rationality,

> then I can replicate your practical reasoning to figure out what you will probably do, so that I can act appropriately. (Lewis [1969, p. 27], quoted by Guala, 2018, pp. 362-63)

As we will assert, ST makes focal point particularly relevant for coordination in strategic thinking. This assertion will be developed largely in the last chapter. But the process of simulation will make the salient outcome of the player more likely to be attributed to her co-player, which ultimately raises the probability of occurrence of this salient outcome in players' beliefs. We could also refer to Guala (2018, p. 356) who defends the relevance of focal points for coordination by using the ST.

A last dimension explaining players' capacity of coordination is a form of social cognition that, contrary to mindreading, does not require mentalization to be acquainted with someone else. From that perspective two accounts can be mentioned: (i) what we identify as the Direct Social Perception (DSP) literature, and (ii) Zawidski's mindshaping theory (Zawidski, 2013). To a certain extent Mindreading more focuses on a psychological dimension to explain individuals' capacity to be acquainted with each other, while the DSP literature and mindshaping link this capacity to a social dimension. According to the latter, being acquainted with someone else can be explained by simple perceptions and the knowledge of the context of interaction which in many cases provides enough information to coordinate with each other. This conception though bypasses the mere introduction of framing. The experience of sociality entails the knowledge of some behavioral patterns, of some norms and conventions which ensure coordination. This social dimension of cognition and of coordination has also been developed alongside the mindreading dimension in both Schelling's and Bacharach's work and it plays a considerable role for focal points and coordination. Besides the degree of homogenization brought about by the existence of norms and conventions as we will see a considerable role played by mindreading. It increases the reliability of the complex task that mental states attribution requires in mindreading. It reduces the set of possible perceptions, beliefs, or **intentions that** someone can attribute to the other.

Recall that the way coordination is understood in this thesis is broader than the paradigmatic example of coordination games like the Hi-Lo game for instance. Coordination is understood as the convergence of player's perceptions, beliefs and behaviors and such processes rely on the many dimensions that Schelling in particular developed in his theory of interdependent decisions.

2. On intersubjectivity and empathy in game theory: a very restrictive integration

According to Kirman and Teschl (2010, p. 304) the developments in experimental economics and behavioral economics and more specifically in neuroeconomics have induced a revival of interest in empathy in economics (for reviews of this literature see e.g. Decety and Lamm, 2006; Singer, 2006; Singer and Lamm, 2009).

This new interest could have induced fertile thinking on intersubjectivity in game theory. However, as this section will show, empathy understood as a mentalizing capacity, i.e. enabling players to be acquainted with others' mental states (beliefs, intentions or preferences) and their choices, remains a minority interest which ultimately does not bring deep thinking on intersubjectivity in game theory.

In behavioral game theory, empathy is often, on the one hand, considered as a determinant for pro-social behaviors and, on the other hand, as an emotional response. The extensive use of neuroimaging induces the investigation of the role of emotions – as considered as an empathetic dimension – in games such as the ultimatum or dictator games (e.g. see Singer and Fehr, 2005; Singer et al., 2006). Such literature therefore nurtured one dimension of the literature on framing effects (i.e. emotions entail a deviation from a rational choice) rather than a deep foundational thinking in game theory. This emotional dimension is identified as an "ethical use" while from our perspective the most fruitful one is the "epistemic use" (Kirman and Teschl, 2010, p. 304). Indeed, "the only persistent use of empathy in economics has been as an instrument of interpersonal comparison of utility with the aim of constructing social preference orderings." (Kirman and Teschl, 2010, p. 304) In a minority of cases however, and very recently, neuroimaging has been used in economics to investigate the epistemic dimension of empathy (Kirman and Teschl, 2010, p. 305). Even if it is acknowledged that "understanding the process of how people put themselves in the shoes of others will help to clarify what knowledge people can reasonably have about other people's beliefs, intentions and motives" (Kirman and Teschl, 2010, p. 305) and even if game theorists consider that this epistemological question is more and more important, the link between empathy (understood in its epistemic dimension, i.e. as the attribution of mental states to the other) and strategic knowledge remains underestimated, and fragile. The fact is that even nowadays it is mainly the 'motivational' aspect of empathy, i.e. linked to other regarding concern, that has driven the interest of behavioral game theorists and neuro-economists. And the inclusion of empathy into the realm of economics and game theory has been mainly along the line of pro-social behavior, altruistic preferences and other regarding motives. This situation is mainly explained, as Sugden (2002) argues, by the RCT paradigm. The latter leads empathetic concern to be integrated into the individual preferences or utility function. This has ultimately left no room, apparently, for considering empathy as a means to form beliefs towards others' mental states (beliefs, intentions and preferences) and choices.

The rare mentions of empathy considered in this section show how there are still improvements to be made in the broad understanding of this concept and on the enhancement that this concept can bring into game theory. This explains why we intend to give a detailed presentation of the ToM and of the knowledge that can be imputed to the players from the ToM with respect to the others' preferences, beliefs or intentions for instance.

2.1. Binmore's tentative to bring empathy in the realm of game theory

Binmore (1994), is one of the few, if not the only one amongst acknowledged game theorists, to early value the importance of considering 'empathetic identification' in interactions (Fontaine,

1997; Sugden, 2002; Kirman and Teschl, 2010). This does not acknowledge, as we will argue (in the section 2.3.) that Schelling (1980[1960]) long before Binmore (1994) and Bacharach (1991, 1993, 1997, 2006) without referring explicitly to the term empathy, both saw its pivotal role for strategic interactions. They both have actually, a very contemporary view of empathy that is in accordance with its modern account as developed by the ST in ToM.

Binmore, who strongly attacks Bayesian rationality in game theory (see chapter 1), militates for the introduction of empathy in the realm of economic theory and advocates for putting empathy at the center stage of economic decisions (Binmore, 1994). He defines empathy as "feeling oneself to be in the other one's place." (Binmore, 1994, p. 56) Empathy is for him a process that "stops short of the point where we supposedly cease to separate our interests from those with whom we identify." (Binmore, 1994, p. 56) At this point this becomes sympathetic identification. More specifically, when Binmore (1994, p. 51) defines empathy; he declares "I intend empathy to refer to the process through which we imagine ourselves into the shoes of others to see things from their point of view, but without the final step envisaged by Hume and Adam Smith. That is to say, the process stops short of the point where we supposedly cease to separate our interests from those with whom we identify. Thus, when Adam empathizes with Eve, the understanding is to be that he puts himself in her position so that he can reason things out from her point of view, but without this having any impact on his personal preferences."

In this definition Binmore therefore assesses empathy in its epistemic dimension and more specifically in what will be presented later as a simulation process, which involves putting oneself in the other's shoes.

However as will be developed below, his account of the role that empathy can play for economics and game theory is confusing. He indeed often refers to empathy when mentioning moral values and for referring to interpersonal comparison of utilities (see for instance Binmore, 1994, p. 65; and Sugden, 2002). Binmore seems indeed much more concerned with empathetic preferences as a means to reintroduce moral questions in economics and more specifically as a means for fairness judgments (see Binmore, 1998, p. 178). Such empathetic preferences are according to him forged by history, i.e. by what he identifies as "the medium-run past" (Binmore, 1994, p. 65); they are "shaped by the forces of social evolution" (Binmore, 1994, p. 65)

In spite of that, his approach is worth mentioning as Binmore belongs to the very restrictive range of game theorists that have a broad conception of game theory and of individual decision-making in interaction. He criticizes the principles of Bayesian game theory as principles of rational decision-making in strategic contexts (see Binmore, 1987, 1990). One of the reasons he puts forward relies on the problems conveyed by "Bayesianism" for players to attribute beliefs to the other players. This entire chapter aims specifically to bring attention to the insights that the ToM can carry on for game theory on this belief ascription. From that perspective, Binmore (1987) conceptualizes what he calls the "open universe problems" which are later taken up by Bacharach in his VFT in a slightly different vein and translated in to "variable universe games". Recall that the variable universe games entail a broadening of standard non-cooperative games through the integration of players' perceptions. The consequences of Open universe problems and variable universe games are that the set of the possible states of world considered in games is no longer finite – contrary to incomplete information and epistemic games, which ultimately for

Binmore threatens the validity of Bayesian rationality. Yet, strangely, Binmore does not link this open universe problem to the use of empathy or mindreading ability, ultimately enabling players, when the definition of the states of world is not a closed system, i.e. finite, to form beliefs on others' reasoning, beliefs and then choices. Therefore, contrary to both Schelling and Bacharach, in Binmore's view empathy does not intervene in the kind of uncertainty involved in strategic context as understood in terms of open universe problems. His conception of empathy seems on the contrary to be integrated into the RCT and the preferences utility conundrum, which ultimately lead to consideration of empathy as a form of altruism or eventually as a form of individual utility comparison as in Harsanyi (1977, 1987). The problem is also that such empathetic preferences must satisfy the von Neumann utility function; they must satisfy its axioms (Binmore, 1998, p. 220), they must be consistent (Binmore, 1994, p. 290).

In fact, Binmore (1994, 1998) uses empathy in his distinction of two types of games; 'the games of life' and 'the games of morals'. In the former each player acts according to her subjective preference, and when an equilibrium exists it is a Nash equilibrium; but in the latter, a problem of equilibrium selection exists which can only be solved by resorting to a convention and in this case players act according to empathetic preferences, in the sense of ethical preferences (Sugden, 2002, pp. 67-68). In games of morals, acting according to empathetic preferences allows players to reach an agreement and accordingly to select one of the multiple equilibria (see also Teschl and Kirman, 2010, p. 304). Binmore therefore uses empathetic preference as a device of equilibrium selection (Sugden, 2002, p. 228).

But in detailing the content of the empathetic preferences of Binmore, Sugden (2002, p. 68) defends the idea that the Paretian and utilitarian turn in rational choice theory necessarily distort empathy into sympathy, as empathetic concern understood in terms of an emotional acquaintance with others must be integrated in player's preferences or utility function. Sugden states the point as follows:

> "In Binmore's account, the distinguishing characteristic of sympathy is that it is registered in the sympathizer's utility function – that is, that Adam's choices are affected by his sympathy for Eve. There is no way of saying that Adam's feelings are affected by his perception of Eve's feelings, without also saying that Adam is motivated to perform actions which benefit Eve. Once the Paretian turn has been taken, this feature of rational choice theory is unavoidable, because feeling, like all other psychological concepts, has been stripped out of the conceptual scheme. But the utilitarian version of rational choice theory faces a similar problem, as a result of its one-dimensional psychology. In the utilitarian scheme, the only way that Adam's feelings can be affected by his perception of Eve's feelings is for him to gain pleasure from his perception of Eve's pleasure. Since individual rationality is understood as the maximization of pleasure, Adam's sympathy for Eve must also be a motive for action for him." (Sugden, 2002, p. 68)

The purpose of this chapter is to show that such restriction of empathy to an emotional dimension and therefore its confusion with sympathy, as induced by the Utilitarian and Paretian turn of RCT, as Sugden suggests is yet avoidable. We should not consider that empathy and sympathy should be confined to the universe of preferences. On the contrary, empathy can intervene in players' knowledge and beliefs, so that the structure of preferences is left unaffected,

but the elicitation of players' beliefs and their ascription of beliefs, intentions, or preferences to the others is explained, contrary to standard and Bayesian rationality in which they remain unexplained (see chapter 1 for an explanation of this claim). However, I concede, this necessitates a considerable amendment of Bayesian rationality. But as already stated in the Chapter 1, the equation between rationality and Bayesianism is far from making a consensus and from ensuring rational choices in strategic contexts (see e.g. Bacharach, 1985; Kadane and Larkey, 1987; Binmore, 1990, 1993, 1994, 1998; Mariotti, 1995, 1996). If Bayesian thinking can explain beliefs updating, in any case it cannot justify the players' prior beliefs as is standardly assumed (again see chapter 1).

In the same way, Kirman and Teschl (2010) argue that empathetic concern should not be confined to a problem of equilibrium selection as we could think in reading Sugden (2002) and they adopt another interpretation of the role assumed by Binmore for empathy. They claim that "[a]s Binmore (1994, p. 289) points out, empathy must not be considered as 'some auxiliary phenomenon to be mentioned only in passing', but rather as something basic to humanity which can enable us to understand the nature of strategic interactions between individuals. Hence, 'Homo economicus must be empathetic to some degree' (Binmore, 1994, p. 28)." (Teschl and Kirman, 2010, pp. 303-304) They point out that Binmore as in Binmore and Shaked (in press a, b) for instance, does not plead for the replacement of the standard economic agent driven by her own interest by an other-regarding economic agent (see Kirman and Teschl, 2010, p. 306). Besides, they add in a footnote that "Binmore's idea of homo economicus is, of course, far from the traditional isolated maximizing individual since he interacts consciously with others. This leads him to contemplate the reactions of others and for this he has to be aware of what their utilities or payoffs are." (Kirman and Teschl, 2010, p. 315) However in reading Binmore (1994, 1998) this assertion is hardly present. Though, again, it is on this dimension that Kirman and Teschl emphasize that empathy can offer the main enhancements of game theory. And again it is on that latter dimension that the remainder of the chapter and of the thesis will focus, offering a model of games encompassing empathy from that perspective. Regarding their reading of Binmore, it seems that his account of empathy is much more epistemic in nature than was early suggested. The integration of empathy in individual preferences as Binmore (1994, 1998) does nevertheless seems really question begging with respect to Kirman and Teschl's interpretation.

2.2. The other-regarding preferences literature

Still in a traditional fashion, the results of experimental games have led to the incorporation of new determinants in players' utility functions or preferences explaining other-regarding behaviors. A variety of models exist, each of them explaining different empirical regularities. In fact, two main ideas are conveyed in other-regarding preferences: altruism and iniquity aversion, which could be some form of empathetic or sympathetic considerations in the sense of being affected by other's well-being. We will only focus on the more popular models within the other-regarding preferences literature to mainly state the extent of the use of empathy in behavioral economics. We will show that such literature remains anchored to a closed system epistemology.

It simply rationalizes behavioral data but without thinking on empathetic considerations and their meaning in the other social sciences. Psychology, social psychology or cognitive sciences are never mentioned as deepening the knowledge of the empathy concept and its meaning for strategic interactions in games. Such literature does not bring interdisciplinarity in economics. This is therefore different from our account of game theory as an open-science, and of Schelling's and Bacharach's account.

For altruism we can refer to Andreoni's (1989, 1990, 1995) 'Warm Glow Theory' (WGT). He assumes that people reveal in their behavior an "impure altruism", i.e. they consent to give money to the others or to the community in order to increase their own wellbeing. The act of giving – for good reputation – gives them satisfaction. Therefore, in the WGT, people's utility functions depend on the one hand on their private consumption of good, and on the other hand on the amount they give to finance the public good. They arbitrate between their own consumption and financing public goods according to their "altruism coefficient" (everyone has a different "altruism coefficient"). Subsequently, for Andreoni (1989, 1990, 1995) and Andreoni and Miller (2002) altruism is rational in the standard meaning of this term.

A second category of models symbolizes altruistic preferences such as Levine (1998), Rotemberg (2008) and Gul and Pesendorfer (2005, 2016). They model players who are not only concerned by the payoff distribution but mainly by the intention underlying the others' behavior. When evaluating the payoff distribution players judge how altruistic the others players are. Consequently, players' utility functions depend on their own payoff, others' payoff and an altruism parameter (which symbolizes how much they care about the others' payoff). The value of the players' altruism parameter depends on the others' altruistic parameter (evaluated in the same way). In Levine's model how much a player weights the other players' payoff in her utility function depends on her own payoff, the value of the others' altruism parameter – as revealed by her own payoff –, and the value her own altruism parameter. He claims to demonstrate in this particular case that players judge others' intentions through the payoff distribution and in particular through the payoff that they earn according to the others' behavior. In Rotemberg's model (2008) players' altruism parameter depends on the action realized and not only on the gains. Furthermore, unlike Levine he only admits (according to him) "weak altruistic behaviors" because players weight their own payoffs more than the others' one. Another interesting model in this category is that of Gul and Pesendorfer (2005, 2016). Intentions in their formalization are a function of players' personalities, so is their degree of altruism or egoism. Players' personalities are revealed step by step over the course of a game. And this personality depends on other players' personalities, so it depends on the result of players' interactions in each round. Basically, altruism is built in the same way through interaction in each round.

As Kirman and Teschl (2010) emphasize many experiences have attempted to determine the distribution of altruistic preferences in the population. As they claim:

> "If we assume [other-regarding] preferences do exist, then the next important issue for an economist to clarify is the stability and distribution of other-regarding preferences. The question is, are there any particular 'other-regarding preference types' in the population? Some neuro-scientific experiments have been claimed to show that there exists a heterogeneous expression of empathy across experimental subjects (Singer et al. 2006).

> The idea is that each person is endowed with a certain degree of empathy or other-regarding preference. This could therefore be translated into a particular distribution of other-regarding behaviour based on these intrinsic other-regarding preferences, assuming that it is this type of empathy that leads to such behaviour. This is good news for standard economists as they generally assume preferences to be given and stable, and it is on the basis of these that they are then able to construct models that lead to specific behavioural predictions over time." (Kirman and Teschl, 2010, p. 305)

Such stability in the distribution of altruistic preferences is for Kirman and Teschl (ibidem), far from being assured and therefore the incorporation of altruism in determinate utility functions or preferences by an established coefficient is very doubtful (see Hichri and Kirman, 2007).

Within the same constraints numerous models have flourished to formalize iniquity aversion. Those widely cited are Fehr and Schmidt (1999), Bolton and Ockenfels's ERC (Equity, Reciprocity and Competition) model (2000), and Charness and Rabin (2002). These models differ according to the type of iniquity aversion the agents' utility function symbolizes and according to how much players are concerned by the others' payoff, i.e. by the weight of their other-regarding inclination.

In Fehr and Schmidt's model each player compares her payoff to every other individuals payoff, twice by twice, so each player is only preoccupied by the iniquity between them and the others. Players have to know every reward in the reference group – so that preferences and utility functions can be defined. In the ERC model, players compare their own payoff to their relative ones; they assess if the distribution of the individual payoffs exhibit iniquity by comparing their own gains to the average gain of everyone else. Another difference between both models is the fact that, in Fehr and Schmidt's model players can weight their own gains and others in the same proportion, whereas in the ERC model players weigh their own gains more than the others. In Charness and Rabin's model players are at the same time altruistic and iniquity averse, and this is formalized by 'quasi maximin preferences'. When players evaluate the payoffs distribution, they simultaneously manifest concerns for the total surplus – i.e. the overall gain – and for the player who receives the lowest payoff. In comparing each individual gain they over-weight the less lucky player in their utility function. Nonetheless, like the ERC model players prefer receiving more than less all things considered. They prefer winning for themselves more than for the others until a certain iniquity level.

In summary, the other regarding preferences literature is characterized by the adding of an ad hoc parameter in individual utility to rationalize the empirical regularities. As integrated in utility function, the other regarding inclination takes the form of sympathetic concerns but without any serious psychological or sociological foundations, i.e. without any references to psychological or sociological theories. As argued at the beginning of this section such literature is merely an ad hoc rationalization of behavioral regularities that does not rely on interdisciplinarity. An ad-hoc parameter in players' preferences or utility has simply been added. It means that in this account game theory remains a closed system. Thus, ultimately, it does not provide fertile thinking on the broad role that empathy can play in game theory. Besides, the other regarding preferences (ORP) cannot explain why players effectively coordinate in coordination game as the experimental

results show. As Lecouteux (2018a) shows, ORP also lead to puzzling results in games like the Prisoner's dilemma.

2.3. The Schelling-Bacharach's perspective

On the contrary, Schelling and Bacharach adopt a completely different perspective. Their respective approaches make room for empathy and mindreading: the ability to attribute to other peoples some mental states like beliefs, desires or intentions; and in fact, their account is very much inclined to the cognitive dimension of empathy in the modern sense of the ToM, and in particular of the ST approach.

Schelling (1980[1960]) frequently refers to the importance in strategic interactions of seeing the decision problem situation from the other's angle. This requires putting oneself in the other's shoes. This is not a matter of guessing what the average man generally thinks or want, but precisely what the man in front of you or with whom you interact wants. The meeting of minds, i.e. the convergence of players' perceptions, intentions and beliefs is ultimately ensured only in this way. Recall that in standard and epistemic game theory such a device is unnecessary as common knowledge of rationality or common beliefs entail that players knows or possess defined beliefs on others' knowledge, on what they want, and ultimately on their choices. Since in Schelling's view the formation of players' beliefs on the others' perceptions, intentions and beliefs is endogenous to the game, their ability to see the problem from the other's perspective becomes the main determinant of their capacity to coordinate. Besides, since players are heterogeneous (in their perceptions and mode of reasoning), they are ultimately able to form beliefs about the others only through this cognitive process, i.e. only through their ability to put themselves in the others' shoes. Resorting to social and exogenous focal points, as we have insisted in the section 2 of chapter 2 and contrary to the widespread understanding of this concept in Schelling's contribution, does not prevent this empathetic identification, which is again, the pivotal determinant of players' capacity to successfully coordinate. In order for players to be confident that everybody will follow the pattern of behavior of a social – and exogenous – focal point, they have to put themselves in the others' shoes. This depends on the uncertainty of the situation that the players face. Thus, empathetic identification is not restricted to cases where no exogenous coordination device exists. In any case it is the main determinant of players' meetings of minds. It is in the very nature of a strategic interaction, i.e. in the interdependence of choices, that empathy matter.

In the same way, in Bacharach's VFT and then TR theory, empathy is a matter of acquaintance with the other's perceptions – i.e. frames – and beliefs. According to Bacharach's methodology players have to account for others' perceptions and beliefs to guess what is the others' salient option; so that they can compare the different individual saliences and ultimately compute if saliences are shared or not. If saliences are not shared there will be no coordination. Accordingly, players must have some knowledge of the others' perceptions and then of their beliefs. As for Schelling, since players are heterogeneous – because they can have different perceptions and subsequently beliefs – in order to be acquainted with the others' frames and beliefs players put

themselves in the others' shoes. From a formal point of view, Bacharach translates this by a principle stating that players attribute to the others a subset of their own frames and beliefs. According to their own frames and to the set of beliefs they handle within these frames, players attribute to the others a part of their own frames and beliefs. As we will see this methodological device to translate empathetic consideration in formal terms is very close to a cognitive mechanism identified in the ST as the egocentric bias, which is a widespread phenomenon in the third person attribution of mental states – i.e. the attribution of mental states to someone else other than ourselves: your co-player for instance (cf. section 4.3). Thus, in Bacharach's view, like for Schelling, empathy is a matter of information gathering for coordinating, nothing more. And like for Schelling, in this cognitive process players can rely on a kind of social background. This assumption seems far from standard and epistemic game theory as it implies (i) that games are no longer closed-systems but instead that some determinants of players' choice are determined by external elements to the games (i.e. to other elements than the rules of the games) and (ii) that common knowledge or common belief in rationality is insufficient for players to anticipate each other's choice. Such social background ultimately induces the sharing of some frames and beliefs towards these frames. We will see that such assumptions rely on mindshaping devices, i.e. social practices that ultimately bring homogeneity among people and that will be detailed in the section 5.2 of this chapter. Players can rely on this background in their reasoning process as it raises the probability of successful coordination. But again, the confidence that the players can have on the convergence of their frames and beliefs – through social coordination devices – is generally sustained by empathetic identification.

Therefore, even considering the differences existing in Schelling and Bacharach in their integration of empathy within game theory and its methodological and theoretic translation, none of their views entail other regarding motives and the need to break with individual's personal interest. Quite the contrary, in order for players to reach their goal they have to empathetically identify with the other. It is a matter of information gathering, a pivotal dimension in strategic interaction and on that point, none of the game theorists could disagree. Empathetic identification allows players for both Schelling and Bacharach to integrate in their strategic reasoning the eventual perceptions, beliefs and in turn choices that the others may eventually make. Empathy is grasped in its epistemic dimension. And by no means is confusion between empathy and sympathy is at stake; by no means is the frontier between oneself and the other blured.

It is now time to turn to the insights that the ToM can bring into game theory. The reminder of this chapter investigates the problem of apprehending others, i.e., on what kind and what form of knowledge of the others can players rely, and on what extent of knowledge of the others can the players ground their reasoning when they have to make strategic decisions. In other words we propose to scrutinize the part of the players' epistemic states, thanks to the ToM, about the other players, their belief, reasoning and choices. This attempt is accordingly in the line of both Schelling's and Bacharach's account of game theory and rationality. It adds to their view a philosophical and cognitive underpinning to (i) justify their original approach and (ii) extend their work. It also sets, from that perspective, the formal contribution that can be made using the ToM and the ST in particular (see the next chapter). The extent of the players' epistemic states concerning the other players supposed by Schelling and Bacharach is in fact supported by the

ToM approaches, however they do not elaborate so much on the psychological and cognitive basis of these specific statements (for Bacharach mainly since Schelling does not make a stance on the psychological theory on which he draws).

3. The cognitive approach of mindreading and the rise of the Theory-Theory (TT)

3.1. The premises of the TT

3.1.1. The philosophy of mind and common sense psychology

The 1950s established the roots of the "modern" history of the ToM and mark the entrance within the landscape of the philosophy of mind of what became later the TT approach of mindreading. Within the Philosophy of Mind, the tennets of common sense psychology, which is the underpinning of the TT approach, had controlled the field for almost 30 years before cognitive sciences burst on the stage, in the mid 1990s – and modified the established order (Goldman, 2006, p. 4; Zahavi, 2014, p. 97; Gallese, 2003a, p. 519). This philosophy of mind has been strongly influenced by two epistemological trends: logical behaviorism and functionalism. Consequently, during this early stage of the philosophy of mind, mental states: (i) cannot be appraised without behavioral data and (ii) they are conceptual, addressed by a quasi-scientific or a commonsense psychology, and they simply play a causal role for behaviors and utterances. The TT account of mindreading is strongly influenced by these trends which, from our point of view, as we will see, involves very profound limits to its use in game theory.

First, appraising mind requires a behavioral insight. Mental states are merely dispositions to act. They can only be grasped by their physical manifestation, i.e. behaviors or utterances. This echoes the behavioral interpretation of payoffs and utilities in games which, as exposed in chapter 1, leads to considerable methodological difficulties. Mental states are "commonsense concepts": identified and defined through a conceptual analysis (Goldman, 2006, p. 6). In other words mental states remain theoretical concepts appraised by a common sense psychology as manipulated by the everyday man, and not necessarily a scientific psychology. Some debates however exists among the early tennets of the TT in regards to the epistemic accuracy or not of this common sense psychology (e.g. on this position see Ryle vs. Heider); (Maraffa, 2015). Referring to these mental states "perceivers bring order and meaning to the massive stream of behavioral data." (Malle, 2011, p. 73) Thus, common sense psychology allows people to infer others' mental states such as wishes, intentions, beliefs, sentiments etc. from overt behaviors.

Second, mental states merely play a causal role for behaviors and utterances. This is the position defended in particular by Lewis (1966) and Armstrong (1968). For both of them mental states relate to environmental stimuli, internal states and behaviors (Goldman, 2006, pp. 5-6). The

child-scientist theory within the TT account has been highly influenced by this functionalist premise (Goldman, 2006, p. 7).

Mental states therefore became defined by theoretical states established by a "commonsense psychological theory", or by a "folk psychology", which sets lawful relationships between unobservable and observable states and between the unobservable states themselves (see Ramsey, 1931; Sellars, 1997[1955]; Lewis 1980[1972]). For instance, Lewis (1966, 1972) defines three types of psychological laws: (i) one for linking observable inputs from mental states, (ii) one for linking the different mental states, and (iii) one for linking mental states to observable outputs such as behavior and utterance. The way Lewis explains how people come to be acquainted with mental states is as follow

> Think of commonsense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. … Include only platitudes which are common knowledge among us – everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that the names of mental states derive their meaning from these platitudes. (Lewis, 1980[1972], p. 212)

The 20th century witnesses the rise of behaviorist approaches and the purpose of the early experimental psychology to provide objective measures of mind. The emergence of cognitive science taking place after the midst- 20th century has been largely influenced by the behaviorist approach of mind, and the analytic philosophy of mind embraced the naturalistic turn (Gallagher and Zahavi, 2012 [2008], p. 4). This progressively opens the way to the modern TT and a bulk of paradigmatic experiments attempting to show the existence of a theory of mind to seal the matter.

3.1.2. The "false belief task": the paradigmatic experiment setting the cognitive turn in mindreading

Premack and Woodruff's (1978) paper "Does the Chimpanzee Have a Theory of Mind?" on the set of experiments they conducted on a female chimpanzee marks the birth of the modern TT. Showing the chimpanzee a video of an individual involved in a problem-solving task, they test her understanding of the scene in her interpretation of purposeful behaviors. The tests revealed that the chimpanzee was in fact able to understand the mental states behind the behavior, to interpret and predict human behavior through these mental states.

Premack and Woodruff interpreted these results as proof that the chimpanzee possessed a "theory of mind." They explain why they refer to a theory of mind as follows

> "A system of inferences of [mental states] is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to

> make predictions, specifically about the behavior of other organisms." (Premack and Woodruff, 1978, p. 515)

They offered two possible interpretations for the chimpanzee's understanding: (i) this is classical associationism which would state that the behavior of the chimpanzee does not rely on mentalization but simply on mimicry and association, and (ii) the "empathy theory" (Goldman, 2006, p. 11). They maintain as an explanation of the chimpanzee's response, instead of mimicry, that she holds a theory of mind: they maintain the "empathy theory". They define the "empathy theory" as follows:

> "The empathy view starts by assuming that the animal imputes a purpose to the human actor, indeed understands the actor's predicament by imputing a purpose to him. The animal takes over the actor's purpose, as it were, and makes a choice in keeping with that assumed purpose. The empathy view diverges only in that it does not grant the animal any inferences about another's knowledge; it is a theory of mind restricted to purpose." (Premack and Woodruff, 1978, p. 518)

The second stage of the experimental template pinning down the TT account is explained by the comments brought about by Premack and Woodruff's article. Philosophers such as Dennett (1978a), Bennett (1978), and Harman (1978) suggested making a further step towards the establishment of the theoretical content of mindreading (Goldman, 2006, p. 11). They propose to assess the ability of the chimpanzee to possess a belief concept and for that, they independently propose to set an experimental design testing the attribution of false beliefs (e.g. see Harman, 1978, pp. 576–577)

The developmental psychologists Wimmer and Perner (1983) first conducted the new type of experiment proposed in the aftermath of Premack and Woodruff's one. This lead to the groundbreaking experimental paradigm institutionalizing the TT approach: the "false-belief task".

In the paradigmatic version of this task, a child watches a room (or a tape) in which there are two dolls (Sally and Anne). Sally puts a toy in a location A and leaves the room. While Sally is out of the scene, Anne takes the toy and puts it in another location, B. The experimenter then asks to the child where will Sally look for the toy (i.e. in the location A or in the location B?). Almost all the results – despite some variations among children – show that while the four and five years old children tend to pass the test and answer that Sally will look in the location A – where she left the toy, the three years old children fail and answer that Sally will look in the location B – where the toy has been placed by Anne. Therefore it is claimed that after approximately 5 years old children succeed in the false belief task, while before 5 years old the children fail in the false belief task.

To explain the results of the 'false belief task' experiment, Wimmer and Perner (1983, p. 126) postulate that "a novel cognitive skill seems to emerge within the period of 4 to 6 years. Children acquire the ability to represent wrong beliefs". Thus, they can distance themselves from reality and observation to make inference on someone else's behavior. They can form counterfactual reasoning. This level of abstraction is a primary form of theorizing. It means that children can make inferences on the other children's perception (i.e. a mental state). It thus reveals that they possess a (rudimentary) theory of mind.

Wimmer and Perner's experiments initiated a stream of experiments in developmental psychology testing the infant understanding of the mind which exactly replicate the experimental design mentioned. And the conclusion widely held from the findings of this 'false belief task', is that before five years old children do not possess a full understanding of mental states. They cannot fully dissociate reality from representation and perceptions. Even if they dissociate reality from dreams, desires, and imagination, they do not have "a fully representational model of mental states." (Goldman, 2006, p. 12)

From that moment, two frames of explanation emerge for the early infant "conceptual deficit" (Astington and Gopnik, 1988):

i) The modularistic explanation, stating that 'mentalistic skills' result from the progressive maturation of an innate "theory of mind module" or "theory of mind system of modules" (e.g. Baron-Cohen, Fodor, and Leslie) and for which the inferences driven by the ToM rely on automatic and unconscious processes.
ii) The Child-scientist explanation, against the innate postulate, considering that the naïve ToM that infants possess is developed in a comparative way to scientific theories (e.g. Gopnik, Perner, Wellman), and for which children have an evolutive ToM which is progressively built and enriched.

This state of the art in the late 1980s sets the background of the TT. Since then, various disciplines such as primatology, developmental psychology, cognitive neuropsychiatry and philosophy (based on empirical support) collaborated to build the TT. They settled the extent of study of the TT approach of mentalizing. Each of them comply either with the modularist view or with the child-scientist view which are the two approaches nurturing the TT account of mindreading.

This explains why the landscape of the theory of mind was dominated until the end of the 1980s by the functionalist and the TT and RT approaches of mind reading. Folk psychology was still dominant in philosophy and developmental psychology. However, the contributions of Gordon (1986), Heal (1986) and Goldman (1989) mark the entrance of the ST onto the scene in the late 1980s. Since then, the ST has progressively reached a critical mass and the debate between the two major stances TT and ST has never faded. The double issue of Mind and Language in 1992 sets this debate. Philosophers like Stich and Nichols, and developmental psychologists like Perner, Gopnik and Wellman defended the TT account against Gordon, Goldman and the developmentalist Harris who supported the ST. More recently, the discovering of mirror neurons by Rizzolatti, Gallese et al. in 1996 sheds a new light on the simulation mechanism reinforcing its scope within the ToM, and strengthening the link between neuroscience and the theory of mind.

3.2. The Theory-Theory paradigm (TT)

The mental concepts in the TT paradigm are described by causal and functional links between external stimuli, behavioral or utterance output, and other mental states. The tree main theses

defended in the TT paradigm are the following: (i) mental states are assessed by causal laws that establish the relationship between external events (the stimuli input and the behavior or utterance output) and other mental states, (ii) the attribution of mental states (in first or third person mindreading) relies on "law guided inferences" from the observation of external events (again stimuli or behaviors), and (iii) these laws are acquired by experience, they are empirically driven "by means of general-purpose scientizing procedures." (Goldman, 2006, p. 27)

It is on the latter point that the two TT schools differ. It is in particular on the way the children acquire and revise their theory of mental states that the sub-approaches of the TT differ. The child-scientist theory theorists argue that the way children revise their ToM by revising their theory in light of new evidence is similar to a scientific procedure, while the modularist theory theorists postulate that some modules are progressively activated according to the incremental complexity of the social interactions involved. For the latter it is not so much the possession of a theory that explain the inferences driven by the mental state attribution, but actually the deployment of such theory.

However, in each case, understanding oneself and others requires the existence of a "folk theory" of human psychology, namely, the existence of a 'naïve' theory of our own and other's mind functioning (Goldman 2012, p. 3). In the TT, mental states are "theoretical states of a common sense psychological theory, (Goldman 2006, p. 7). Mental states are appraised and defined by theories. Any reference to a mental state therefore requires a theoretical inference. These theories and theoretical inferences are incrementally refined with experiences of live (Goldman, 2006, 86)

The sticking point between modularist and child scientist proponents is not on the existence or otherwise of a folk psychology but rather on its acquisition and deployment, i.e. on the innate vs. acquired.

3.2.1. The modularist theory

Fodor (1975, 1983, 1987), then Leslie (1987, 1994, 1997, 2000), and Baron-Cohen (1989, 1995, 1999) are the forerunners of the modularist theory. Fodor sets its theoretical formulation and establishes the theoretical inference explanation of the common sense psychology on which Leslie and Baron-Cohen ground their work. Thereupon they respectively enrich this theoretic background and refine it. They specify the functioning of the modules responsible for the mentalizing and attribution capacities, and their categorization. Leslie in particular lays the ground for empirical research.

Fodor (1987, p. 1) indeed characterizes folk psychology as an "implicit, non-demonstrative, theoretical inference." For him

> "When such [commonsense psychological] explanations are made explicit, they are frequently seen to exhibit the "deductive structure" that is so characteristic of explanation in real science. There are two parts to this: the theory's underlying generalizations are defined over unobservables, and they lead to its predictions by iterating and interacting rather than by being directly instantiated." (Fodor, 1987, p. 7)

This common sense psychology is an implicit and 'science like' theory. From that perspective, Fodor considers that people attribute mental states from the inferences they make from their common sense psychology (Goldman, 2006, p. 96)

Much of the developments of modularist theory are based on experiments scrutinizing the link between autism and mind-reading deficits (e.g. see Baron-Cohen, Leslie and Frith, 1985, 1986; Goldman, 2006, pp. 13-14). These experiments reveal that autism damages a "domain specific capacity": the mentalizing capacity but not the mechanical one (Goldman, 2006, p. 13). While autistic children do not have an impaired motor capacity and can reproduce observed movements and behaviors, they cannot mentalize upon these observed behaviors and infer the goal, purpose, desires or beliefs behind them. The modularist theorists accordingly conclude that a specific module is devoted to mentalizing. This module, which is impaired in autism, is the ToMM (Theory of Mind Mechanism): a "domain specific mentalizing module or mechanism" (Leslie, 1987; Goldman, 2006, p. 15)

The ToMM is for Leslie (1987) an innate mechanism that matures just before two years old and on which the acquisition of a theory of mind is based. This mechanism "employs a proprietary representational system that describes propositional attitudes" (Goldman, 2006, pp. 101-102). Recall that propositional attitudes refers to the mental states held by agent A toward a proposition p, such as A believes that p, A desires p, etc. Thus, the ToMM provides the ability to represent propositional mental states, such as 'pretending, thinking, knowing, believing, imagining, dreaming, and guessing' (ibid, p. 17). Leslie, and the modularists following Leslie (e.g. Baron-Cohen), maintain that the ToMM establishes meta-representation for mental states. It generates "representations of propositional attitudes of the form Agent-Attitude-Proposition" (ibidem). These meta-representations explicate and link information concerning an agent, her attitude in relation to a situation – an 'anchor' – and pretense, i.e. a counterfactual situation (ibid, p. 107).

For understanding the specificity, but also the limitations of the ToMM, it is of prime importance to distinguish the representational properties of a mentalizing mechanism – i.e. the ability to represent a mental state – from its attributional capacity – i.e. the capacity allowing an individual to attribute to someone else some specific mental states and a specific content to these mental states such as 'Anne believes that p'. It is of prime importance to distinguish the ability to state that a belief and an intention can together lead to a behavior from the ability to believe that if someone else believes that p and has the intention to do m she will perform the action m (Goldman, 2006, p. 108). It is indeed not the same thing to generate representations of some mental states as to generate beliefs on these representations, like 'I believe that the other believes that I am risk averse'. From that view, the fact that the ToMM generates meta-representations does not grant that it also generates beliefs about these representations of mental states. This is thus of little help to explain the set of beliefs that a player holds in games with respect to the other players' beliefs for instance.

Leslie in fact argues that the belief fixation – i.e. I believe that you believe p for instance – is not accomplished by a modular mechanism but by the 'nonmodular central system' (see Scholl and Leslie, 1999, p. 147). The mechanism of attribution in mindreading, i.e. the formation of beliefs on someone else's mental states, therefore relies on a 'nonmodular activity'. The explanation of

this mechanism must thus be found outside this theory. To sum up, the ToM remain an issue for the modularists, no strong modularist justification of the empathetic capacity has emerged (Goldman, 2006, p. 109). The ToMM has no attributional property, which is obviously the prime determinant for empathetic identification. But more importantly it is therefore of no help to define in game theory what a player believes regarding the other player's beliefs. It is of no help for a player to anticipate the other player's choice.

3.2.2. The Child-Scientist theory

We do not intend to detail the debates prevailing in the child-scientist view of the TT over the period of acquisition of the mentalizing and attributional capacities but we will mainly point out the claims made towards these inferential capacities in third person attribution and the extent of the use of the common sense psychology involved. For our purpose, the child-scientist approach is interesting only in respect of its conception of the acquisition of folk principles as a progressive construction and revision of a theory, pretty much in the same way as scientific theory is built and enhanced.

Child-scientist theorists take the results of the false belief task as the evidence that children undergo a conceptual evolution. More specifically: they undergo a progressive understanding of mental concepts as their environment and informational basis grows and becomes more complex. As their understanding of mental concepts develops, so does their theory of the mental states. That is why scholars take as granted that mental states must be theoretical concepts and that the attribution to the self and others of these states is inferential, i.e. executed by a theoretical reasoning (see Gopnik, 1993).

In particular, the success of the false belief task reveals for these scholars a transition from a 'nonrepresentational' to a 'representational' conception of belief because the concept of misrepresentation is finally handled (see Goldman, 2006, pp. 70-71). They cannot understand the concept of belief because they cannot understand representational concepts: "they cannot represent that something is a representation." (Perner, 1991, p. 186)

Additional studies extend the investigations of children's conceptual change to mental states other than beliefs, such as desires. Again, for these scholars, these new studies show an evolution in children's understanding of the concept of desire (Perner, 1991; Bartsch and Wellman, 1995; Repatcholi and Gopnik, 1997).

In addition to the conceptual changes – regarding beliefs or desires – enhanced in infant development, child-scientist theorists focus on the types of laws that lead to the infant's knowledge and to the infant understanding of the knowledge concept. They generally consider that for children seeing something implies knowing something; this is what they call a "seeing = knowing rule." (Goldman, 2006, p. 81; see Wimmer, Hogrefe, and Perner, 1988; Perner, 1991) For developmental psychologists this rule reveals a lack of inferential power by the lack of a psychological law relating behavioral data or stimuli to different mental states. This rule of knowledge acquisition is very problematic for game theory and incompatible with the level of uncertainty generally involved in games. In one shot games in incomplete information the rule

seeing = knowing is inapplicable as nothing can be observed (there is no past behavior). This is therefore of no help in defining the set of beliefs that the players can hold with respect to the other players' beliefs and choices. Such a rule is too restrictive to be applied to game theory, as it can be relevant only if there is repetition: if players can observe the others' behavior. The acquisition of the nonrepresentational conception of mental states is problematic. It requires many observations to be sufficiently refined in order to be operative, i.e. to be accurate. It necessitates a considerable amount of time, of energy, of cognitive and computational resources… It necessitates a considerable repetition of interactions. This is therefore specific to repeated games potentially played infinitely.

The last statement made by the child-scientist theory is that the way children develop and revise their theory of mind is accomplished by a "general-purpose causal reasoning" like for scientific theories (Goldman, 2006, p. 83). This causal learning is a proxy of Bayes reasoning, called the "Bayes nets". This system allows children to progressively build a set of causal maps (see Pearl, 2000; Spirtes, Glymour and Scheines, 2001; Gopnik, Sobel, Schulz, and Glymour, 2001; Gopnik and Glymour, 2002; Gopnik et al., 2004). Bayes nets are sorts of graphs establishing causal relationships between variables with probabilistic measurements to come to predictions (Goldman, 2006, p. 84)

Bayes net representations are used for both observable and unobservable variables such as mental states. These representations in particular allow children to learn and predict unobservable mental states from 'observable data' such as observable contingencies and correlations. This helps them to build new "causal representations" from new observations of interaction, of correlation, etc. (Goldman, 2006, p. 84; referring to Gopnik et al., 2004). Obviously a serious objection can be raised. Do children really think as Bayesian statisticians i their early days? The most important problem is however the following. If this rule of learning, the "Bayes net", can be accepted to explain how players revise their nonrepresentational theory of mental states, and lets say more specifically their beliefs towards someone else's mental states, this is of no help in explaining their initial beliefs before revision, this is of no help in explaining where priors come from and what their content is. Recall that the main difficulty of epistemic game theory, as shown in chapter 1, is to explain the prior beliefs that the players handle. As Singer and Fehr (2005, p. 344) accurately point out, the players' prior probability distribution over the other players' type, i.e. their prior beliefs regarding others' type is "a huge black box". Bayes rationality cannot justify the set of individual prior beliefs. Explaining these prior beliefs is one of the main purposes of the model proposed in the chapter 5. The model indeed proposes a theory of the formation of these prior beliefs.

Therefore, in both the modularist and the child scientist theories the mechanism of mental states attribution remains ill defined. This is this mechanism that is of prime importance for game theory. We have in game theory first to incorporate a cognitive device enabling player to form beliefs mainly toward the other players' beliefs and choices, second this cognitive mechanism must allow us to say something on the content of the beliefs that the players attribute to the others, i.e. on the specific beliefs they believe the others hold. For example players must be able to believe in a prisoner dilemma if the other believes that she is a cooperator or a defector. In both instances, if they provide an explanation of the concept of mental states and of the way individuals can represent and acquire a representational system enabling them to understand such

concepts and the role they play, it remains far from clear that the mechanism of attribution is also solved and even more that a specific content to the other's mental states can be attributed. General law hardly cover this dimension of mindreading.

3.3. A representation of the mechanism of attribution according to the TT

To understand how the mechanism of third person mental state attribution as developed by the TT paradigm should operate let us refer to two schemas.

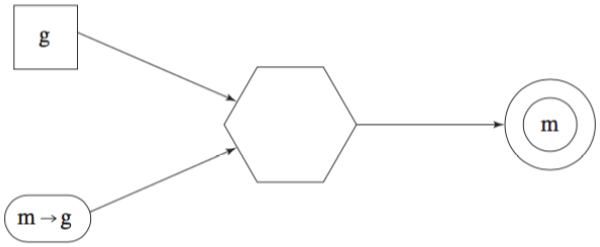


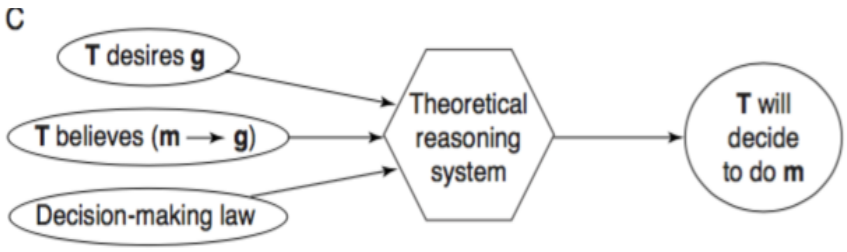
Figure 2.1. Decision (by target) to do m. (Adapted from Gallese and Goldman, 1998, with permission from Elsevier.)

The first schema describes the target's decision process. It represents both the mental states and the cognitive mechanisms (or operations) involved in the decision and behavior of the target.

The ovals represent the beliefs, the squares the desires. In this example, the target has a 'desire' for realizing the goal g, and believes that the action m will allow her to accomplish that purpose. The hexagon represents the decision-making system, and the double circle a decision. Thus, when these beliefs and desires are inputted into the target's decision-making process, the output represented by the double circle is a decision to perform the action m. The general laws supposed by the TT account regarding decision making and behavior work in this way: an interaction between a belief and a desire imputed into a decision making system lead to an action.

Now, how could an observer predict the decision of the target?

The way the observer predicts the decision of the target according to the TT, i.e. the 'theory-based inference', is described by the schema below. As for the previous schema ovals symbolize beliefs, the hexagon a decision-making mechanism and a circle a decision.



In this schema, this time the ovals represent the beliefs that the attributor – i.e. the observer – has on the target's mental states, i.e. her desires and beliefs. Another kind of belief that was not present in the previous schema enters the game. It concerns the attributor's beliefs in the folk psychological laws and it plays the central role for the TT account of third person attribution. It implies that the attributor believes in a "decision-making law" linking the mental states themselves and the mental states and the behavioral output. The 'decision-making law' holds as follow:

> "Whenever an agent wants a certain outcome (more than any competing outcome) and believes that a certain action is the best means to that outcome, the agent decides to perform that action." (Goldman, 2006, p. 28)

Set in these words, it implies that the decision-making law is a form of optimization rule. You perform the action that you believe satisfies your interest the most. This decision making law can take the form of an instrumental reasoning, a profit maximizing rule, a risk averse decision rule, a satisficing rule, etc. The type of decision-making law that is supposed can be ambiguous. It can be learned through experience and therefore if individuals meet other individuals that generally adopt a best reply mode of behavior they will tend to suppose that such decision making law is of this sort. The three types of beliefs that the attributor has, i.e. the beliefs she has regarding the target's belief, desire and decision making law, inputs into her reasoning process, and more precisely her "factual reasoning mechanism". Why this 'factual reasoning' label? Because contrary to the previous schema beliefs are at the same time the input and the output of the reasoning process. The output of the reasoning process is not an effective decision but a belief about the target's decision.

The general law entails that nothing in principle comes to disrupt the mechanism.

As presented in this section a question however remains; how is the observer able, in the first instance, to attribute to the specific target in front of her the specific desire to do g and the belief that doing m will lead to g? As soon as beliefs and desires are known the commonsense psychological theoretical inference naturally lead to the prediction of a decision. However the very fundamentals of such mental states attribution, i.e. the attribution of the inputs of the target's decision process, are left undefined in the TT. The main difficulty of the TT account is to explain this process of attribution and to define content in the mental states that are attributed to the targets. If the observer is told the initial mental states of the target, if she knows the target, or if they have discussed together, etc. she can effectively run a theoretical inference concerning the target's decision. However in the other cases how she came to be acquainted with these mental states is quite uncertain . Besides, beliefs and desires are not the type of mental states that can be translated by facial or bodily expression. In this case, how can the TT help to solve the problem of second order beliefs (beliefs on the others' beliefs) in game theory? How can the TT provide a methodological device to define the set of second order beliefs that a player can hold? The extent of the set of beliefs that a player can hold has to be found outside the TT. Since defining that the prior beliefs that the player holds cannot be sustained by Bayesian rationality (see chapter 1), we

are tied to a mechanism of attribution setting the players' possible beliefs concerning the others' mental states, such as perceptions, intentions and beliefs.

In addition to this serious shortcoming, why not rely on the TT account for enhancing GT? For a few reasons. The TT account of mind reading presents some difficulties when applied to game theory. First it outpaces individual cognitive capacities. In fact, as we can understand through this short presentation, it requires a heavy cognitive load. In order to explain and predict someone's behavior you have first to possess knowledge of human psychology and cognition (even rudimentary) and second to build a theory of her mind. You have accordingly to regularly observe her, in order to make generalizations. How could this be possible in one-shot coordination games when you do not know your co-players? How is this possible in incomplete information games when you have uncertainty about who the other player is (e.g a risk averse, a best reply reasoners, an optimistic player etc.), on her preferences, her set of strategies? Etc. What do you do when you can rely solely on your beliefs because the level of uncertainty is so high that you cannot rely on your knowledge of a general decision making law, as the elements on which you can rely are so thin that you cannot define your target's mental states to input such a decision making law.

Translated into game theory this raises serious difficulties. Because of the eventual differences in perceptions, cognitive lengths, and epistemic conditions among players, the TT account can mean that some players fail to build a theory of the other's mind. Besides, a players' revision of beliefs on the others' mental states, such as their beliefs, is much more difficult. To revise her beliefs a player needs to revise her theory of the others mind functioning. This revision can impact many dimensions: it can be the mental states and the interaction between the different mental states that are erroneous, the causal role they play, or the general decision-making law, etc. Revising the beliefs would require building another general decision-making law. Again this process requires high cognitive resources, time, new experiences, new inductive inferences, etc.

3.4. The Rationality theory

The RT is considered as a specific instance of the TT. The common sense psychology, i.e. the general decision-making law, assumed is rationality. It is supposed that each individual is rational and therefore that her mental states and behaviors or utterances are the ones supposed to be handled and performed by a rational man.

There exist two versions of the RT: (i) a strong version and (ii) a weak version. The strong version of the RT entails strong assumptions toward the 'logicality', 'consistency, and 'truthiness' of propositional attitudes in general and beliefs. The target is interpreted as an agent conforming to all the normative principles of rational reasoning and choices, i.e., the axioms of logical omniscience, deductive closure and logical consistency (Goldman, 2006, p. 54). Recall that deductive closure means that if someone knows the proposition P she can deduce all the logical implications of knowing P; and if Q is logically implied by P then she knows Q (Brueckner, 1998). Logical omniscience entails that people know all the logical consequences of what they

know. "If '$\Phi \rightarrow \Psi$' is valid so is '$\Box \Phi \rightarrow \Box \Psi$'" (Stalnaker 1991, p. 425). Many scholars attack this assumption for being unrealistic; that is why logical omniscience is often interpreted as a mere idealization (ibidem). And finally, logical consistency regarding a set of propositions implies that none of the propositions contradict each other, they must all be true at the same time. The weak version relaxes some of these assumptions and accepts a weaker version of rationality, i.e. a "minimal rationality" (e.g. Cherniak, 1986; Stein, 1996).

Dennett, the most famous forerunner of the RT with his Intentional-Stance Theory (IST) initially seemed inclined to the strong version of the RT (see Goldman, 2006, p. 54). Though, he progressively admitted that rationality is a normative principle and that people can depart from this "myth" of rationality (ibidem). On the other hand, Cherniak (1986) and Stein (1996) are spokesmen for the weak version of the RT.

Dennett defines his account of the RT: the IST as follow

> "The intentional stance is the strategy of prediction and explanation that attributes beliefs, desires, and other "intentional" states to systems - living and nonliving - and predicts future behavior from what it would be rational for an agent to do, given those beliefs and desires. Any system whose performance can be thus predicted and explained is an intentional system, whatever its innards. The strategy of treating parts of the world as intentional systems is the foundation of "folk psychology"." (Dennett, 1988, p. 495)

Dennett (1987) sees his IST as an explanation of mindreading capacity (Dennett 1987, 1988, 2008; see Goldman, 2006, p. 53). He refers to the capacity of belief attribution – which is mindreading – in his contribution to the RT through the IST. To predict someone else's behavior, he declares that we have to "[c]onsider first how we go about populating each other's heads with beliefs" (Dennett, 1987, p. 17), and he adds: "[d]o people actually use this strategy? Yes, all the time." (Dennett, 1987, p. 21)

Dennett however does not subscribe to either of the TT accounts, i.e. the child scientist theory or the modularist theory. He indeed declares that our folk psychology may be innate or "aided by innate perceptual or dispositional biases" but that it is also "learned from experience" (Dennett, 1988, p. 495). It is hard to believe that we learned from experience that people are rational in the sense of the RCT or of Bayesian rationality when experiments show that most of our behaviors violate these same principles of rationality. Dennett also argues that building a general quasi-scientific law on human decision-making that provides a good anticipation and explanation of other's behavior is quite impossible. It is too complicated so that in the majority of cases it leaves people with no prediction or no explanation of the specific behaviors under the circumstances scrutinized; "it is actually hard to generate a science-fictional scenario so novel, so unlike all other human predicaments, that people are simply unable to imagine how people might behave under those circumstances." (Dennett, 2009, p. 6) In this way he disagrees with Fodor who favors scientific psychology against folk psychology (Cussins, 1988, p. 509).

The elementary premise for mindreading behind Dennett's Intentional Stance is therefore that the target is considered as a rational agent, i.e. at the same time a rational thinker and the bearer of a set of rational mental states. In the IST

> "one treats the system whose behavior is to be predicted as a rational agent; one attributes to the system the beliefs and desires it ought to have, given its place in the world and its purpose, and then predicts that it will act to further its goals in the light of its beliefs. Generally, the beliefs any system ought to have are true beliefs about its circumstances, and the desires any system ought to have are those that directly or instrumentally aim to secure whatever the system needs to preserve itself and further any other projects it has." (Dennett, 1988, p. 496)

Being a rational agent for Dennett (1978b) in particular entails logical omniscience and deductive closure. He indeed declares

> "The assumption that something is an intentional system is the assumption that it is rational; that is, one gets nowhere with the assumption that entity x has beliefs p, q, r . . . unless one also supposes that x believes what follows from p, q, r. . . . So whether or not the [intentional agent] is said to believe the truths of logic, it must be supposed to follow the rules of logic." (Dennett, 1978b, pp. 10-11)

What does it imply for the attribution of beliefs? When an attributor imputes to a target a set of beliefs she also supposes that this target believes all the logical consequences following from her initial set of beliefs. This leads to the problem of the possible infiniteness of beliefs, and as already emphasized it necessitates considerable cognitive resources that mostly bypass human capacities. Beliefs must be consistent each other and they must be true. This premise is hardly compatible with a view of players in which their perceptions and frames are taken into account as in Schelling's and Bacharach's respective work and as is more generally supposed in the thesis. Recall that when frames are integrated in Bacharach's VFT logical omniscience vanishes (see the section 5 of the chapter 3 of the thesis).

Many experiments, among which some are well known in economics, show that agents do not respect these principles of rational decisions and behaviors, so that it seems even more dubious than these very same agents attribute beliefs or other mental states conforming to all of these principles while they do not respect them. Experimenters indeed report that in general people are very much inclined to be incapable of drawing all the logical consequences following from their prior beliefs, in the same way as they tend to forget or even ignore in some circumstances those same priors (e.g. see Stich, 1981; Kahneman, Slovic, and Tversky, 1982; Kahneman and Tversky, 2000; Gilovich, Griffin, and Kahneman, 2002; etc.). Therefore, in attributing beliefs to the others', the agents neither assume that they respect the axiom of logical omniscience nor the axiom of consistency (Harman, 1973; Stich, 1981; Goldman, 1986).

That is why, when dissipating the worries concerning the RT and belief ascription, Dennett (1987, pp. 94-95) describes "what rationality is not": "It is not deductive closure. … Nor is rationality perfect logical consistency". In denying deductive closure, Dennett therefore concedes that individuals are not necessarily able to infer all the logical consequences of their prior beliefs. He even goes a step further in what could be identified as the weak rationality theory (WRT), when he claims

> "Of course we don't all sit in the dark in our studies like mad Leibnizians rationalistically excogitating behavioral predictions from pure, idealized concepts of our neighbors, nor

> do we derive all our readiness to attribute desires to a careful generation of them from the ultimate goal of survival. We may observe that some folks seem to desire cigarettes, or pain, or notoriety (we observe this by hearing them tell us, seeing what they choose, etc.) and without any conviction that these people, given their circumstances, ought to have these desires, we attribute them anyway. So rationalistic generation of attributions is augmented and even corrected on occasion by empirical generalizations about belief and desire that guide our attributions and are learned more or less inductively. … I would insist, however, that all this empirically obtained lore is laid over a fundamental generative and normative framework that has the features I have described." (Dennett, 1987, pp. 53–54)

Therefore the empirical generalizations made by agents, as is supposed in the TT account, interfere in the attribution of mental states consistent with the normative principles of the SRT. Such empirical generalizations tend to predict that the other's behavior will show lapses from perfect rationality. This statement makes it difficult to integrate the IST in games in which experiences of gaming are insufficient to determine the type of relaxation that a player can make regarding the others' capacity to conform to rationality in her mental states and behavior.

One of the critics raised against Dennett's IST is indeed the difficulty of accounting for lapses of rationality. It is hard to determine to what extent and to what kind of failures of rationality people are subject (see Stich, 1981). Another type of criticism (see also Stich, 1981) relies on his imprecise notion of rationality. He does not define precisely what he means by rationality. He indeed sometimes defines rationality with precise concepts like deductive closure but after criticism he then makes concessions and argues that rationality does not imply deductive closure. That is why Dennett defends his reference to rationality in the IST as a "myth", a normative principle that is a good approximation, in first instance, of our and other's behaviors. For him we all aspire to rationality

"However rational we are, it is the myth of our rational agenthood that structures and organizes our attributions of belief and desire to others and that regulates our own deliberations and investigations. We aspire to rationality, and without the myth of our rationality the concepts of belief and desire would be uprooted. Folk psychology, then, is idealized in that it produces its predictions and explanations by calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do." (Dennett, 1987, p. 52)

It is also interesting to emphasize that for Dennett, one explanation of the miss-attribution of beliefs lays on a social and cultural dimension: "Sometimes attributions of belief appear entirely objective and unproblematic, and sometimes they appear beset with subjectivity and infected with cultural relativism." (Dennett, 1988, p. 496) Thus, what in general tends to induce cognitive similarities (i.e. aligned beliefs, common inductive reasoning; see chapter 2 and in particular the reference to community-based reasoning) and what makes mindreading accurate is discarded in Dennett's theory.

Different strategies to weaken the SRT are possible and have been examined. For instance Stein (1996, pp. 133–134) suggests two possible ways: (i) attributors assume that their target always

obeys some principle of rationality and (ii) attributors assume that their target sometimes obeys some principles of rationality (Goldman, 2006, p. 58) The question is therefore what do these 'some' and 'sometimes' really mean and imply? This uncertainty can induce several serious flaws in the WRT, and lead to even more difficulties in the process of attribution. How can an attributor evaluate in their attribution of beliefs the extent and amount of 'some' principles? And how frequent is the sometimes? This opens a serious indeterminacy in the definition of players' beliefs and in beliefs attribution in games. Are these lapses from rationality determined by general laws? By inductive inference and generalizations from experience, as would have supposed by Dennett when defending his IST against criticisms regarding a strong version of rationality? Cherniak (1986, p. 16) offers a possible answer by considering that a "minimally rational" agent will attempt to eliminate the inconsistencies in her set of beliefs when and if these inconsistencies arise. If this is the case, this can explain how players progressively refine their set of beliefs when information is revealed, but this supposes the repetition of players' interactions. This principle is however less helpful in defining the players' prior beliefs. It can, nonetheless, allow an attributor to restrain the set of beliefs she could possibly impute to a target. For instance, this implies in a prisoner's dilemma that a player cannot believe that her co-player believes at the same time that she will cooperate and that she will defect. She must handle only one of these beliefs: either she cooperates or she defects. Believing both at the same time reveals an inconsistency. Less trivially, a player cannot believe that the other believes that she is risk averse and that she intends to cooperate; both beliefs are inconsistent. Nevertheless this still leaves clear breaches in the conceptual and pragmatic use of the WRT, particularly when we integrate player's mental states into games like their perceptions, as is the case in this thesis, which can violate this minimal principle of rationality. As Bacharach (2001) explains, framing can imply some inconsistencies (see chapter 3 section 5). And the problem of the different epistemic states in which an attributor and a target can be reinforces this weakness. If players do not have sufficient information on the other players and in particular on the way they conform to the principle of rational decisions as supposed in the RT, this can lead to serious flaws in their mental states attribution and open them to indeterminacy in their set of beliefs.

Subsequently, in such cases of insufficient information for instance, Dennett seems to adopt a position directed towards a more simulationist account. If the target does not seem to conform to the normative principles of rationality, how can we interpret and predict her choices or behaviors? Dennett answers

> "I am backing into...the view that when we attribute beliefs and other intentional states to others, we do this by comparing them to ourselves, by projecting ourselves into their states of mind. (Dennett, 1987, pp. 98–99)

The ability of others to conform to the rationality principle is therefore determined by our own ability to conform to it. From this perspective, it entails that we project our own level of rationality of mental states and reasoning onto the others.

Other difficulties are involved in Dennett's intentional stance and in the RT in general. First it does not exhaust all of the possible mental states. It remains silent on direct social perceptions and on mental states attribution regarding emotions, feelings, etc. Second it is silent too on the content of the mental states attributed. The latter are only supposed to follow the norm of

rationality (Goldman, 2006, p. 64). Mental states remain to a great extent, "black boxes", like in the TT account more generally (Dennett, 1987, p. 58). Dennett avows that his IST is "literally a black box theory" (Dennett, 1988, p. 498). These mental states matter insofar as they play a causal role for ascribing to the target an intention behind a behavior – again conforming to the rational norm (Goldman, 2006, p. 64). Intentional attribution in Dennett's view is abstracted "from the realization of belief, desire, and the other attitudes" (Goldman, 2006, p. 64) It is rather clear in the following quotation that Dennett sees mental states in a purely instrumental manner: "[w]hat makes some internal feature of a thing a representation could only be its role in regulating the behavior of an intentional system." (Dennett, 1988, p. 497)

The intentional systems theory also remains silent on the cognitive processes that drive peoples reasoning and mental states. "The intentional stance is thus a theory-neutral way of capturing the cognitive competences of different organisms (or other agents) without committing the investigator to over specific hypotheses about the internal structures that underlie the competences." (Dennett, 2009, p. 10) Others' brains remain black boxes and we refer to their mental states such as their beliefs, preferences or desires solely as the triggers of behaviors or utterances; solely as involved in a causal mechanism in decision making; thus only in order to make sense to behaviors. But generally we just need to assume "unconsciously" in Dennett's words (2009, p. 5) that others are rational. And it provides according to him a good approximation of their behavior (ibidem).

> "The central epistemological claim of intentional systems theory is that when we treat each other as intentional systems, using attributions of beliefs and desires to govern our interactions and generate our anticipations, we are similarly finessing our ignorance of the details of the processes going on in each other's skulls (and in our own!) and relying, unconsciously, on the fact that to a remarkably good first approximation, people are rational." (Dennett, 2009, p. 5)

The interest that Dennett sees in his theory is its generality and neutrality with respect to the organism under scrutiny (see Dennett, 1988, p. 487; Dennett, 2009, pp. 19-20). This theory can be applied to a human, an animal, a robot, etc. In this respect the IST hardly fits our attempt to integrate in to game theory the players' mental states as real and human, and to provide a device explaining how players form their beliefs towards others' mental states and behaviors. A theory in which peoples cognitive mechanisms and mind functioning remain a black box is of no help. Recall the major difficulties, emphasized in chapter 1, that epistemic game theory faces with a purely instrumental and mathematical use of players' beliefs. As the reference to people's mental states, such as their beliefs, is solely of such instrumental use in Dennett's intentional system theory, this cannot be applied in our approach. Dennett indeed sees his IST as "idealizing, holistic, abstract, instrumentalistic" (Dennett, 1988, p. 497) while at the same time, strangely, attempting to argue that his theory is not only instrumental; he indeed declares

> "Although the earliest definition of the intentional stance (Dennett 1971) suggested to many that it was merely an instrumentalist strategy, not a theory of real or genuine belief, this common misapprehension has been extensively discussed and rebutted in subsequent accounts (Dennett 1987, 1991, 1996)." (Dennett, 2009, pp. 14-15)

The insights that can therefore be drawn from the RT concerning both the attribution mechanism and especially the content of this attribution (i.e. supposing that an agent believes that p or desires m etc.) are limited. Again, for methodological purposes we need, in game theory, a mechanism of attribution providing a clear device (i) for specifying a defined set of players' beliefs (mainly regarding the others' preferences, intentions, beliefs and choices) and (ii) a specific content to these set of individual beliefs. As they remain "black boxes" in the RT, our theoretic solution must be found elsewhere. This elsewhere is the frame of the ST. We indeed need to define a set of beliefs for every player (which is impossible in the case of the RT or the TT when players do not have sufficient information on the other players) and in order for this set of beliefs to be defined players must be able to determine what the other players specifically believe. In a prisoner dilemma they must be able to determine if the other players believe that they are cooperative or not, for instance.

4. The Simulation Theory (ST)

Within the early roots of simulationism, Lipps (1909) is one of the first who introduces the term empathy to characterize the ability of individuals' to understand someone else, and to see others as 'minded creatures' (Zahavi, 2014, p. 103; Goldman, 2006, p. 19). Lipps (1909, pp. 225-237) defines empathy as a psychological and social concept consisting in the projection of oneself into an external object. This vision of empathy, projecting oneself into someone else to understand her, establishes? the essence of the ST. For Lipps, a large part of the capacity to be acquainted with someone else is in some relevant aspect explained by self-experience. Empathy is, in his view, a form of 'internal mimicry' (Goldman, 2006, p. 19). For example, if I have experienced anger, and see the expression of anger in someone else it evokes the feelings I once experienced, I understand her facial expression as anger, and I am therefore able to understand this someone else (Lipps, 1907a, 1907b). Simulationist ideas then diffuse somewhat and are found in the thoughts of diverse scholars such Hume (1958[1738]) Quine (1960), Smith (1976[1759]) and Dilthey (1977) (see Goldman, 2006).

The common idea within the early simulationists, which contrasts with the TT account of mindreading, is that attribution of mental states to the other does not resort to a 'scientific' or 'quasi scientific' exercise of psychology (Goldman, 2006). There is no need for theorizing about psychological law, rational beliefs, preferences or desires to come to a prediction of someone else's behavior. In Guala (2018, p. 363)'s words, simulation is "more a matter of doing – of replicating, in particular, the reasoning of other agents." Simulation therefore offers a much more pragmatic solution for games. It provides a good and 'plausible' explanation of how players attribute to the other players' some mental states. It offers a pragmatic approach to games and strategic reasoning in which a player's limited cognitive capacities are taken into account.

One of the main arguments in favor of the ST is also that contrary to the TT, which focuses on the attribution of propositional attitudes (i.e. beliefs, intentions, goals, etc.), it can account for a variety of mental states in addition to the propositional attitudes, such as sensations, emotions

and feelings (like anger, fear, joy etc.). Accordingly, the ST is a much more general mechanism of attribution, potentially very effective in our everyday life since understanding others is not reduced to the beliefs, intentions and purposive types of mental states they handle. In addition, as already mentioned, coordination is manifold; emotionally based interpersonal understanding is one of its mediums. There is in fact much empirical and experimental evidence showing that the ST mechanism is very effective for reading others' emotions and feelings and which favor successful coordination. As Goldman (2006, p. 20) points out:

"Rival theories should be tested by their ability to account for the mindreading of all categories of mental states. This immediately poses difficulties for certain contenders, such as the rationality theory, that are ostensibly directed exclusively to the attitudes."

However, even in the case of the attribution of propositional attitudes, one of the key virtues of the ST compared to the TT is its computational parsimony. The body of knowledge that the individuals are supposed to hold in folk-psychology to ultimately mind read someone else imposes a heavy burden on mental computation (Gordon, 1986; Goldman, 1995; Stich and Nichols 1992, 1995). Again as accurately summed up by Guala (2018 p. 363)

"In order to attribute beliefs, from a theory-theory perspective, I must try to figure out what information the other players have about the situation. Then, using a theory about the way they reason, I must try to infer their beliefs. But in a coordination game the actions of each agent depend on her beliefs about the actions of the other agents, and vice versa. So I cannot fix one variable until I have fixed the others. Using a theory-theory perspective I cannot issue a precise prediction, because the model is underdetermined by the available data."

However, if instead of computing such a body of knowledge as in the TT the mindreaders only have to use their own mind, this burden vanishes: a "mindreader makes special use of her own mind in assigning mental states to others." (Goldman, 2006, p. 40) They just have to run a simulation of someone else's decision, assuming the – pretend similar – underlying mental states of a common decision-making mechanism. Simulation is accordingly, potentially a much more effective heuristic.

We will detail in the next subsection how this simulation process is deployed, and in particular, on what cognitive resources and mechanisms it is drawn by Goldman (2006). We will mainly develop Goldman's account of mindreading in this section and will explain this choice. At the end of the section we will nevertheless contrast Goldman's with Gordon's account to show how the former is more suitable for game theory and less problematic from both an epistemological and a methodological perspective . As for the TT and the RT, simulation can involve first or third person mind reading. As previously, I mainly interpret the use of simulation for third person mindreading as being interesting primarily for the investigation of the way Bayesian theory can be replaced by another methodological device allowing it to explain and formalize players beliefs on the other's eventual behaviors and beliefs.

4.1. The ST paradigm

Contrary to the commonsense psychology account and therefore both the TT and RT:

> "A fundamental idea of ST is that mindreaders capitalize on the fact that they themselves are decision makers, hence possessors of decision-making capacities. To read the minds of others, they need not consult a special chapter on human psychology, containing a theory about the human decision-making mechanism. Because they have one of those mechanisms themselves, they can simply run their mechanism on the pretend input appropriate to the target's initial position. When the mechanism spits out a decisional output, they can use the output to predict the target's decision." (Goldman, 2006, pp. 19-20)

There is no need to possess some knowledge of the laws of psychology, no need for making general laws such as 'wanting and believing induce a decision' or such as 'people are rational in an instrumental manner: they attempt to achieve the most desirable outcome provided their desires and beliefs', or as supposed by Dennett that 'people hold the belief and the desires that they ought to have as rational people', etc. It means that the ST provides a most pragmatic approach to the mechanism of attribution and strategic reasoning, as we will show in chapter 5 of the thesis.

In the ST, mindreading, for instance the prediction of someone else's decision, is a three step process. The first step is the attributor's simulation process per se. The attributor first brings forth 'pretend' or 'imaginary' mental states in her own mind which are supposed to correspond to or 'mimic' those of her target. The simulator therefore pretends to have the same perceptions, beliefs or desires as her target. Then, she feeds these pretend states into her own cognitive scheme, and then into her decision-making system. Her decision-making system runs an output: a decision. However, the decision is not performed. The decision-making system is disengaged from the motor system. That is why many scholars in the ST talk about "off line" simulation. This means that, whatever the cognitive mechanism involved, it runs offline (ibid, p. 20). The decision generated by the simulator's decision-making process is finally attributed to the target. This is what Goldman (2006) calls "projection". This is the last step of the simulation routine.

Projection does not only apply to the output that the decision-making process runs but also to the initial pretend mental states. "In acts of projection, an actual state of the self is projected onto the target, whether the state is an evaluation state, a somatic state, or whatever." (ibid, p. 173) The whole process is accordingly, for Goldman (ibid, p. 40), a "simulation plus projection routine". The projection stage of the process as a final stage implies an "exit" from the two first stages of the simulation routine (ibidem). So at the end of the process, if the attributor's decision-making process and pretend initial mental states are sufficiently similar to that of the target, so is the decision projected to the target. And the attributor can ultimately make a reliable prediction of her target's decision. Therefore, the simulator clearly uses her own cognitive apparatus as a 'model' of the other's cognitive apparatus.

In the ST paradigm two main characteristics determine the simulation process: (i) the role of pretense and of pretend states, and (ii) the similarity postulate, i.e. the attributor counts on the fact that she uses the same mechanisms of thought and decision-making as the target (Goldman, 2006, 2012). As was already partly developed in the chapter 2 and 3 through community based reasoning in particular, and as will be developed in the next section through the mindshaping hypothesis, belonging to a common community supports this similarity postulate. More precisely it supports the similarity of mental states (perceptions, desires or intentions and beliefs), and of decision-making.

The act of pretense is fundamental in Goldman's account of the process of simulation. It allows people to feed their decision-making system with the mental states they attribute to the others and which are supposed to mimic the mental states effectively handled by these others. The way Goldman defines pretense is more precisely as follows:

> "Pretend states—in the sense intended by ST—are states produced by enactment imagination (E-imagination). A pretend desire is the product of enacting, or attempting to enact, desire; a pretend state of fear is the product of enacting, or attempting to enact, fear; and so on. Pretend desire is quasi desire produced by E-imagination, pretend fear is quasi fear produced by E-imagination, and so forth. … What simulationists call "pretend states" are states like make-believe, make-desire, and so forth. They are states produced by an operation of mental pretense, or E-imagination." (Goldman, 2006, p. 48)

This idea of pretense or imagination, is confirmed by experimental results which show that when a player is playing games and has to form beliefs towards others' beliefs and choices (e.g in McCabe et al., 2001; Gallagher et al., 2002; Bhatt and Camerer, 2005) while their brain is being scanned, the area of the brain that is activated (in particular the medial prefrontal lobe or the anterior paracingulate cortex) is the area involved in the creation of "decoupled representations of beliefs about the world", i.e. "decoupled from the actual state of the world." (Singer and Fehr, 2005, pp. 340-341; see Frith and Frith, 2003). As there is a distancing, a detachment from reality in pretense we can assume that it can involve, as supposed by Goldman, a form of imagination.

> "E-imagination" in Goldman's view is a 'psychological construct' that can be instantiated either as a process (i.e. a decision making mechanism) or for its inputs (i.e. beliefs, desires, etc.) and outputs (i.e. the decision), and may be both voluntarily or involuntarily, conscious or unconscious (this marks the distinction between 'high-level mindreading' and 'low-level mindreading'). E-imagination is used for a wide spectrum of mental states (Goldman, 2006, p. 151).

More specifically,

> "To E-imagine Xing, where X is some kind of mental state, it does not in general suffice merely to suppose or hypothesize that Xing occurs in you. To enactively imagine seeing something, you must "try" to undergo the seeing—or some aspects of the seeing—despite the fact that no appropriate visual stimulus is present." (ibidem)

E-imagination implies a form of visualization or at least mental visualization. Lots of experimental settings have focused on this imagination and visualization part of mindreading

therefore providing a justification of the simulation heuristic. Psychological and neuroscientific findings also show that the use of visualization and accordingly E-imagination yields accuracy in the attribution process (ibid, p.157). These researchers additionally point out the extensive use of 'self-reflection' and 'self-reference', which again confirm the role of E-imagination and at the same time of simulation in mindreading (see Kelly et al., 2002; Gilbert, Gill and Wilson, 2002; Mitchell, Banaji and Macrae, 2005).

> "Self-reflection, or self-reference, is a natural subactivity of third-person mindreading according to ST. For one thing, simulational mindreading requires an attributor to monitor her own genuine states so as to identify those that should be inhibited and excluded from a simulation routine. In addition, a paradigm simulation routine, at its next-to-last step, involves detecting and classifying an output state of a cognitive mechanism (e.g., a decision). This detection-and-classification step requires self-reflection." (Goldman, 2006, pp. 147-148)

When an attributor imagines what her target's perceptions, feelings, desires, thoughts, intentions, or behavior in the situation of her target would be, she necessarily uses self-reflection to understand the mental states of her target and the options she may have, and distances herself from her own mental states in order not to attribute her own mental states to her target. For instance I might be risk averse and play a secure strategy in the prisoner's dilemma, i.e. defection, but I can be perfectly aware that the other may possibly be a risk taker and hence that she is more inclined to cooperate. Thus I can attribute to her the intention to cooperate while my own is to defect. In Goldman's account of the ST, self-reflection allows the attributor to simulate and project some mental states – in an 'anchoring phase' – and to adjust this 'simulation plus projection' – in an 'adjustment phase' – when the prediction or explanation of the target decisions do not seem accurate (see Gilovich, Savitzky, and Medvec, 1998; Epley, Morewedge, and Keysar, 2004; Epley, Keysar, van Boven, and Gilovich, 2004). This process of adjustment will take the form, in the next chapter, in the game theoretic model proposed, of what we have identified as the massaging process which is a form of rationalization of the others' beliefs. As the game theoretic model presented in the next chapter is an attempt to get rid of the standard prior beliefs assumption on the one hand, and as it relies on the integration of players' subjective perceptions on the other hand, resorting to self-reflection in mental states attribution is particularly relevant. For instance, the beliefs that the players attribute to their co-players are determined in first instance according to their own but the adjustment phase, i.e. the massaging process, which is the form of rationalization of these beliefs, possibly makes these beliefs and decision different from what the players initially attributed to their co-players. It provides a methodological device to establish endogenously and from the players' perspective the beliefs they hold.

The second point of particular importance after pretense for Goldman is the similarity postulate between the attributor and her target. The pretend states that are fed into the attributor decision-making system must be similar or at least sufficiently similar to that of the target. On this condition only, can the simulation of the target's mental states, decision or behavior, be accurate. According to Goldman (2006, p. 49), three kinds of similarity between pretend and real mental states justify such supposition: (i) 'introspectible', (ii) 'functional' and (iii) 'neuronal'. The similarities that Goldman assumed are therefore of a cognitive basis. In the next section we will

see that similarities of mental states and behaviors among people is also of a social basis. Goldman refers to numerous empirical findings confirming the resemblance he asserts. Yet it is evident that any simulation may not be accurate. Many elements can prevent accuracy. The mental states attributed to the other can be wrong, "they may have been chosen badly out of ignorance." (ibid, p. 48) Insufficient information about the target can interfere (ibid, p. 175). Both attributor and target can be in very different epistemic conditions, etc. That is why, for Goldman,

> "a reasonable version of ST would not hold that the mental processes of mindreaders always match, or even approximately match, those of their targets. ST, like any plausible theory of mindreading, should tolerate highly inaccurate specimens of mindreading. … What ST essentially maintains is that mindreading (substantially) consists of either successful or attempted mental simulations." (ibid, p. 38)

In the end, the aim of a simulation is merely to be as close as possible to the target's mental states, decisions, etc. A simulation process intends to mimic, 'duplicate' or 'match' a target (Goldman, 2006, p. 38). And it is not always successful. It offers, in the first instance, a pragmatic heuristic of choice. In fact, numerous experimental findings report that there are pervasive egocentric biases in simulation that preclude accurate mindreading. This point will be discussed in the section 4.3., as studies on this egocentric bias provide an important source of empirical findings that tend to favor the ST explanation of mindreading as compared to the TT and RT approaches. Simulation is in many cases a heuristic, and a very pragmatic principle for anticipating and explaining other's choice although not always accurate. It offers a very pragmatic principle for decision making in strategic contexts even with great uncertainty regarding other's informational state, other's preferences, other's possible decision rule, etc. In many case this heuristic of choice is largely sufficient to come to a decision with very little information regarding the others. As will be exposed in more detail in the next section with mindshaping, thanks to social life and different community memberships, a sufficient degree of similarity among people is ensured so that the prediction to which the player comes using simulation is in many case accurate enough to bring coordination. Mindshaping adds to the cognitive similarities postulated by Goldman (2006) social similarities, which reinforce the potential accuracy of prediction.

Another important dimension that will be discussed in more detail in the subsection 4.2., is the capacity of the ST to account not only for high-level mindreading, i.e. for the attribution of propositional attitudes and metacognitive dimensions such as preference, beliefs or desires, but also to account for low-level mind-readings, i.e. for the attribution of feelings, or emotions for instance (contrary to the TT and RT which only range for propositional attitudes). Feelings or emotions can be an important source of coordination in face-to-face interactions. Recall that coordination is understood as a more broad and general phenomenon than mere coordination games like the Hi-Lo game. Coordination is understood as the convergence of individual perceptions and beliefs in real and constructed (as in experiments for instance) strategic interactions. In that way, the ST, again contrary to the TT and RT, can encompass both conscious (through the attribution of propositional attitude) and unconscious simulationist mechanisms (through the recognition of facial emotional expression) (Goldman, 2006, p. 40; Jeannerod and Pacherie, 2004, pp.128–9; Zahavi, 2014, p.171).

A last dimension that naturally follows from the ST thesis and which is of particular importance for beliefs attribution in game theory, is the capacity for players to use the simulation process to attribute propositional attitudes to others and a content for such propositional attitudes.

> "Discussions of high-level mindreading typically focus on "whole" propositional attitudes, that is, attitude-content pairs. ... [ST explores] the way(s) that mindreaders construct contents for others' intentional states." (Goldman, 2006, pp. 175-176)

In other words simulation makes it possible for players to believe that the other players believe that p or believe that q. They assign a specific content to the belief they attribute to them. They not only assume that a desire and a belief together lead to a decision as in the TT account. An attributor simulates her target's mental states and builds in her own cognitive scheme a set of pretend mental states she considers to hold for her target to come a specific choice; in this process, the propositional attitudes that she builds have a content. This dimension, as already mentioned, is much more uncertain in the RT and TT, which moreover, does not drive much of the attention of the scholars within these frames. To underpin this epistemological dimension of the ST, some experiments have investigated this capacity of content attribution (see Spelke, 1990, 1994; Soja, Carey and Spelke, 1991; Casati, 2003; and see also Bloom, 2000, chapter 4).

4.2. Simulation with and without introspection: the distinction between high-level and low-level of mind reading

There exists a distinction between high-level mindreading supposed to be about the conscious attribution of propositional attitudes such as desires, intentions or beliefs and low-level mindreading, which is about the unconscious share of emotional states or feelings via the automatic activation of the same neural networks.

My interest is obviously primarily in high-level simulation and the attribution of propositional attitudes such as beliefs, desires, or intentions, etc. to the others. Yet this does not mean that in a broad apprehension of the informational device valuable in real strategic interactions, low-level mindreading does not play a great role. This means that exhausting the extent of the forms of attribution that the ST (i.e. of high or low level) offers finds a place in this thesis. When strategic interactions occur in the real world, without communication, face to face encounters (like in repeated games for instance) provide an informational basis for feelings or emotions that may help players to attribute to the others some perceptions or intentions, that will eventually matter in their decisions. This is in particular the ground for the Direct Social perception thesis that will be exposed in the next section.

The discovery of mirror neurons triggered a surge of investigations expanding the pivotal place that the ST has in the ToM framework. Its impact is manifold. First and foremost, it justifies the prime postulate of simulation: i.e. the attempt to mimic or mirror a target. As Goldman (2006, p. 37) acknowledges

> "it is doubtful that all simulation is purposeful. Some simulation may be automatic and nonpurposeful. Even without purposefulness, however, a phenomenon intuitively counts as a simulation of another if it is the function of the former to duplicate or resemble the other."

Second it leads to the identification of the different neurocognitive architectures involved in low- and high-level mind reading. Third it shows how effective simulation is in various interactional circumstances.

However this discovery also leads to tremendous criticisms. Mirror neurons entail an unmediated mechanism. While it has been indicated that a simulation involves the attribution of mental states and propositional attitudes to a target, such conscious and voluntary mental acts do not prevail in mirroring. The problem is that many scholars speak of empathy and simulation when referring to the activation of mirror neurons (e.g Singer et al. 2004; Singer and Fehr, 2005) while empathy should imply mindreading and accordingly the activation of neural bases involved in deliberation (Gallese, 2003a; Iacobini, 2007). But many proponents of the ST account do not hold such a strong claim. They only assert that mirror neurons can be one of the bases of simulation routines but that it does not exhaust its mechanism. As Goldman argues:

> "mindreading involves mental attribution to a target. This requires two mental acts by the attributor: selecting a mental-state category, or classification, and imputing an instance of that classification to the pertinent target. Let's call these acts M-classification and imputation … mirroring per se entails neither M-classification nor imputation. It requires only that a receiver undergo a matching event, which doesn't guarantee that she has a repertoire of M-classifications. Even if she has such a repertoire, she may not deploy it on the present occasion. Moreover, mirroring doesn't entail mindreading because the receiver may not impute anything to the sender. Although mirroring doesn't entail mindreading, as a matter of definition, it is entirely possible that mirroring is used as a basis of mindreading." (Goldman, 2006, pp. 133-134)

While mirroring does not involve the mechanisms of a conscious simulation process as described in the previous subsection, Goldman is not hostile to the idea that it may however be a part of mindreading. Thus the question remains open concerning the link between mindreading and mirroring. As Gallese and Goldman (1998, p. 498) speculate that " [the mirror neurons] represent a primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mindreading." Some others hypothesize that the brain "mirror modus operandi" (Kirman and Teschl, 2010, p. 306) can contribute to understand others' action, intention, or emotions (see Wicker et al. 2003).

4.2.1. Low-level mind reading and mirror neurons

Goldman defines low-level mindreading as "simple, primitive, automatic, and largely below the level of consciousness (Goldman, 2006, p. 113). It mainly concerns what is called "Face-based emotion recognition" (FaBER). It is documented by many experimental results.

There are different mechanisms of simulation for emotional attribution but the one which has recently received more attention is mirroring, i.e. the activation of mirror neurons. Such activation provides the grounds for the attributor to access the emotion of their target. The other types of mechanisms are largely described in Goldman (2006, pp. 124-128). Being mainly interested in high-level mindreading, an exhaustive presentation of the different types of routines involved in emotional attribution is not appropriate for this thesis. However, since the discovery of mirror neurons pushed the investigations of simulation routines in interpersonal interactions considerably and put the ST at the center of the ToM, they are worth mentioning here and we consider it important to draw attention also to the debates they engender.

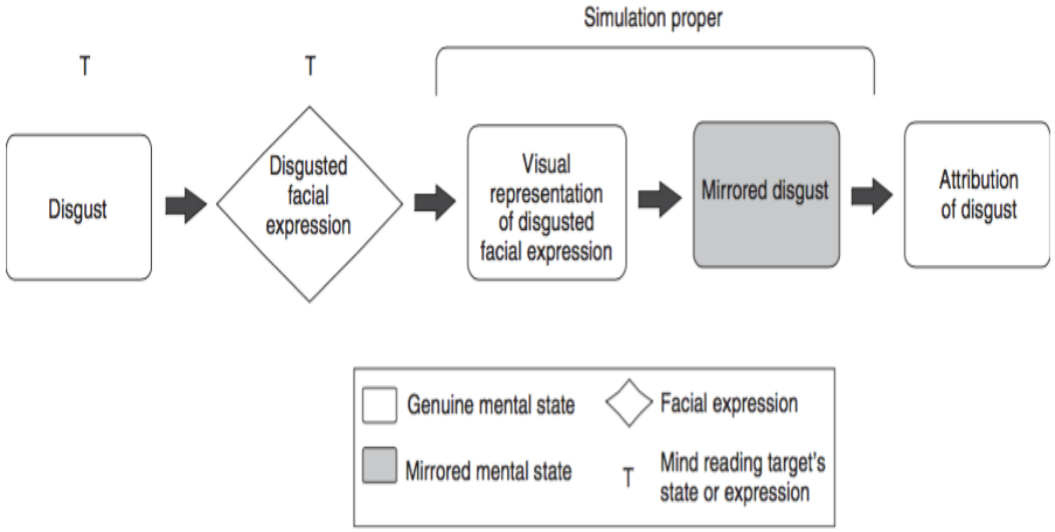
Since the discovery of the existence of mirror neurons and a "mirroring system" in the ventral premotor cortex of monkeys, by Rizzolati et al. (1996) and Gallese et al. (1996) many studies have been conducted on humans proving the existence of a wide range of mirroring mechanisms, so that many types of cognitive states of the observed are mirrored by similar cognitive states in the observer. Mirror neurons constitute a "matching system", or a "resonance system"; the neural activity of the observed when performing an action or showing an emotion (such as pain or disgust) instantiates in resonance the activation of the same neural activity in the observer, though actions or emotions are not effectively accomplished or triggered in the observer, but are inhibited.

As previously emphasized some scholars consider mirroring as a kind of empathy without a conceptual task or 'mental classification' (e.g. Gallese, Keysers and Rizzolatti, 2004). Viewed from that perspective, mirroring is a form of 'social understanding' in the same way as high-level mindreading is (Goldman, 2006, pp. 136-137). Besides, some studies show that the mirroring system is more than a simple emotional contagion (see Singer et al., 2004, Jackson et al., 2004; Iacoboni et al., 2005). For instance the motor mirroring system can serve as the basis for intention ascription (see Iacoboni et al., 2005; Singer and Fehr, 2005). All of this leads Goldman to conclude that

> "mirroring-based attribution of mental states is quite strong, at least in three categories: emotions, feelings, and intentions. Because mirroring is one species of interpersonal simulation, this lends weighty support to the thesis that low-level mindreading typically proceeds by simulation. I do not claim that whenever there is mirroring, there is also low-level mindreading. But wherever there is mirroring, the potential for simulation-based mindreading is there, and creatures with the requisite conceptual resources, especially humans, seem to exploit this potential extensively." (Goldman, 2006, p. 140)

How then do mirroring and mindreading function?

The schema below taken from Shanton and Goldman (2010, p. 2) represents the role of an observer's mirroring system in the assignment of an emotion of disgust to her target showing a facial expression revealing disgust (the same schema could have been applied to other emotions such as fear, anger, sadness or happiness, etc.). The observation of disgust in the target's facial expression generates an automatic mechanism activating in the observer the same neural basis as the target experiencing disgust. On these grounds the observer can understand the emotion of her target and identify the emotion as disgust. In other words it allows the attribution of an emotion based on a facial expression. The re-experiencing of disgust in the observer is a kind of simulation process (see also Goldman and Sperida, 2005). And for Goldman, this is a form of low-level mindreading.



Except the simulation part of this low-level mindreading, this process is close to the perceptual interaction at the core of direct social perception (DSP) (cf. section 5.1.). DSP entails that most of our interactions and ability to coordinate in the real world does not pass through a conscious reasoning and through deliberation involving the attribution of propositional attitudes but is first and foremost unconscious. Facial recognition of an emotion may be sufficient to be acquainted with the target and thus to understand her. In the DSP thesis, facial recognition, the context, the existence of social customs, routines etc. are much more effective coordination devices.

4.2.2. High-level mindreading

On the other hand, high-level mindreading involves (i) the attribution of complex mental states such as propositional attitudes and (ii) a – predominantly – conscious process, i.e. voluntary control (Goldman, 2006, p. 147). Its neural bases are therefore different than for low-level mindreading.

This process of attribution through simulation can be used either for predicting or for explaining someone else's decisions or behavior. To illustrate how a simulation routine works in each case we refer to some schemas detailing the whole processes.

First: the prediction by an attributor of her targets' decision.

The situation is the following. A target has a decision to make. The attributor will intend to predict her decision by simulating her decision making process and for that the attributor will simulate her eventual mental states. The situation depicted is that of a decision making process in which the target is supposed to have a 'desire' for realizing the goal g, and a belief that the action m will allow her accomplishing this purpose.

This means that the attributor puts herself in the other's shoes and generates in her own mental scheme (i) a 'pretend desire' for a goal g (that she supposes to be the goal of the target), and (ii) a 'pretend belief' that m is an effective way to reach such goal g (again that she supposes to prevail for her target). Then the attributor feeds her own decision-making process with the pretend desire and belief. She comes to a 'pretend decision' to do m. The result of the attributor's simulation routine is a – real or 'genuine', i.e. non-pretend – belief that her target will accomplish the decision m to which she comes when running her simulation. And she finally imputes, or projects, to her target the effective decision to perform m. In other words she predicts that the target will do m (ibid, pp. 28-29).

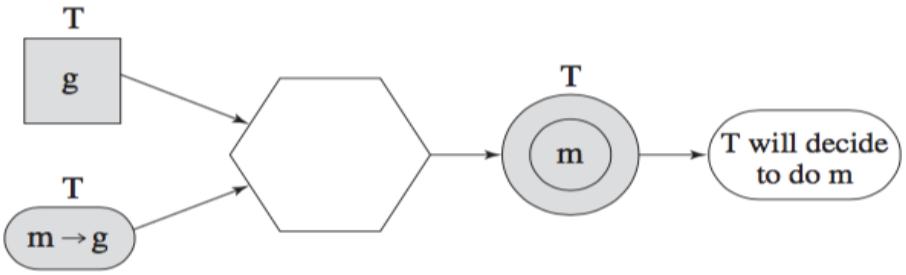
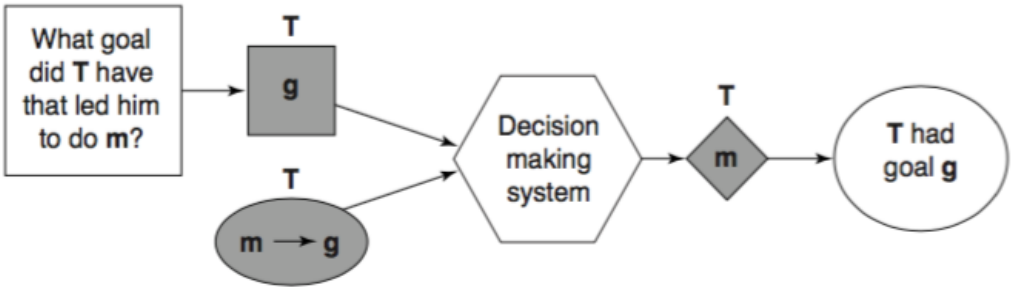


Figure 2.3. Decision attribution reached by simulation. (Adapted from Gallese and Goldman, 1998, with permission from Elsevier.)

The shaded square describes the pretend desire; the shaded oval represents the pretend belief; and the hexagon the decision-making mechanism. The 'pretend decision' run by the attributor is depicted by the shaded double circle, and the ultimate real belief on the action that will accomplished by the target is represented the non-shaded oval (as it is not a pretend belief).

Second: the explanation of a target's decision. This is what Goldman (ibid, p. 183) calls the "retrodictive use of simulation".

The decisional situation of the target is exactly the same as in the case of prediction. The schema is taken from Gallese and Goldman (1998, p. 498).



From a known behavior the attributor attempts to determine the target's mental states causing the behavior: the "sought-after causes." (Goldman, 2006, p. 183) The attributor therefore thinks backwards. It is for Goldman (ibidem) a "generate-and-test strategy." The "generate phase" determines the pretend, or "hypothetized" states that led to be the situation observed, and this phase is for Goldman (ibid, pp. 183-184) "executed by non-simulative methods." The "test phase" is an assessment of the way the hypothesized states combine each other to produce the situation observed, i.e. the choice or the behavior. The simulation routine runs during the "test stage." The attributor imagines being in the initial mental states she imputes to the target, feeds her decision making system with these pretend mental states, and comes to a pretend decision. Then she compares the pretend decision triggered with the effective decision made by her target. If it matches, she has found a possible explanation, if not, she runs another simulation with alternative pretend initial mental states, and so on.

Such retrodictive use of simulation is particularly relevant for repeated games for instance, in which players may attempt to explain the others' past behaviors in order to understand the conditions (i.e. the emotions, intentions, or desires) that led them to such behaviors and thus anticipate their future behaviors. If players better understand the cause of the others' past behavior they can better predict their future behaviors (providing that the same intentions or desires prevail).

In these processes of simulation for high-level mindreading, a question arises about the inferential power needed for attributing choices or beliefs and desires, etc. to a target and therefore if it has something to do with a form of theory as in the TT. Regarding that kind of concern, Goldman answers

> "I may have a tacit set of principles that guide my own inferential steps. But these principles don't constitute a folk-psychological theory; they don't tell me how people in general make inferences. They only instruct me what conclusions to infer from various premises, not what other people will infer from those premises. In other words, they are the rules that drive, or govern, the operating procedures of my own reasoning mechanism. It is possible, of course, to apply these principles to the task of predicting other people's inferences. But such an application would simply consist in a simulation routine. It would be a familiar matter of pretending to have the specified beliefs of the target and feeding them into my reasoning mechanism to see what it outputs. Then I attribute that output to the target."(ibid, p. 181)

Lots of experimental setups report systematic errors either in the prediction or the explanation of individual decisions and behaviors (e.g. see Stich and Nichols, 1992, 1995; Nichols et al., 1996). For the simulationists, such errors can occur either because of the difference between the attributor's 'pretend states' and the 'genuine states' of the target or because of a difference in their respective decision-making mechanisms or 'executive systems' (see Maraffa, 2015). In these cases contrary to the TT account, the simulator has only to try other simulations and to test with different pretend mental states or with different decision making systems (e.g. instead of running the simulation with maximization as the rule of decision a satisficing rule can be tested, etc). In the ST framework, an important literature exists on what is called egocentric biases in mindreading. Such egocentric biases induce the failure of simulation and the inaccuracy of prediction or explanation. The theory theorists disagree with the explanation of the simulationists, arguing on the contrary, that failures of prediction are better explained by an incomplete folk psychology. Lack of valuable resources in an individuals' – naïve – theory of psychology is a source of mistakes (again see Stich and Nichols, 1992, 1995; Nichols et al., 1996; and also Saxe, 2005). The disagreement between both approaches – TT and ST– is well settled between Saxe and Goldman (see Saxe, 2005; and the reply of Goldman, 2005). The next section however records numerous experimental findings tending to justify the ST explanation of these errors of mindreading.

4.3. Failure of mindreading: egocentric biases and lacks of quarantine

An egocentric bias occurs when we do not quarantine our own genuine mental states and feed into our simulation routine these genuine states instead of the pretend mental states supposed to mimic those of the target. These genuine mental states can be very different from pretend mental states. A successful simulation requires such quarantine (Goldman, 2006, pp. 28-29)

Failure of quarantine entails simply the projection of our own genuine mental states to our target and the risk that there is no overlapping between the target's real mental states and our own. Here genuine states must be opposed to pretend mental states. The projection of our genuine states onto our target thus indicates a failure of quarantine, and the failure of the simulation routine proper, as simulation entails feeding our own decision-making, or other cognitive mechanisms, with pretend states. In the existence of such bias, an indirect answer to the content attribution problem of mental states is also found, since people tend to attribute to the others their own mental states. These attributed mental states are thus necessarily content-laden. For instance, we assume that the beliefs we hold are the same for our target. This echoes Bacharach's assumption in the VFT in which he assumes that the others' beliefs must be at least the same or a subset of our own but never different. This will also be of particular importance in the next chapter as players' prior beliefs regarding the other beliefs and then choices will be determine by their own. We thus suppose an egocentric bias in each simulation process in our model of games. On the contrary, pretense means the acknowledgment that the other's mental states must be – at least partially – different from our own.

Goldman therefore asks the following question: "Should this be considered an instance of simulation, or simulation-plus-projection?" and he answers "Given our definition of projection, taking a genuine state of one's own and ascribing that state to another is clearly a case of projection." (ibid, p. 41)

How does the mechanism of quarantine function?

Below is a schema taken from Goldman (ibid, p. 30), which shows a simulation routine and the quarantine of the attributor's own mental states (and presently her genuine desires and beliefs). The mental states under quarantine are represented in the bottom of the schema. At the bottom this figure the current desires of the attributor, which are not g – i.e. the target's desire for reaching the goal g –, but a desire for h, and a belief that does not entail that the action m will lead to g, i.e. the simulator does not believe that the action m will lead to the consequence g. The top of the schema depicts exactly the same story as detailed in the section 4.2.2.

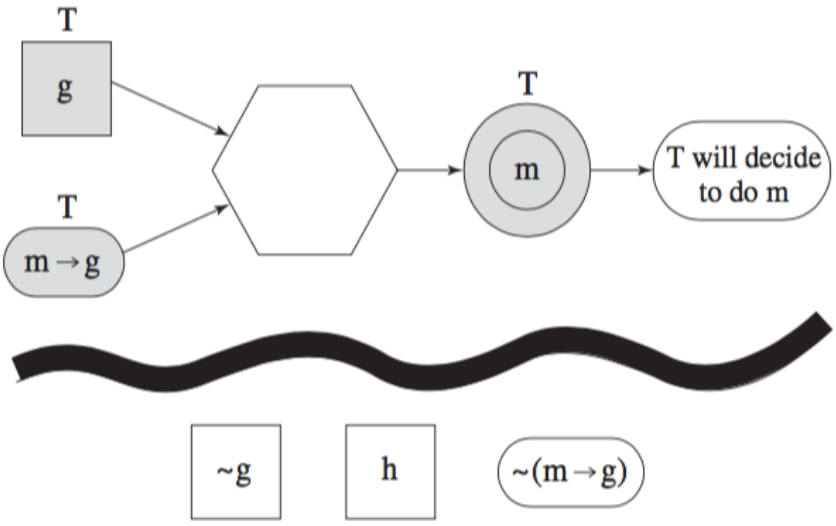


Figure 2.4. Decision attribution reached by simulation, showing quarantine.

Empirical evidence shows that a wide range of mental states are affected by egocentric biases and therefore that failure of quarantine is an extensive phenomenon in attribution and mindreading – as accounted by ST (ibid, pp. 41-42). Psychological studies show that this is the case for knowledge and belief attribution (Krauss and Glucksberg, 1969; Camerer, Loewenstein and Weber, 1989; Newton, 1990; Keysar, 1994; Keysar and Bly, 1995; Nickerson, 1999, 2001; Keysar, Lin and Barr, 2003; Birch and Bloom, 2003, 2004); for valuations and preferences (Ross, Greene and House, 1977; Van Boven, Dunning and Loewenstein, 2000); for desires (Loewenstein, Prelec and Shatto, 1998; see also Read and Van Leeuwen, 1998); for feelings (Van Boven and Loewenstein, 2003), for attitudes, somatic states, etc. (see Goldman, 2006, pp. 167-168 for the details and explanation of all of these experimental findings). For Goldman (2006, p. 167), contrary to what Nichols and Stich (2003) argue, "default attribution is not confined to default belief attribution". Besides, the other forms of default attribution, like for feelings, are hardly

compatible with the explanation of the TT in terms of an insufficient folk psychology theory since the latter is restricted to propositional attitudes among which feelings do not appear.

This egocentric bias will be an important dimension of the model proposed in the next chapter. It will provide a justification of the initial set of players' beliefs regarding the others' choice (and eventually beliefs). In the first instance, players assume that the other players perceive, want, believe and reason in the same way as themselves. They can however revise these beliefs, and more precisely they will rationalize such beliefs through what we call a massaging process. All of these aspects of the model will be discussed in the next chapter

4.4. The different forms of ST

As Nichols and Stich claim, (see Stich and Nichols, 1997; Nichols and Stich, 2003) so many variations of the ST account exist that gathering approaches so different within the same label, ST, can be confusing. For instance Gordon and Goldman, who are considered as the two leaders of this account are opposed on many points concerning the role of simulation in mindreading and its mechanisms (see also Goldman, 2006, 2010, 2012; Maraffa, 2015).

On what dimensions do the different ST approaches vary? Principally on three dimensions of 'strength' for Goldman (2002a, 2006) and Jeannerod and Pacherie (2004): (i) its 'frequency of use', (ii) its 'source', and (iii) the range of mental states up for simulation. In the strong version, scholars argue that (i) simulation is always used for third person mindreading; (ii) simulation is the main basis of 'interpersonal mentalization'; and (iii) all types of mental states are up for simulation. Weaker versions, on the contrary, state that (i) simulation is used occasionally for third person attribution, or that it is a 'default mode'; (ii) simulation is a 'shortcut mode' that ultimately tends towards more theorizing; and (iii) only a subset of mental states are up for simulation (e.g. only propositional attitudes, or conscious states, etc.), (Goldman, 2006, p.141). While Goldman is inclined towards a weak version of the ST, Gordon is representative of the strong version. For instance, Goldman admits a mix of simulation and 'theory-driven' inferences (see Maraffa, 2015). Another kind of disagreement among the weak and strong versions of ST appears around the uniqueness versus multiplicity of systems involved in simulation routines. As we saw, Goldman identifies different systems or 'operations' for simulation routines, as his distinction between the low- and high-level mindreading shows (Goldman, 2006, pp. 42-43).

The disagreement between Goldman (2003, 2004, 2006, 2012) and Gordon (1995a, 1995b, 1996) is even greater considering Gordon's rejection of what he considers to be the main premise of that version of the ST: the ascription of mental states to others' by introspection, and the required analogy between oneself and the others, which therefore necessitates that the attributor possesses the mental states before attributing them to someone else (Gordon, 1995b, p. 53). Gordon departs from all of these commitments of the standard – or weak – version. He argues that when we simulate someone else's mental states, when we intend to predict her decision or behavior, we do not imagine ourselves in her position and transfer or project a part of our cognitive states onto the other. Simulation requires for him an "egocentric shift", a "recentering"

(Gordon, 1995b). I "transform" myself into someone else: I cease to be me to become, at least transitorily – during the simulation process –, someone else. And "[o]nce a personal transformation has been accomplished, there is no remaining task of mentally transferring a state from one person to another, no question of comparing [the other] to myself" (Gordon, 1995b, p. 56). When simulating I do not see the situation as a kind of 'what would I think or do being someone else', but merely and directly what the other thinks and intends to do, without any analogical inference from me to the other.

Gordon's view implies serious issues. What becomes of the attributor during the simulation process? Is her identity preserved? The subject of a simulation remains an individual that does not cease to exist. How does she attribute content to the other's mental states? What amount of information about the target is necessary? When individuals are in very different epistemic states or epistemic levels, it can be very difficult to become this other, to see the situation from her perspective. Besides, while Gordon rejects the postulate of analogy between the attributor and the target, empirical evidence (reviewed in the previous section) on egocentric biases shows that this is on the contrary pervasive in simulation. And again, empirical evidence points towards the use of similarity judgments in third person mindreading (see Goldman, 2006, pp. 162-164). Besides, as will be developed in the next section, similarity between the simulator and the target, not only from a cognitive point of view but also from a social point of view, is necessary to come to an accurate prediction of the target's behavior for instance.

The rejection of the introspective identification as a 'classification' process allowing oneself to differentiate from the target therefore seems dubious.

For these reasons we rely on Goldman's view of ST. In addition to the fact that his view is much more moderate and much more developed, and empirically grounded, it does not break with a subjective account of individuals and does not endanger the identity of the subjects. On these grounds intersubjectivity is essentially grasped by the capacity of an individual to become acquainted with others thanks to her own mental states and decision process and not by becoming transitorily another person. As we will show in the next chapter this makes the information basis that the players have in incomplete information games sufficient to define their set of beliefs regarding the other players' beliefs and choice. They merely use, in first instance their own perceptions, intentions, preferences and modes of reasoning, to define the others' ones.

Finally, on the critique expressed by Gordon on the collapse of the ST into the TT, in its weaker version and in particular in Goldman's version, Goldman answers that mixing theoretical inferences and simulationist routines does not endanger the role of simulation in third person mindreading. The acquaintance with the other's mental states or decision remains a process of simulation even if at some point of this process theoretical inferences intervene. As he puts it:

> "simulation for mindreading is (arguably) a "control" operation that directs a variety of processes or operations. When one simulates a decision process, for example, the control operation directs the creation of pretend desires and beliefs and, once these are created, selects the decision-making system as the system into which to feed these pretend desires and beliefs as inputs. So there is a higher-order process—the simulation control

process— and a number of lower-order processes, such as the decision-making process. Either of these types of processes might be implemented by a tacit theory. … In none of these cases would simulation be threatened by theorizing." (Goldman, 2006, p. 43)

A form of theoretical reasoning can be used by the simulator to impute some mental states to her target. The process through which the simulator will come to a prediction of her target will however remain a simulation process by feeding her decision-making system with the pretend mental states that she attributes to her target. Thus, there can exist a combination of theorizing for selecting the pretend mental states and of simulation (ibid, p. 44).

5. Intersubjectivity without mentalization

In the opposite stream of the philosophy of mind there are theories supposing that coordination does not require mentalization. Being acquainted with the others, understanding them is not necessarily mediated by mindreading and the attribution of complex propositional attitudes. On the contrary it relies on a fast and frugal mode of coordination requiring only a perceptual dimension and in many cases it relies on the existence of social devices of coordination and cooperation such as norms or conventions that regulate behaviors and interactions. In the latter case individual behaviors are more easy to interpret and thus predictable. In the following sections we will review what is called the Direct Social Perceptions (DSP) literature and the mindshaping hypothesis. Both accounts suppose that coordination does not necessarily require the attribution of propositional attitudes.

5.1. The Direct Social Perception thesis (DSP)

A branch of social cognition contradicts the very idea of mentalizing and considers social cognition to be a direct interactional matter without resorting to mindreading. This is what is now identified as the DSP (Direct Social Perceptions) literature. In some aspects, DSP, as already suggested, can be identified as a form of low-level mindreading as presented in section 4.2.1.

The 'Direct Social Perception' thesis (DSP) criticizes the very fundamentals of both the TT and ST, i.e. the unobservability of others' mental states. This unobservability postulate explains why mentalizing in the ToM relies on cognitive processes. While many of the studies in social cognition have been carried out to identify the 'cognitive architecture' of mindreading, in the recent years the proponents of the DSP have challenged this dominant trend. They claim to the contrary that others' mental states are directly perceivable (e.g. see Maibom, 2003, 2007; Gallagher, 2004, 2008; Bermúdez, 2005, 2006b; Ratcliffe, 2007; Reddy, 2008; Hutto, 2008, 2009; Gallagher and Zahavi, 2012 [2008]; Zahavi, 2011, 2014).

Two traditions exist in the DSP literature: the phenomenological led by Gallagher, Ratcliffe and Zahavi, and the analytic philosophical by Bermúdez, Hutto and Maibom. In each case – as previously suggested – interpersonal understanding does not necessarily require mindreading. For instance, Bermúdez even argues that the role of mindreading must be reevaluated (Maraffa, 2015). Although the phenomenological tradition and the analytical philosophy have overlapping interests they have not been particularly connected on the topic of social cognition (see Smith, 2003; Gallagher and Zahavi, 2012 [2008], p. 3).

The DSP approach criticizes the three lines of justification of the theory of mind for the unobservability assumption: (i) the metaphysical – i.e. mind-body dualism, (ii) the phenomenological – i.e. 'what it is like' to experience the others' mental states, and (iii) the psychological – i.e. the view that empathy relies on cognitive rather than perceptual processes (de Bruin and Strijbos, 2015). DSP's proponents reject what Hurley calls the "sandwich model", i.e. the view of "perception as input from the world to the mind, action as output from the mind to the world, and cognition as sandwiched in between." (Hurley, 2008, p. 2) The DSP thesis clearly opposes this vision. "We should not fail to acknowledge the embodied and embedded nature of self-experience and we should not ignore what can be directly perceived about others." (Gallagher and Zahavi, 2012[2008], pp. 201-202) From that perspective, mental states can be perceived. Despite this, the proponents of DSP recognize the distinction made between cognition and perception that is emphasized in the sandwich model. They just simply focus on the perceptual dimension of social cognition by arguing that understanding someone else does not necessitate an inferential process (de Bruin and Strijbos, 2015, p. 4).

The argument for the DSP account of empathy therefore runs as follows

> "our primary mode of understanding others is by perceiving their bodily behaviour and then inferring or hypothesizing that their behavior is caused by experiences or inner mental states similar to those that apparently cause similar behavior in us. Rather, in empathy, we experience the other directly as a person, as an intentional being whose bodily gestures and actions are expressive of his or her experiences or states of mind." (Gallagher and Zahavi, 2012 [2008], pp. 202-203)

For Bermúdez, most of empathetic identifications rely on what is called the 'low-level' mechanism of attribution, i.e. either on emotional sensitivity or on mimicry (Bermúdez, 2006a, p. 55). One of the examples he gives for the latter claim is particularly relevant from a game theoretic point of view. Bermúdez refers to the iterated prisoner's dilemma and relies on Axelrod (1984) to explain that in this context of interaction, players merely use the "tit for tat" strategy and do not intend to predict the other players' behavior through a process of mentalizing or mindreading. For him the 'tit for tat heuristic' simply requires mirroring the co-players' previous moves at each round. It is therefore merely sufficient to know the move available for the other players (i.e. either cooperation or defection in the case of the prisoner's dilemma) and to remember their choice at the previous rounds. On these grounds, there is no need to infer other players' mental states but simply to respond symmetrically, i.e. to use mimicry (Bermúdez, 2006a). He goes further, arguing that in reality, much of our – recurring – social interactions are supported by the exploitation of behavioral regularities. It is indeed commonly accepted that the role of imitation in social interaction is very important to facilitate coordination (Lieberman 2007;

Iacoboni 2009). Imitation tends to trigger empathetic response and acquaintance with the other (Chartrand and Bargh 1999; de Waal, 2008; see Kirman Teschl, 2010, p. 307).

Even if this position is highly justifiable in repeated games, in the case of one-shot games, and in particular, of coordination games, this heuristic of behavior can be insufficient. In the latter case prediction can require what is called 'high-level' mentalizing, i.e. the attribution of mental states to others in order for a player to adopt in turn, a particular strategy instead of another one.

Embodiment in the DSP thesis is of particular importance (see Clark, 2008; Gallagher and Zahavi 2012[2008], 2011; Chemero, 2009; Shapiro, 2011; Zahavi, 2011, 2014). The notion of embodied cognition within cognitive science began to gain importance in the 1990s. For instance for some scholars such as in Varela, Thompson, and Rosch (1991), Damasio (1994), or Clark (1997), cognition cannot be disembodied. Decisions and actions, or body language occur in a context that helps to delineate what has mattered before them and what was intended after them. The context or environment of an action is never neutral.

> "The environment, the situation, or the pragmatic context is never perceived neutrally (non-semantically), either in regard to our own possible actions, or in regard to the actions and possibilities of others. … we see things in relation to their possible uses, and therefore never as disembodied observers. Likewise, our perception of the other person, as another agent, is never of an entity existing outside of a situation, but rather of an agent in a pragmatic context that throws light on the intentions (or possible intentions) of that agent." (Gallagher and Zahavi, 2012 [2008], p. 211)

This assumption strongly echoes Schelling's and Bacharach' claim that players' perceptions are primarily triggered by a specific context of interaction. Recall that in Bacharach's formalization of the variable universe games, players must coordinate by choosing the same object in matching games. The specific objects under scrutiny and their characteristics, specificities, etc. determine player's frames, i.e. players' perceptions. The context of the game therefore matters and not only the standard rules of the game. Recall also the example given by Schelling about the paratroopers in which the geography of the location where they are dropped provides clues for coordination (i.e. for meeting up). Actions are meaningful in and for a specific context; without the context, people cannot understand others' actions, expressions, etc. Deprived of their environment, actions do not have any specific meaning. One important point on which the phenomenologists and analytical philosophers representing the DSP account agree, is that direct perceptions plus contextualization is one form of interpersonal understanding. Intersubjectivity is always embedded in a specific game, a specific context, a specific set of persons, etc. Again this echoes Schelling's claim that the solution of a game depends on a time, a place and a set of players.

Nevertheless, although their prime focus is on the direct understanding of mental states through experiential self-reference, both traditions grant that intersubjectivity overlays different forms of interpersonal appraisal that are more complex than this direct access:

> "[I]t is a mistake to consider directness and contextuality as excluding alternatives. One can concede that our typical understanding of others is contextual without endorsing the view that our engagement with others as minded creatures is primarily and fundamentally a question of attributing hidden mental states to them. Likewise, it is a mistake to

> consider directness as necessarily opposed to complexity. Saying that we can be directly acquainted with certain mental states of others is consequently not to argue that the process that allows for this direct apprehension must necessarily be simple. The crucial point, and this is what the term 'direct' is supposed to capture, is that the object of my apprehension, the mental state of the other, is my primary intentional object. It is the state itself that I am facing, there is nothing that gets in the way, and the state is experienced as actually present to me. This is precisely what distinguishes empathy from other, more indirect forms of social cognition." (Zahavi, 2014, pp. 179-180)

Directness works for some mental states but cannot exhaust all of them. DSP has a limited power (Zahavi, 2014). Perceptual and contextual acquaintance with others may not be sufficient and to understand what someone else is feeling or desiring it can necessary to "to consider the larger social, cultural, and historical context" (Zahavi, 2014, p. 168). Thus an informational basis larger than mere perceptions and observations of the other is necessary. This is the case, in particular, when the interactional context is not simple, when for instance, no behavioral regularities and no knowledge of some social conventions exist, when players do not belong to the same community or more specifically when there is no overlap among the different communities to which they belong, etc.

In fact, the limitations of empathy as understood in terms of a direct experiential understanding (provided contextualization) begin when complexity increases: "social cognition does not involve high-level mindreading when the social world is "transparent" or "ready-to-hand," … [but] when we find ourselves in social situations that are "opaque," … we cannot help but appeal to the type of metarepresentational thinking characteristic of intentional psychology." (Maraffa, 2015; referring to Bermúdez, 2005, pp. 205-206-225) In the latter case, grasping the other's beliefs, intentions or goals for instance, requires mindreading.

In fact, in the end, whatever the tradition concerned (RT, ST, DSP) each one finally tends to recognize that empathy, i.e. intersubjectivity, is manifold. According to the context and in particular due to its uncertainty and complexity, various mechanisms involved in empathetic identification can be involved: "empathy [works] along a continuum." (Zahavi, 2014, p. 169)

For the purpose of this thesis, DSP is however relevant for understanding coordination and explaining the role of the focal point, understood as norms or conventions and implying a regular pattern of behaviors. As Bermúdez argues, many interpersonal interactions rely on behavioral regularities. When following these regularities a behavior can therefore be directly understandable. It justifies some readings of Schelling's account of the focal point when the latter are understood in terms of social conventions and patterns of behavior. But as we showed in a previous part of this thesis (cf. chapter 2), the concept of the focal point in Schelling's thought cannot be reduced to a mere convention. As Schelling accurately argues, direct social perception is not enough in many cases (see the section 2.2.2. of the chapter 2). The aim of this section was also to show that coordination, as understood in a broad sense in this thesis, i.e. as requiring the convergence of individual perceptions and beliefs, rests on manifold processes. According to the complexity and foremost the uncertainty of the interactional situation, individuals will appraise the situation and coordinate according to different modes. In some cases relying on behavioral regularities will not bring coordination and a favorable issue. Thus, the experiential acquaintance

and direct perceptions would not be sufficient; instead, real inferences and cognitive processes will be needed.

Our main goal in this thesis is of course the identification and the implementation in non-cooperative game theory of individual cognitive processes in strategic reasoning and the impact of these cognitive processes on the convergence of individual perceptions and expectations required. However achieving this goal must not be at the expense of the acknowledgement that intersubjectivity and empathy which are the basis of coordination understood in terms of convergence of individual beliefs, in the real world, are manifold. Incorporating an effective process for the convergence of players' expectations, as it will be done in the model discussed in the next chapter, will not dismiss the insights of the DSP. Quite the contrary; it will show that they are both compatible and modeled in a common conceptual and formal framework. This presentation of the DSP thesis moreover explains why in the next chapter we will argue that the context, which is a particular type of game, leads to certain specific beliefs and not to others. It explains why some beliefs, and not others, will be selected and attributed to the others, as the context of interaction influences their plausibility. Ultimately we will see that the model proposed in chapter 5 can explain the different forms of coordination (either based on behavioral regularities or cognitive and inferential).

5.2. The mindshaping hypothesis

Contrary to the mindreading approach of coordination due to the attribution of mental states to others, the mindshaping hypothesis entails that

> "coordination problems emerge automatically due to default interaction strategies that most group members are shaped to follow. … Coordination emerges not because individuals are preselected to be good mindreaders but because groups with better mindshaping practices that ensure that most members engage in superior coordination strategies by default have selective advantages over other groups." (Zawidski, 2013, p. 127)

The argument that Zawidski develops to justify mindshaping practices more than mindreading practices relies on an evolutionary argument. He indeed declares that it is because communities which perform better in coordination (in order to accomplish cooperative tasks like hunting megafauna) outperform other communities. This capacity requires rapidity of coordination and good tacit (or explicit) communication which cannot be ensured by complex mental states attribution requiring cognitive resources, computations and thus time and energy (Zawiski, 2003, pp.101-102). Zawidski indeed declares that

> "attributing propositional attitudes should be computationally more demanding than tracking behavioral regularities or mental states that are closely tied to behavior. If any finite sequence of observable behavior is in principle compatible with any finite set of propositional attitudes, given appropriate adjustments to background propositional

> attitudes, then discovering a conspecific's propositional attitudes should be more computationally demanding, requiring more time and energy, than tracking behavioral regularities or mental states closely tied to observable behavior." (Zawidski, 2013, pp. 68-69)

Like Bacharach, Zawidski, assumed that natural selection operates both at the individual and the collective level. And for him, natural selection favors individuals that use directly accessible information (without resorting to the complex attribution of mental states) which allow to quick adaptation to situations. Natural selection thus also favors individuals who are easy to interpret and with whom coordination is facilitated, again without resorting to mindreading and mental states attribution (Zawidski, 2013, p. 69; see also Godfrey-Smith, 2002; Sterelny, 2003). In describing one form of imitation (which is, as described below, a mindshaping practice) that is identified as the "chameleon effect", Zawidski emphasizes that it "involves a direct perception-behavior link, with no mediating processing." (Zawidski, 2013, p. 51 referring to Chartrand and Bargh, 1999) The chameleon effect does not require mindreading but instead "an automatic, unconscious" process of matching (Zawidski, 2013, p. 52). Such a process is therefore much more effective, fast and frugal than mindreading.

Through the mindshaping hypothesis Zawidski proposes another alternative to mindreading to explain acquaintance with someone else and to justify coordination that is less based on psychological mechanisms and more on social mechanisms. This however does not imply incompatibility between these two mechanisms; quite the contrary. They are complementary. This is one of the main messages delivered by both Schelling and Bacharach in their respective contribution. The social dimension that impinges on individuals' reasoning cannot be appraised without its psychological dimension, especially in strategic contexts which involve the attribution of mental states to the other which necessarily involve at least a minimal form of cognitive mechanism.

More specifically, mindshaping is defined by Zawiski (ibid, p. xiii) as "a group accomplishment, involving simultaneously interpretive and regulative frameworks that function to shape minds." Mindshaping is induced by a set of social devices such as norms, conventions or institutions, or the existence of social roles and a set of social practices such as imitation, pedagogy, and "narrative self-regulation" like "self- and group-constituting narratives." (ibid, p. 29)

For instance, social norms regulate individual behaviors and social practices. Everybody knowing and conforming to these norms expects everybody else to act and reason in the same way in the contexts in which the norms apply (see also Hédoin, 2014, 2016). In these cases, people's behavior is therefore easily interpretable. In Zawidski's words the other's mind has been shaped in the same manner as yours, i.e. the others' perceptions, reasoning mode and beliefs conform to the same 'model' as yours. When following norms the way we interpret each other's behavior provides also, and simultaneously, instructions from a normative point of view, i.e. instructions regarding the way we are all supposed to act (Zawidski, 2013, p. 52-53). An important (experimental) literature points out that individuals are intrinsically motivated to abide by the norm, i.e. to adopt the pattern of behavior prescribed by the norms "over and above (and to a substantial degree over and above), what would be predicted from instrumental reasons alone" (Sripada and Stich, 2006, p. 285). The internalization of social norm induces "a lifelong pattern of

highly reliable compliance." (Sripada and Stich, 2006, pp. 285–86) It is in this manner that norms are simultaneously interpretative and regulative frameworks. Following behavioral regularities and knowing that others follow the same behavioral regularities thus makes it useless to resort to mindreading. If we accept that social norms do not require mentalization and the attribution of complex propositional attitudes, as Zawidski (2013, p. 57) argues, and considering that norms govern so many aspects of our social life, the role of mindreading therefore appears considerably reduced in our everyday capacity to coordinate.

Mindshaping induces similarities among people's minds so that they tend to focus on the same features of reality (ibid, p. 86). This makes interpretations and expectations of individual behaviors generally accurate: "human varieties of mindshaping have in common mechanisms for making human minds and behavior more homogeneous and hence easier to predict and interpret." (ibid, p. 29) Mindshaping induces "cognitive homogenization" (ibid, p. 65). Seen from this perspective, individuals ultimately become symmetric reasoners. The proper functioning of norms and institutions generates common understanding and hence symmetric reasoning among people sharing the same institutional heritage (Hédoin, 2014, 2016).

Therefore, according to Zawidski, we are able to understand each other accurately and rapidly because we have been exposed to the same "cultural mind-shaping influences." (Zawidski, 2013, p. 81; see also Apperly, 2011, pp. 29-160) This is reinforced by the fact that "[w]e tend to interact with people who have been subject to the same kinds of pedagogy, who have tried to imitate the same kinds of role models, and who seek to conform to the same norms as ourselves." (Zawidski, 2013, p. 81) Accuracy of prediction relies on intercommunity interaction, i.e. on interactions among people who have been subject to the same mindshaping practices. For instance, referring to "self-constituting narratives" Zawidski (2013, p. 59; referring to Ross, 2007) declares "[they] make social coordination possible: without conformity to publicly known narratives, anticipating an individual's behavior in real-world interactions is extremely difficult." But to facilitate coordination, these narratives must be widely known among the people (Zawidski, 2013, p. 60; Ross, 2007). In other words, it depends on community membership. This echoes Hédoin's community-based reasoning and the concept of community-based salience: the knowledge of publicly known events that ultimately induces symmetric reasoning.

One important aspect of mindshaping practices pointed out by Zawidski which will be of particular importance in the last chapter of the thesis is that they "drastically reduce, and hence make manageable, the space of interactions in which we engage: we restrict the games we can play with each other by adopting culturally afforded narratives as self-regulating regimes. Because such narratives are public, the culturally available games are known by our likely interactants, and coordination is dramatically facilitated." (Zawidski, 2013, p. 58) This again echoes Hédoin's claim with respect to community-based salience which is induced by publicly known events (which are of common knowledge); from these publicly known events players draw common inductive inferences (see chapter 2). This dimension of Zawidski's mindshaping will explain in the next chapter the prior beliefs that the players will hold before the game. This explains the perceptions they have of the game and the beliefs regarding the other players' perceptions, intentions and eventual behavior. Players assume, in first instance that they have been shaped by the same social practices so that they tend to focus on the same characteristic of the game and will thus intend to play the same strategy. But this is a first approximation and players will massage their prior beliefs

to check their coherence. In other words, they will assess if the first similarity hypothesis is credible and whether other possibilities exist.

The previous quotation also means that even if complete homogeneity is not acquired there exists a sufficient degree of homogeneity to induce a space of interaction in which the set of information and the set of beliefs regarding the others' eventual behavior is manageable. This is of particular importance considering that the way we conceive games, like Bacharach, entails an open universe problem. A situation in which, contrary to standard game theory, the set of the possible world is not finite.

One main argument in Zawidski's thesis regarding the importance of mindshaping practices is that without these practices humans would have been incapable of mindreading. Mindshaping is a prerequisite to mindreading (e.g. see Zawidski, 2013, pp. xii-xvi). For Zawidski, mindshaping ultimately allows mindreading, i.e. the attribution of 'full blown propositional attitudes' such as beliefs or intentions, reasoning mode, etc. to other people. It is indeed for him "only with prior, sophisticated mind-shaping, ensuring cognitive homogeneity in populations of likely interactants, can distinctively human mindreading be reliable." (ibid, p. xvi) Zawidski argues that without mindshaping practices humans would have been too cognitively heterogeneous for mindreading to be possible i.e. to lead to some reliable predictions (ibid, p. 65). The key role of mindshaping is therefore to lead to the "reliability of social cognition" with mindreading as a form of social cognition (ibidem).

The problem of the intractability of propositional attitudes attribution is explained as emphasized by Zawidski by the holism problem:

> "The reason propositional attitudes have such a tenuous connection to observable evidence is holism: any finite set of propositional attitudes is compatible with any finite set of observable evidence because a propositional attitude's relations to stimuli and behavior depend on other background propositional attitudes. In other words, as Quine (1960) might put it, an agent's beliefs face the tribunal of experience as a whole, not piecemeal, and an agent's beliefs and desires direct behavior in concert, not individually." (ibid, p. 66)

Therefore without mindshaping the holism problem entails that mindreading and the attribution of mental states to the others is an intractable cognitive task. People must have been shaped by the same social practices in order to become "cooperative and easily interpretable." (ibid, p. 28) The purpose of mindshaping is homogenization and it is only under this condition that sophisticated mindreading can lead to accurate predictions (ibid).

Recall that according to Goldman's simulationist approach, people simulate the reasoning of the others to ultimately predict or interpret their behavior. They use their own mind to model, i.e. to assess, the other player's mind – e.g. their frames, beliefs and reasoning (Goldman, 2006; Goldman and Shanton, 2012). The devices of the homogenization of mindshaping therefore become decisive: in order to be correct this simulation process requires homogeneity among people. Mindshaping accordingly provides the grounds for an accurate fast and frugal procedure of mindreading in simulation (Zawidski, 2013, p. 81). As Zawidski emphasizes "the accuracy of such simulation heuristics obviously depends on extreme homogeneity in human populations:

interpreters and their targets must prioritize problems in similar ways and make decisions based on similar information and heuristics." (ibidem)

Thanks to mindshaping, simulation becomes tractable from a computational point of view (ibidem). As people first tend to attribute to the others their own frames or propositional attitudes, this limits "the search space" of the eventual propositional attitudes – e.g. beliefs – that the other can hold (Goldman, 2006, p. 184).

6. Conclusion

Dating back to the Scottish enlightenment (e.g. see Hume (1958[1738]), empathy was seen as an epistemic device allowing individuals to tacitly communicate. Empathy provided the grounds for an intersubjective understanding, playing an important role in interaction. It was a testimony of the vision of economics as an interactional system, in which the intersubjective dimension was inescapable. This conception of empathy has however been dismissed, raising problems for neoclassical economics in terms of subjectivity and in terms of the need for psychological theories to impinge on the realm of economic theory and game theory. On the one hand, empathy and sympathy have been conflated and integrated in individual utility function as a means to rationalize other regarding behavior, or on the other hand to eventually formalize interpersonal comparison of utility, but without psychological foundations. The latter use led to considerable epistemological issues in welfare economics. However if we agree with Kirman and Teschl (2010) that the new paradigm of neuroeconomics, by providing data on players' reasoning, can bring considerable insights for reassessing individual decision-making in both decision theory and game theory, empathy can again be at the core of interpersonal interactions.

It is true that for scholars who consider economics as a closed system and attempt to deprive economics of all the determinants resorting to its neighboring social sciences, returning empathy to the realm of economic theory is an issue. However, if taking a broad view of economics and considering that economics cannot be appraised independently from the other social sciences, empathy offers valuable methodological improvements for economics and even more for game theory. If the methodological issues of modern non-cooperative game theory are taken seriously, a close look at these methodological issues shows that empathy offers manifold solutions.

In non-cooperative game theory the intersubjective dimension of strategic interactions is usually defined by common knowledge of instrumental or Bayesian rationality. It allows players to form beliefs on others' beliefs and choices and therefore to predict their choices, but as has been shown in the previous chapters of this thesis this methodological device is insufficient in many games and coordination contexts to lead to clear recommendations in terms of individual strategy and to ultimately provide determinate solutions for games. As emphasized in Chapter 1, many critics have been raised on the way individual beliefs are formalized in game theory. The conception and very fundamentals of the use of prior beliefs have been numerously contested. Those prior beliefs are defined prior to the game by the theorist, provided a given solution concept, a given concept of rationality and a given objective informational basis are provided by

the definition of the game. The subjective dimension of these beliefs has accordingly been questioned. Besides, priors are considered to be the rational beliefs that the players can hold provided the given informational basis, and the justification of this claim relies on Bayesianism. Again this premise has led to considerable debates, if it is acknowledged that Bayesianism and Bayesian rationality provide a clear and potentially – from a normative point of view – non-challenging account for players' revision of beliefs, it is much less guaranteed that players' priors can be defined by Bayesian rationality. Some scholars highlight that even defining those priors as rational is due to Bayesian rationality and assuming Bayesian revision of beliefs in strategic contexts leads to inconsistency and therefore to the fall of strategic rationality. Ultimately it is claimed that players' prior beliefs and more generally their set of beliefs are solely mathematical artifacts and are not mental states (Gul, 1998).

Empathy as assessed and defined in Goldman's account of ST provides one possible methodological solution for all of these shortcomings. Indeed, it offers a methodological device to define players' beliefs from their own perspective not from the perspective of the game theorist as an outside observer. If we accept seeing coordination in a broad sense, i.e. as understood in terms of convergence of players' perceptions, intentions, and beliefs, and taking into account the players cognitive processes leading to such state of convergence, empathy in its epistemic understanding becomes a core determinant of individual decision-making in strategic interactions.

Empathy as understood by the standard ST provides an explanation of the way players form their own perspective, i.e. from their own perceptions of frames appraising the others' perceptions, intentions and beliefs. They simply assess these mental states from their own and simulate what the others' mental states may be. Ultimately, through this cognitive process they can predict the others' decisions. The set of these frames, and beliefs attributed to the others is circumscribed by the players' own beliefs and frames, and by the context of the game, i.e. the particular game under scrutiny. Some perceptions and beliefs will indeed be considered as non-relevant under the conditions faced by the players, again according to their own perceptions. From this perspective ST explains at the same time the mechanism of attribution (like the TT and RT) and the content of the mental states or perceptions attributed (which the TT and RT hardly provide). Besides, in the ST, there is no ambiguity about the extent of information needed by players to infer the others' mental states while in the RT and TT it is remains in question. In the ST, players clearly use their own epistemic states. They use their own decision system and mental states as an informational basis, or in other words, as a model of the others. The mindshaping thesis ensures that a sufficient degree of homogeneity among players exists so that using our own mental states and decision-making system as a model of the other's ones generally provides reliable predictions.

The ST also shows that there is a wide range of mechanisms involved in interpersonal understanding (contrary to the TT and RT). Perceptual dimensions, face-to-face encounters and the recognition of emotional signs matter in the ST account of mindreading for being acquainted with someone else, as is also supposed in the DSP literature. All of these dimensions can matter for successful coordination. Interactions are necessarily embedded in a specific context: this is one of the main messages of the DSP. Cognition is situated. This contextual dimension will play an important role in the next chapter as perceptions, which are necessarily situated (i.e. embedded in a specific context), are integrated in games.

ST justifies the use of and the reference to focal points in game theory. On the one hand ST is easily accommodating to frames and the salience of frames, and in this perspective to endogenous focal points – i.e. as understood in terms of salient subjective perceptions. Namely, what is salient or obvious for players, will be attributed to the other players; what comes into their mind first will be supposed to instantiate first in the others' mind, and in these pretend states that players feed into their decision making system, it is very likely that they will put in those salient states. This assumption is justified by the mindshaping hypothesis. As mindshaping brings homogeneity among people, and as mindshaping practices serve both as interpretive and regulative frameworks, it makes perfectly sense to assume that the others will hold the same perceptions and intentions as ours. On the other hand when referring to a social focal point, i.e. to an exogenous focal point, ST explains why in the attribution of mental states like intentions, beliefs and in their prediction of the others' behavior, the players will refer to the ones entailed by the focal point. The Focal point understood as a social norm or a convention entails community-based reasoning, i.e. common perceptions, common intentions, common behavior and common inductive inferences (see Hédoin, 2014, 2016). According to a specific interactional context, the existence of a focal point understood in terms of a social convention will instantiate specific mental states that will be then attributed to the others. This is ensured by community-based reasoning which states that people come to the same inductive inferences (Hédoin, 2014, 2016).

Empathy as understood by the ST is compatible with a conception of sociality in which convention and regular pattern of behaviors matter, as social coordination devices can orient and influence the process of mindreading and of mental states attribution. Attribution and accordingly empathy is mainly grounded on introspection and on self-reflection. The personal identity and subjectivity of the subject does not break down. But such subjectivity does not prevent being acquainted with the others. Mindshaping is of prime importance in this perspective as it ultimately brings homogeneity among people who have been shaped by the same mindshaping devices, i.e. the same norms, conventions, social roles, pedagogies or group-constitutive narratives. This compatibility between players' personal experience or personal identity and collective identity or community membership provides one of the justifications of the implementation in game theory of Goldman's view of ST and not Gordon's view. The main justification is, however, that without supposing common knowledge rationality (or common belief in common rationality) as is usually supposed in incomplete information games, it is possible to define the set of beliefs that the players hold with respect to others' beliefs and choices. The other approaches are insufficient to define the players' beliefs in the same manner without any additional information.

However it is true that bringing empathy into game theory requires the acknowledgement that economics cannot be conceived independently from the other social sciences since this mechanism of interpersonal understanding is at the core of the interdisciplinary cognitive sciences. It requires a conception of economics as an open science (see chapter 2 for an explanation of economics seen as an open science). It also implies that the players' cognitive processes must not be considered as mere black boxes, contrary to standard and epistemic game theory and that any account of game theory should not be immune to an inquiry into such black boxes. From this perspective, it seems worth quoting Singer and Fehr (2005, p. 344) who declare that in epistemic game theory "the question about the determinants of [players'] prior probability

distribution has not been addressed. In fact, the assumption of a prior distribution over types constitutes a huge black box." As shown in this chapter the ToM opens this "huge black box" and the ST more specifically provides an explanation of the determinants of this prior probability distribution, i.e. of the prior beliefs that the players handle. The next chapter will more specifically demonstrate how it explains the set of prior beliefs that the players handle using the ST.

Chapter 5:

On the use of mindreading and mindshaping in game theory: how to incorporate players' mental states and to endogenize players' beliefs

1. Introduction

This chapter presents a theory of games in which the mental states of the players have been incorporated and in which a theory of the formation of players' beliefs is proposed by relying both on mindreading, and more specifically on the Simulation Theory (ST) of Goldman, and the mindshaping hypothesis of Zawidski (both presented in the previous chapter). In line with Kadane and Larkey (1982a)'s suggestion, the object of this chapter is to offer a psychological theory of the formation of the players' prior beliefs. Such theory of beliefs formation is 'vital' in Bacharach (1993, p. 258)'s words for game theory. As demonstrated in the thesis, this is from that perspective only that the explanation of how and why an equilibrium occurs can be provided: this is on that perspective only that the player's reasoning process toward the equilibrium can be explained. This chapter thus tackles the issue of the conditions under which the players' representations and beliefs can converge on an equilibrium so as to ultimately coordinate. It provides an explanation of the existence of a particular equilibrium. In that perspective we will integrate in the model of games proposed in this chapter some of the enhancements of standard non-cooperative game theory proposed by both Schelling and Bacharach. More specifically, we will introduce, like Schelling and Bacharach, and thanks to framing, personal and socio-cultural determinants in the players' reasoning process.

The integration of players' reasoning process and of their mental states like their perceptions, intentions and beliefs as already mentioned in the chapter 2 and 3 of the thesis requires considering that the strategic decision problem they face is no longer a closed system but an open system in which the standard Bayesian decision theory is no longer applicable. We have therefore to formally deal with games that are not described by closed states spaces (i.e. by situations in

which all the possible consequences of the players' choice are finite. We rely on the distinction made by Savage between small and large worlds to explain how players transform open decision problems into more tractable problems in which they can decide and act. The explicit introduction of the processes by which the players represent real and complex decision problems as simpler, more tractable, ones, however requires tackling the problem set aside by Savage (1954) of developing a theory of choice in *large* worlds (in his terminology, i.e. in open decision problems). The approach we choose to close the universe will conceptually rely on Bacharach's variable frame theory, though the formalism will be slightly different to better catch Savage's distinction between small and large worlds. In a nutshell, we will consider that some features of the decision problem will be more *salient* than others, and the players will be 'guided' by those salient features to frame the initial large world as a tractable problem – a small world – for which Bayesian methods can be applied.

Humans routinely coordinate on complex activities, whether it be playing a matching game in a lab experiment, driving, working within a team, co-authoring an academic paper, or meeting in a crowded place such as a train station or during a concert. An explanation of this ability to coordinate with others is that we share a capacity to correctly represent each other's mental states and beliefs, and can thus accurately anticipate the behaviors of others. Different theories of *mindreading* – the capacity to 'read' each other's minds – have been suggested in the literature,[89] though they systematically start from the principle that the key to human coordination is our ability to form complex epistemic states about each other's mental states, especially beliefs and desires (Zawidski, 2018). This is also the approach implicitly endorsed in the epistemic program in game theory, according to which the analysis of strategic behaviors can be reduced to Bayesian decision theory, while taking hierarchies of beliefs as an input of decision-making.

There are however some methodological issues with the project of developing rigorous Bayesian foundations for game theory.[90] We can in particular question the usefulness of an approach that requires players holding complex beliefs about each other's actions and beliefs, without explaining *how* players could form such beliefs. The lack of a theory of belief formation implies that Bayesian game theory – just as classical game theory – is unable to explain successful coordination in simple matching games for instance. Without an explanation of the elicitation of players' beliefs we are condemn to assume, as in epistemic game theory, a behavioristic interpretation of beliefs which implies as explained in the chapter 1 that the beliefs held by the players are equilibrium beliefs, once their equilibrium choice is made. It is thus impossible to explain coordination as *de facto* the game describes the choice of the players in which coordination is a priori excluded. The main issue of Bayesian game theory is indeed that the beliefs that the players are supposed to use as an *input* of their reasoning, are by construction the beliefs at the equilibrium, i.e. the *output* of their reasoning (e.g. see Lecouteux 2018a).

---

[89] The two main approaches are the 'theory-theory' and the 'simulation theory'. See Larrouy and Lecouteux (2017) and Guala (2018) for a brief presentation and a discussion in a game theoretical context.

[90] In addition to the points discussed here, see also Mariotti (1995, 1997) on the problems of extending Savage's axioms to game theory.

One possible strategy used by game theorists to solve coordination puzzles is to suppose the existence of *focal points*, and to suppose that people are more likely to believe that the focal point is the 'correct' choice. Apart from a few attempts to develop a rigorous analysis of *salience* and *focal points* (Bacharach 1993, Sugden 1995, Casajus 2000, Janssen 2001), those notions remain loosely defined and merely offer an *ad hoc* explanation. The conceptual issue here is that game theory is 'an internally closed procedure which operates according to fixed rules' (von Neumann, 1983 [1931], pp. 61-62, quoted in Giocoli, 2003, p. 15), and therefore that all the relevant information for the players *must be* included in the mathematical description of the game. Developing an empirically robust theory of games would however require including the processes by which the players use the 'background', of the game to choose their strategy, i.e. would require including the role of the *context* in game theoretical analysis, and more generally within this context the personal and social background of the players. When considering decision problems faced by actual players, the game is necessarily embedded in a particular context (a lab experiment, a specific social setting, etc), and – while the game as a mathematical object aims at capturing the objective strategic features of the interaction only – the context will provide clues to the individual about the choice to be made (i.e. drive on the left side of the road in England, but on the right side in France). This implies that a non-purely mathematical account of game must be considered. The objective and mathematical characteristics of the game, as claimed by both Schelling and Bacharach are insufficient to explain coordination. The role of the 'context' in decision making is extremely intuitive – both as a way to intentionally discriminate between alternatives in matching games, but also as an unintentional factor that impacts our decisions (see e.g. Kahneman and Tversky (2000) on framing effects) – though remains a fundamentally vague theoretical object, as the background in which the game takes place is necessarily *external* to the game understood as a closed and mathematical object. Referring to such external factors (such as norms of language in the description of the game in Bacharach (1993)'s 'variable universe games') has two methodological consequences: it implies (i) that players' mental states – mainly their perceptions and beliefs – must be included in the description of a game and (ii) that the choice problem faced by the players is 'open' rather than 'closed' or 'completable' (Binmore, 1988, 1993). This distinction between a closed and completable and an open decision problem will be formally appraised by Savage's distinction between a small and large world that will presented in section 2.2 of this chapter.

One of the aims of this chapter is thus to develop a theory of coordination that considers the role of the *context* within which the individuals interact and in such context their personal and socio-cultural background. In particular, we will emphasize the role of mindshaping in the formation of a common intersubjective background among the players, allowing them, *in fine* to identify and coordinate on the same focal points. An alternative explanation for human coordination that has indeed been highlighted by cognitive scientists is the process of mindshaping, according to which it is the ability to *shape* each other's mind that is the key to our capacity to coordinate. Recall that Zawidski (2013, p. xiii) defines mindshaping as involving both "interpretive and regulative frameworks" such as norms, conventions, or institutions, that shape minds. The aim of such devices is to generate a 'cognitive homogenization' (Zawidski, 2013, p.65), so that in addition to regulate individual behaviors the latter ones become easily interpretable. From that prospect, successful coordination is not the result of a highly intellectualized process that requires people to solve the complex and sometimes intractable

problem of correctly inferring the mental states of others. It is rather because we share some common perceptions *ex ante*, and that we have been exposed to the same norms and conventions, that we can easily anticipate each other's behaviors, based on our own experiences. Mindshaping – by shaping our minds in similar ways – is a prerequisite to successful coordination, since it guarantees that our own mental states (whether they be experiences, perceptions, reasoning mode or beliefs), are likely to be similar to the mental states of the other individuals with whom we interact. Besides as we will show in section 4 the set of equilibria in our model of games are quite large when only considering simulation although such a set can be considerably reduced by mindshaping. Although mindreading may offer a justification of the formation of common belief of events and rationality of the players (see proposition 1 in section 3 of this chapter), it remains insufficient to determine unambiguously to which beliefs the players will *in fine* converge.

Recall that a crucial issue of open choice problems is that we cannot *a priori* rely on the tools of subjective expected utility theory. Traditionally, expected utility theory is founded on an axiomatic derivation of probabilities and utilities from a set of observed preferences over acts (e.g. Ramsey, 1926; Von Neumann and Morgenstern, 1953; Savage, 1954, Anscombe and Aumann, 1963; Aumann and Drèze, 2009). Those systems of axioms are designed such that they lead to a unique class of state-independent utility functions and probability distributions, which rank the different acts according to their expected utilities. Von Neumann and Morgenstern (1944) indeed showed that, if the choices of a player respect certain axioms, then it is *as if* the player was maximizing an expected utility function. Similarly, Savage (1954), Anscombe and Aumann (1963), and Aumann and Drèze (2008) show that, if the choices respect certain formal conditions of consistency, then we can define a utility function and subjective beliefs such that the action that maximizes the expected subjective utility of the player is precisely the choice we observed. The primitive of analysis is therefore the *choice* of the players, and the utility is defined *ex post*, as a representation of their behavior. In this framework, preferences are called 'behavioristic'. Although preferences are a central concept in economics, there still exists some confusion about what preferences *actually* are. Heidl (2016, pp. 26-44) indeed suggests that preferences can be interpreted either in a *mentalistic* or in a *behavioristic* way. Recall that according to the mentalistic interpretation, 'preferences are understood as scientific refinements of the folk psychological concepts of desire and preference' (Heidl, 2016, p. 26), while the behavioristic interpretation is that 'preferences are not mental entities but consistent patterns of choices' (ibid, p. 26). The mentalistic interpretation therefore leaves space for the integration of the players' reasoning process – which implies integrating a psychological theory explaining the formation of the players' subjective beliefs – while the behavioristic interpretation understands payoffs in games as Von Neumann Morgenstern (thereafter vN/M) utilities rather than material payoff,[91] for which the psychological dimension has been erased (Hausman 2000, p. 115). Defining the payoffs in games as vNM utilities was an attempt to get rid of psychological and unobservable

[91] We assume here for simplicity that the preference ordering corresponding to the mentalistic interpretation (one's material payoff) is the same than the ranking of monetary gains: a more precise definition would however include other-regarding preferences (in the sense of Vanberg (2008), as preferences over outcomes rather than actions), such as sympathetic concerns for others, rather than restricting mentalistic preferences to the player's self-interest.

variables such as the (subjective) tastes and beliefs of the individuals (Sen, 1973, 1987; Bacharach, 1986; Hausman, 2000, 2012; Lehtinen, 2011; Heidl, 2016). Nevertheless, and following Heidl (2016, p. 29)'s claim we argue that "preferences cannot be defined purely behavioristically by choice, but necessarily involve a reference to the mental states of agents". Behavioristic game theory indeed cannot explain how players compare different alternatives (and therefore cannot explain why selecting a particular strategy profile would be rational). As Binmore (2007, p. 12-13) highlights it is tautological to claim for instance in the Hawk-Dove game, that each player will choose *Hawk* because the payoffs of this game are such that it is already assumed that *Hawk* maximize the players' subjective expected utility.[92] Yet, from a positive perspective, when playing a game each player should be able to compare the desirability of each potential outcome to determine the strategy that will maximize her expected payoff. In the same way, each player should be able to make probabilistic assessment about each possible choice that the other players may make. Nevertheless by definition only equilibrium play are allowed in games. Players thus cannot assess what can be their equilibrium play by assessing all of the possible states of world that may be brought about by the combination of each player's choice. In other words, counterfactual reasoning is precluded.

More importantly, players' beliefs cannot be appropriately defined since in a behavioristic account of games, the individual choices cannot reveal preferences in contexts of uncertainty without any consideration about beliefs (Hausman, 2012, p. 30), the latter must already be assumed in order to define the games, i.e. to state the matrices. In this perspective, Hausman (2000) speaks of "belief-dependent revealed preference theory". Indeed, in the revealed preferences interpretation of payoffs as vNM utilities, "the payoffs … would say how individuals would choose. They would already incorporate the influence of belief, and belief could play no further role. If the revealed-preference theorist were right and payoffs already represented what strategy was chosen, there would be nothing left for game theory to do." (Hausman, 2000, pp. 111-112) Besides, since the player's beliefs are defined according to the priors they handle before the game (e.g. see Bacharach, Hurley, 1991; Morris, 1995; Gul, 1998; Gilboa, 2011; Grüne Yanoff and Lehtinen, 2012), as early emphasized by Bacharach and Hurley (1991, p. 26):

> "What brings me to have the prior probabilities that I do for your deciding on one option and another is a question not answered (and rarely asked) by the Bayesian theory of games. The absence of an independent account of what is in the players' priors is a grave lacuna. There are many games for which, once the priors are given, the identities of the rational acts follow trivially, and then game theory itself is trivialized if it is merely assumed that the prior are such and such. To avoid this trivialization by Bayesianization, we must take the content of the priors in such cases to be the central unknows of the theory, endogenous to it."

The approach to subjective expected utility developed in this chapter thus significantly differs from this traditional approach, as we do not treat choices but *payoff* as the primitive of analysis. To avoid any confusion with the 'utility' defined *a posteriori* from individual preferences over acts,

[92] See also Lehtinen (2011) for a detailed discussion of Binmore's example. For other examples, see (Hausman, 2000, 2012).

we will only use the term 'payoff', defined as 'normalised measures of the values of the relevant outcomes to [the player], *in terms of her own interests, as judged by her*' (Sugden 2015, p.144, emphasis in original). Unlike utility, which merely describes the preferences of the player, the payoff is defined as the *cause* of the preferences of the individual – I prefer the act *f* over the act *g*, because the associated expected payoff is higher[93] (while the expected utility of *f* is higher than *g*, *because I prefer f*).

Taking payoffs rather than choices as the primitive of analysis means that we must investigate how players choose for given payoffs, i.e. how they form their beliefs and intentions. This requires, as already emphasized, a significant departure from the usual Bayesian analysis of strategic interactions, since this kind of problem cannot be considered as a 'small world', which is the only situation – as argued by Savage (1954), and more recently by Binmore (2009) – to which Bayesian decision theory could be applied:

> "[t]he models constructed by game theorists are small worlds almost by definition. So we can use Bayesian decision theory without fear of being haunted by Savage's ghost telling us that it is ridiculous to use his theory in a large world. Bayesian decision theory only ceases to be appropriate when attempts are made to include the minds of the players in the model to be considered" (Binmore, 2009, pp.134-135)

For sake of simplicity after having presented in section 2 the issue of dealing with open decision problem and explaining how to transform complex decision problems, i.e. decision problem in large worlds (in which the set of the possible consequences of our choice is not finite) into more simple and tractable decision problems, i.e. in small worlds (in which the set of the possible consequences of the choices is finite), we present in section 3 and 4 a model of games in small worlds. Section 3 and 4 formalize, the way players rationalize their hierarchy of beliefs through a massaging process thanks to simulation, and once their prior beliefs have been to a certain extent homogenized by mindshaping. We thus start by formalizing the rational strategic reasoning process of the players (in section 3 and 4) before extending the analysis to the unconscious and selective aspect of player's priors understood as gut feelings. We extend the analysis of games to large worlds and resort to mindshaping as the evolutionary approach explaining how players select their beliefs about players' perceptions and beliefs prior to the game.

2. Coordination games as 'open' decision problems

We begin with two illustrations (a standard matching game in Paris, and the Brexit negotiations) to highlight the main features of our analysis of games as 'open' decision problems. We then

[93] An important implication of defining payoffs this way is that they can be observed *prior to the choice*. This means that we can experimentally test the theory that players behave as if they intend to maximise their expected payoff, while it is tautologically satisfied – by definition of the utility function itself – in the behaviouristic approach (see Lecouteux 2018b).

recall Savage's definition of small and large worlds, and highlight the role of focal points as a factor of cognitive homogenization, allowing the players to adopt the same representation of a large world.

This section thus captures the two main problems to explain when games are no longer closed and mathematical representations of players' choice but the representation of the way players appraise their decision problem. It presents: i) how to frame the decision problem from a large world to a small world, this is explained in the first illustration: the Brexit negociation problem; and ii) how we choose as soon as the small world is defined, this is explained in the second illustration: the meeting point in Paris.

2.1. Two illustrations of open decision problems

2.1.1. Brexit negotiations[94]

The example of the Brexit negotiations illustrate how players may frame the initial problem into a much more tractable one. The idea is here to show that – unlike most of game theory that starts with already abstract situations (like meeting a random individual in Paris in a week from now) – the processes we intend to model also govern 'real' interactions. It will also give us the opportunity to introduce the main ingredients of our formal model to translate 'large worlds' into 'small worlds'.

Consider the strategic interaction between the UK government and the European Union about the nature of their future relationship after the Brexit. The precise and exhaustive description of the set of actions (i.e. the details of the possible agreements between the EU, the UK, and third parties, the timing of the negotiation and the possible impacts of national elections, etc.) for both parties is far from being cognitively possible. A way of framing the problem for the two parties is to consider a restricted set of acts, and to focus only on very few features of the consequences of those acts. The UK may for instance consider that only two 'properties' of the possible outcomes are relevant from its perspective: the UK economic performance and migration flows (as this were the two central themes during the 2016 referendum campaign). The UK may also believe that the EU will evaluate the outcomes mainly according to one property, 'economic performance of the EU'. In the game imagined by the UK, the set of possible outcomes is thus framed according to those combinations of properties. This will also determine how the UK perceives the available strategies for both players: the UK will propose an agreement to the EU, and the EU may accept or reject it (the 'no-deal' scenario is detrimental for the economic performance of both players, but is a good outcome in terms of migration flows for the UK). Among the set of agreements that the UK could offer (its set of strategies), the key elements will concern the two relevant properties for the UK, i.e. protecting the free trade between the UK and the EU (with a good outcome in terms of economic performance) while limiting the

[94] This section have been written before the effective conclusion of the Brexit.

migration from the EU to the UK. The UK can represent its available strategies as (i) offer a free trade area, (ii) stay in the Single Market, or (iii) no offer (with a no deal scenario). Assuming that the EU only cares about its economic performance, it is in the EU interest to accept any offer, and to avoid a no-deal scenario. The UK can thus confidently offer only a free trade area, and wait for the positive reply from the EU – this was broadly the rhetoric of the 'Leave' campaign during the referendum, because it was argued that it would be in the interest of the EU to stay in good terms with the UK after the Brexit.

The negotiations (from the UK's perspective) are thus framed as a game in which they can choose between three strategies (free trade area, single market, no offer) and the EU between two strategies (accept or reject the offer). Since rejecting the offer seems to be a weakly dominated strategy for the EU, the best strategy for the UK is to offer a free trade area and to limit migration flows. If, however, the EU also evaluates the outcome according to the property 'deterrence' (on future exits from the EU), then a scenario of 'no-deal' may be acceptable by the EU, since an economic failure following Brexit could dissuade others countries from leaving the EU. Depending on which property (economic performance of the EU, or deterrence) the EU most favors, the evaluation of the outcomes may be radically different from what is framed in the UK's game representation, in which only the economic performance of the EU is considered by the EU.

2.1.2. Meeting in Paris

Suppose that you are asked to meet someone (about whom you have no additional information for the moment – she could be a tourist, a colleague, a friend, etc.) in Paris in exactly a week from now at noon. Where will you wait for the other? This is a standard 'matching game', in which you intend to go to the same location than the other individual, whatever this place is.

Probably one of the first meeting points – if you are not a Parisian – that spontaneously came to your mind was the Eiffel Tower, which is the most iconic place of Paris. Note that you can confidently feel that we would *all* recognize the Eiffel Tower as a focal point for a meeting in Paris (including the other individual you have to meet). This is because, as part of the community of 'non-Parisians' (with a limited knowledge of the city), it seems reasonable to assume that we would all feel the same than how you felt when thinking about a focal point in Paris. The fact that you identified the Eiffel Tower as focal may give some evidence about the focal point that is identified by the other members of the community of non-Parisians.

Suppose now that you live and work in Paris: like most Parisians, the Eiffel Tower is not particularly salient for you anymore, and a more natural meeting point would typically be the station Châtelet Les Halles, the main commuter train hub in Paris. From the perspective of a Parisian, this station stands as a more obvious solution to the game than the Eiffel Tower. However, as many Parisians, you also probably have a favorite café for meeting your friends after work, another for having brunch or for drinking your morning café before going to work, a favorite bistro, etc. Depending on the context (whether you have to meet the other in the morning, at noon, after work, during a weekend, etc.) you may spontaneously identify different places as the most salient ones to meet the other individual.

Given that you must meet the other individual at noon, this particular context may trigger the idea that you should meet in the restaurant you usually go for lunch (for instance the one close to your workplace if it is a weekday). This specific restaurant is then of 'primary salience' for you (Metha *et al.* 1994a,b; Bacharach and Bernasconi, 1997), it is the obvious meeting point that spontaneously comes to your mind. We will define later in the paper the primary salient option as an *individual focal point*.

However, without any additional information about the other, there is little chance that she will also identify your favorite restaurant as a focal point. This means that you have to anticipate how the other will frame the game, and what focal point she is likely to identify. The element of the context that seemed important for you (it is lunchtime, so a restaurant is a good meeting point) may not be the element that the other will consider as important. This means that you have to 'think about salience strategically' (Bacharach and Bernasconi, 1997, p. 39). If it turns out that the other individual is one of your colleagues with whom you regularly go to this restaurant for lunch, then it seems reasonable to think that going to the restaurant is primary salient for her too. But if you simply learn that the other individual is a Parisian (without any additional information), then you have to think about what is of primary salience for a 'Parisian': here Châtelet Les Halles seems to be a good candidate. By supposing that Châtelet Les Halles is of primary salience for the other (as a Parisian), it becomes of secondary salience for you (Bacharach and Bernasconi, 1997, p. 38). We will define later such secondary salient option as a *subjectively based social focal point*.

We could continue the illustrations with many other cases. For instance, if you have no information about the other (you are thus in a complete uncertainty) – she could either be the colleague with whom you regularly go for lunch or a foreign tourist – it seems safer to restrict the set of possible meeting points to what is shared by virtually everybody, i.e. the Eiffel Tower. If you both identified the same meeting point as a subjective social focal point (i.e. the Eiffel Tower with a tourist, Chatelet les Halles with a Parisian, the restaurant with your colleague), then we will call such point an *objective social focal point*.

The existence of objective social focal points (such as the Eiffel Tower as a default option, Chatelet les Halles for Parisians, the restaurant for colleagues) is the product of mindshaping, *via* routines and recurrent interactions with others. It is because many Parisians meet at Châtelet that it became focal for them, and because you usually go with your colleagues to the same restaurant that it is also focal for your smaller community. The initial problem, which was of an infinite complexity (they are uncountable meeting points in Paris), became almost trivial simply thanks to the fact that both you and the other individual belong to a same community ('everybody', 'Parisian', or 'colleagues'), with a set of shared perceptions and norms. As emphasized by Schelling (1980[1960] pp. 57-58):

> "A prime characteristic of most of these "solutions" to the problems, that is, of the clues or coordinators or focal points, is some kind of prominence or conspicuousness. But it is a prominence that depend on time and place and who the people are."

A problem that remains to be solved, however, is the question of the rationalization of focal point play: although it seems 'obvious' that a focal point is a perfect guide to coordination, we still lack Bayesian foundations for such type of play. One of the objective of the remainder of this

chapter – in addition to offering an approach to model how players frame the initial problem into a simpler one – will be to argue that the most salient option in the game (as we think is commonly perceived by the members of the community) can be integrated as our 'gut feeling' regarding the outcome of the game, which constitutes a basis for the prior belief we will use when solving the game.

Our framework aims at capturing the main features discussed here. Each player will frame first her set of strategies (based on the 'salience' of some acts). They will then try to frame the set of strategies of the other player, by a process of *simulation* (by imagining how they would represent the set of strategies, if they were in the other's shoes). Given the definition of the set of strategies for each player, each strategy profile will be assessed on the basis of various 'properties', over which the agents will have a preference relation (e.g. the EU may prefer the property 'deterrence' over 'EU economic performance'). We will model this preference over properties thanks to Mandler et al (2012) model of choice by checklist.[95] Given their set of strategies, and the evaluation of the outcomes, we obtain a game representation (a *small world*, in Savage's terminology) in which the only uncertainty concerns the strategy of the other player.[96] We will then assume that the players form their beliefs about the choice of the other by simulating their reasoning, as in section 2 of this chapter

2.2. Small worlds, large worlds, and the grand world

We slightly adapt Savage's notations to define large and small worlds. We consider a set $S = \{s, s', \dots\}$ of *states of the world*, a set $\mathcal{C} = \{c, c', \dots\}$ of *consequences*, and define an *act* as an arbitrary function $\boldsymbol{A}: S \mapsto \mathcal{C}$ that assigns consequences to states of the world. We denote by $\mathcal{A}$ the set of acts. A *world* is a triplet $W = \langle S_W, \mathcal{C}_W, \mathcal{A}_W \rangle$ that describes the set of possible acts $\mathcal{A}_W$ and their consequences in $\mathcal{C}_W$, given the possible states of the world in $\mathcal{S}_W$.

A world W is *smaller* than a world W' if and only if the set of states of W is a partition of the set of states W' (W' is then a *larger* world than W). For instance, if the states of *W* describe the range of possible temperatures in Paris at noon – which may go from -25°C to 40°C – a smaller world *W'* can for instance be composed of the two states 'below 0°C' and 'above 0°C'. The two states below and above freezing indeed form a partition of the range of possible temperature in Paris at noon.

---

[95] A conceptually related framework would be Dietrich and List (2013) model of 'motivationally salient properties', with a restriction to lexicographic weighing relations over property combinations, as in de Jongh and Liu (2009).

[96] Note that, in the process described above, it is very likely that the small world representation of a player may fail to take into account some important features of the problem. In the case of Brexit, we can mention for instance the property of 'deterrence' – that the UK fails to recognise as salient for the EU – or the 'status of Northern Ireland', which has been set aside in our small world representation. The players may also be unaware of some strategies (e.g. finding a new type of arrangement different from merely remaining or leaving the single market).

Each element $s \in S_W$ offers a more or less precise description of the state about which the decision maker is uncertain. 16°C is for instance more precise than 'above 0°C', though it could have been even more precise if the state described the distribution of temperature at the different weather stations in Paris. Note that we can almost always add arbitrarily detailed elements to the description of the state of the world (e.g. describe the temperature in Paris by the distribution of temperatures in Celsius degree, precise up to the third decimal, for each square meter of the city of Paris). But if my initial problem is whether I should wear a coat or not, I do not need such a precise description – it would be enough to partition the states of the world with a few states such that 'below 15°C' and 'above 15°C'. 15°C seems indeed to be a reasonable threshold for many people to choose whether to wear a coat or not (but different people could consider that the relevant threshold – and then, the framing of the initial problem as a small world involving only two states – is lower or higher).

Developing a theory of choice in large worlds requires modelling how the agents frame the initial problem as a small world, with a 'reasonable' number of states of the world. We do not need to define a threshold in terms of cardinality of the set of states of the world to characterize a 'small' or 'large' world. The distinction is mostly qualitative, i.e. the differences between states of the world are meaningful for the choice under consideration.

Savage uses the two following proverbs to explain the nature of the difference between small and large worlds:

> "The point of view under discussion may be symbolized by the proverb, "Look before you leap," and the one which it is opposed by the proverb, "You can cross that bridge when you come to it." When two proverbs conflict in this way, it is proverbially true that there is some truth in both of them, but rarely, if ever, can their common truth be captured by a single pat proverb. One must indeed look before he leaps, in so far as the looking is not unreasonably time-consuming and otherwise expensive; but there are innumerable bridges one cannot afford to cross, unless he happens to come to them." (Savage, 1954, p.16)

A choice problem in which 'Look before you leap' is a reasonable principle of choice, i.e. in which the individual can anticipate all the possible consequences and plan in advance all her future moves given the possible states of the world, "in so far as the looking is not unreasonably time-consuming and otherwise expensive" (Savage, 1954, p. 16) is called a small world. Otherwise:

> "Carried to its logical extreme, the "Look before you leap" principle demands that one envisage every conceivable policy for the government of his whole life (at least from now on) in its most minute details, in the light of the vast number of unknown states of the world, and decide here and now on one policy. This is utterly ridiculous, not – as some might think – because there might be later cause for regret, if things did not turn out as had been anticipated, but because the task implied in making such a decision is not even remotely resembled by human possibility" (ibid, p. 16)

Our interpretation of Savage is that the distinction between small and large worlds is not of an ontological nature.[97] Although it might seem there exists a fundamental difference between being uncertain about 'Whether a particular egg is rotten' and 'The exact and entire past, present, and future history of the universe, understood in any sense, however wide' (Savage, 1954, p. 8), Savage sets the limit between small and large worlds with a criterion of *bounded cognitive ability*. A game of chess – though it is logically solvable by exploring all the possible sequences of moves – is for instance considered as a large world. Rather than anticipating all the future moves from a given position, a chess player will typically try to reach a position that seems good (a 'dream position'), and will start to think about her next moves only when arriving in that position (problems are thus treated sequentially). Similarly, in the case of your meeting in Paris, although the Eiffel Tower is a focal point, you will quickly realize that it is not clear where you should meet once arrived there. Probably you will then look for a meeting point close to the entrance, or try to find a spot where you can easily be seen and recognized. But when asked to meet someone in Paris a few pages ago, you did not consider all those practical issues: you reframed the initial problem as a succession of simpler problems (first, find a good spot in Paris; second, find a good spot at the Eiffel tower).

Savage's sketch of a theory of large worlds (pp. 82-91) starts from the 'grand situation', and considers that the agent will successively solve tractable problems by focusing on 'isolated decision situations'. Our framework starts from a similar principle, and we define the *grand world* $GW = \langle \mathcal{S}_{GW}, \mathcal{C}_{GW}, \mathcal{A}_{GW} \rangle$ as the 'largest' possible world. GW offers the most complete and detailed description of all the states of the worlds, acts, and of their respective consequences. Describing GW would require a perfect omniscience, so we must keep in mind that we use GW only as a theoretical tool to build a tractable model of framing in large worlds, and not as an actual world for which the agents could distinguish all the different states**.** In the case of the coordination problem in Paris, a GW-act could be something like 'go to a precise GPS coordinate, at a precise altitude'. This means that the act 'go to the Eiffel tower' corresponds to a significant subset of the GW-acts (the world in which this act is available is thus smaller than GW).

According to Savage, "to cross one's bridges when one comes to them means to attack relatively simple problems of decision by artificially confining attention to so small a world that the "Look before you leap" principle can be applied there" (Savage, 1954, p. 16). He however confesses being "unable to formulate criteria for selecting these small worlds," while believing "that their selection may be a matter of judgment and experience about which it is impossible to enunciate complete and sharply defined general principles" (ibidem). Bacharach's (1993, 2006) *variable frame theory* offers a framework explaining how players frame open problems as closed ones:

> "[I]n order to explain how someone acts, we have to take account of the representation or model of her situation that she is using as she thinks what to do. This model varies with the cognitive frame in which she does her thinking. Her frame stands to her thoughts as a set of axes does to a graph; it circumscribes the thoughts that are logically

[97] We do not mean here that there is no ontological difference at all between small and large worlds, but that even some closed systems are large worlds – and any open system may be presumably considered as a large world.

> possible for her (not ever but at the time). In a decision problem, everything is up for framing. The preferences on which she acts, her alternatives ... So far from finding herself with given preferences over outcomes, as traditional theory holds self-evident, these preferences depend upon the evaluative concepts that are uppermost in her mind." (Bacharach, in Gold and Sugden, 2006, p. 69)

Our approach will be conceptually similar to Bacharach's, though we will adopt a slightly different formalism. The main idea is that certain small worlds within the large world are more salient than others. Chess players for instance do not consider exhaustively all the possible sequence of moves, because they can identify *ex ante* only a few patterns which seem more relevant than others (and good chess players will be able to recognize intuitively only the best available candidates moves). This means that the solutions that are 'intuitive' and in the player's mind are of major importance for the individual's decisions, and constitute what we will call *individual* focal points.

2.3. The role of mindshaping and focal points for cognitive homogenization and coordination

Schelling (1980[1960]) defines focal points as everything that is salient: a pattern of behavior, a strategy, an outcome, etc. Anything that leads the players to perceive an outcome as the solution of the game is a focal point. A focal point therefore serves as a guide to coordination: it first *allows* coordination by making individual beliefs converging to a specific solution (Schelling 1980[1960] p. 56), and second *maintains* coordination by reinforcing the coherence of individual beliefs regarding everybody else's behavior (Hédoin, 2014, p. 366). Focal points can be used both as an input and as an output of successful coordination. Seen as a regulatory social device,[98] a focal point generates symmetric reasoning, and may induce stable patterns of behavior.

The existence of focal points is part of the processes of mindshaping to the extent that it eases the cognitive homogenization of the environment, both by aligning players' perceptions of the environment, and by reinforcing those perceptions thanks to the success of past coordination. If Châtelet les Halles is focal among the community of Parisians, it aligns the beliefs of Parisians about what is a good meeting point, and the fact that Parisians successfully coordinated by going to Châtelet les Halles justifies that it constitutes a good meeting point for Parisians. Even though my individual focal point can be to go to the restaurant, my subjectively based social focal point (what I believe I share with the other individual) which is to go to Châtelet les Halles is reinforced if it turns out to be an objective social focal point (i.e. if all the Parisians consider that it as a subjective social focal point).[99]

---

[98] For other aspects of Schelling's account of focal points, see the chapter 2 of the thesis or Larrouy (2018).

[99] This distinction between individual and social focal points means that an individual can differentiate her own frames, intentions or beliefs, from those she believes are common among the group (Orléan, 2004). Framing

The cognitive processes allowing a successful coordination are a combination of *simulation thinking* and mindshaping. Recall that simulation thinking, as conceptualized by Goldman (2006), consists in using one's own mind to simulate the reasoning of others. People anticipate the beliefs and behaviors of others by projecting their own perceptions, beliefs, and mode of reasoning in the mind of others – first by attributing their own beliefs in presence of uncertainty (e.g. I believe that the Eiffel Tower is focal for a foreign tourist, although I have no specific information about her), and second by using their own reasoning as a simulator of the reasoning of others (e.g. if I think it is 'rational' to go to the focal point, I will believe that you think it is 'rational' to go to the focal point too). This means that, in case of a full uncertainty about the other, the individual tends to attribute to the other her own beliefs about what she considers as focal.[100] If I have no reason to believe that you could be unaware that the restaurant is a good meeting point, then my own individual focal point becomes a subjectively based social focal point.

It now clearly appears that simulation thinking is accurate and brings correct anticipation only if our minds have been shaped by the same models – and that it is indeed realistic to identify the Eiffel tower as focal, and that my mode of reasoning (e.g. expected payoff maximisation) is similar to yours. The role of mindshaping becomes decisive: it is only if we share a common and homogeneous perception of the environment that simulation may be effective – and that the points we identify as subjectively based social focal points are *objective* social focal points.

An important point here is that my prior belief about how the other will behave depends on what I think we share in common: apart from the case of full uncertainty (where I tend to attribute my own individual focal point to the other), any information about the other – i.e. you are a Parisian, a colleague, or a tourist – will determine my belief about what we have in common, and therefore what we are likely to both identify as a social focal point.

Given the initial (and possibly infinitely complex) problem of meeting in Paris, mindshaping helps to "drastically reduce, and hence make manageable, the space of interactions in which we engage: we restrict the games we can play with each other … and coordination is dramatically facilitated." (Zawidski, 2013, p. 58)[101] In other words, mindshaping reduces the set of properties that the players can have in common. It first determines our own small world representation (through the different properties that spontaneously come to our minds – e.g we do not think in terms of GPS coordinates, but of larger locations), and then our beliefs about the *shared* small world representation, depending on whom we are interacting with.

Even if we do not share the exact same cultural models and cannot reach a perfect homogeneity, a partial overlapping of those models is in many cases sufficient to bring enough homogenization. And even if the small world representation that the players have built turns out

---

becomes strategic, since the player can distance herself from her own initial framing, and strategically reframe the game.

[100] This is known as the 'egocentric bias' (see Goldman, 2006, pp. 177-179), and the observation of such bias can be used as evidence of simulation thinking.

[101] Similarly, Ross (2007) suggests that it is only because of some social devices shape people's mindso that they become more alike that coordination can occur.

to be erroneous, they will build another small world until congruence is reached. An objective social focal point – by signaling the convergence of small-world representations – reinforces the beliefs and representations of the players, while it is initially the outcome of an evolutionary process of trials and errors.

3. A model of strategic reasoning in small worlds

3.1. Simulation and the formation of players' beliefs

Predicting someone else's behavior requires forming expectations about the mental states leading this other to adopt a specific behavior. In a game situation, this typically means forming expectations about her preferences, beliefs and objectives. Once Player 1 (P1) has formed her beliefs about the choice of Player 2 (P2), (based on what P1 thinks P2's preferences, beliefs and objectives are), P1 can determine her best reply. The simulation theory suggests that P1 uses her own mind to predict the behavior of P2: "our own mental processes are treated as a manipulable model of other minds" (Cruz and Gordon, 2005, p. 10). Simulation is therefore an efficient heuristic for predicting someone else's decision (Shanton and Goldman, 2010) – a recent upsurge of empirical data provided by neuro-imaging indeed contributes to suggest that simulation is a very effective process of mindreading (Goldman, 2006; see also Kirman and Teschl, 2010; Singer and Fehr, 2005). Consider the Stag Hunt discussed above (with material pay-offs rather than vNM utils):

SH	A_2	B_2
A_1	(\$3;\$3)	(\$0;\$2)
B_1	(\$2;\$0)	(\$1;\$1)

Suppose that P1 intends to play a best reply to P2's strategy: P1 must therefore anticipate the choice of P2 (strategy A2 or strategy B2). To do so, she imagines herself in P2's shoes. The simulation process is then a three steps process (P1 is called the attributor or simulator, and P2 the target). First, P1 brings forth 'pretend' or 'imaginary' mental states – like the intention for P2 to maximize her material pay-off, the preference for the profile A1A2, the belief that P1 is likely to choose A1 – in her own mind.

Those mental states are supposed to 'mimic' those of her target: as a simulator, P1 pretends that those imaginary intentions, preferences and beliefs, are those of her target P2.[102] Second, she feeds these pretend states in her own decision-making system, which runs 'off-line'. As a best-reply reasoner, P1 imagines what she would play if she were choosing instead of P2, with the intention of reaching a pay-off maximizing profile for P2, given P2's belief that P1 is likely to choose (say) A1.[103] The output of the simulation process (the best reply to what P1 thinks P2 believes about P1) is therefore A2. Lastly, this output is attributed to her target (Goldman, 2006, p. 20). P1 believes that P2 will choose A2, her best reply to P2's choice is thus to play A1. Simulation thus explains how P1 forms her beliefs about P2's mental states, reasoning and choice. Assuming this alignment of rationality is justified in first instance by the mindshaping hypothesis explained in the previous section. It is also explained according to the simulation process by a pragmatic heuristic which, in the absence of any information regarding the other, allows the player to form beliefs about the other's eventual choice. This however does not mean that the player is not aware that such hypothesis can be wrong. Besides assuming this alignment of individual rationality does not correspond to the standard imposition of priors (correlated or commonly aligned) which again as explained in the chapter 1 are merely mathematical artefacts or notational artefacts that merely means that the players' posteriors are at the equilibrium and thus describe the players' choices at the equilibrium.

If the attributor's decision-making process and pretend initial mental states are similar to that of the target, then the output of the simulation process is a reliable prediction of the target's choice. Simulation may however not lead to accurate predictions, for instance if the mental states attributed to the other are incorrect, and "chosen badly out of ignorance" (Goldman, 2006, p. 48), or if the other's decision process is different from the one of the simulator. Numerous experimental findings indeed report egocentric biases in predicting other's choices, which preclude from an accurate mindreading (Goldman, 2006, pp. 177–179). In the previous illustration, although P1 predicts that P2 will choose A2, it is not certain that P2's actual beliefs about the choice of P1 correspond to the 'pretended' beliefs simulated by P1. Furthermore, unlike P1 who is a best reply reasoner, P2 could be a maximin player, i.e. a player who always chooses her maximin strategy (B2 in the Stag Hunt). Since P1 and P2 reasoning processes are different, it is likely that the outcome of their reasoning will be different. In cases of erroneous predictions, the simulators revise their beliefs about their targets' mental states, and then run other simulations with different inputs.

In any case, even with very little information about the target, such as the information given in a pay-off matrix, an individual can form a prediction, based on her own perception of the game, and her own cognitive scheme and reasoning process. Hence, the ST provides a theory to derive endogenously the beliefs of the players from the structure of the game.

---

[102] Goldman (2006) considers that the role of pretence, which is the core of the simulation routine, is not restricted to mental states. It intervenes either for processes (i.e. a decision-making mechanism) or for its inputs (i.e. beliefs, desires, etc.) and outputs (i.e. the decision). In this paper, we will only consider that players simulate the process of decision-making and its inputs (when players have neither information nor specific prior belief about the beliefs of others).

[103] Note that the belief about P1's action that P1 attributes to P2 is not justified for now: we will determine in Section 3 what beliefs P1 could attribute to P2 if she also believes that P2 is a Bayes rational player.

3.2. The formalization of Simulation Theory in games

We introduce the following belief operators:

- 1st order belief: $B_i(E)$ means '*i* believes E', with $E \in \mathcal{E}$ a proposition, and $\bar{E}$ its negation
- 2nd order belief: $B_{ij}(E) = B_i\left(B_j(E)\right)$, means '*i* believes that *j* believes E'
- Mutual belief: $MB(E) = B_i(E) \cap B_j(E) = (B_i \cap B_j)(E)$ means that 'i and j believes E'
- Common belief: $CB(E) = (\cap_{k=1}^{\infty} MB^k)(E)$ means that 'E is mutual belief', that 'the proposition "E is mutual belief" is mutual belief', and so on *ad infinitum*
- Uncertainty: $U_{ij}(E) = \overline{B_i\left(B_j(E)\right)} \cap \overline{B_i\left(\overline{B_j(E)}\right)}$ means that *i* neither believes that '*j* believes E', nor that '*j* does not believe E'.

We saw that the simulation routine (i) tends to imply an egocentric bias, i.e. that the individual tends to attribute her *own* beliefs and perceptions to the other, and (ii) that the simulator uses her own reasoning process to simulate the reasoning of her target. An egocentric bias means that, when *i* is in a situation of uncertainty regarding *j*'s knowledge of a proposition E (or at least, *i* does not believe that *j* does not believe E), then *i* tends to assume that *j* believes E. Formally, we can translate this property as follows:

$$\text{SIMB}_\text{i}: \quad B_i(E) \cap U_{ij}(E) \Longrightarrow B_{ij}(E). (1)$$

SIMB_i means that, in presence of uncertainty, *i* attributes her own beliefs to player *j*. Think for instance of a coordination game with a focal point: if I believe that an option is more salient than the others, and that I don't have a specific reason to believe that you don't perceive this option as more salient, then I will assume you also perceive this outcome as the most salient (which could then rationalize our coordination on the focal point). This corresponds to the subjectively based social focal point described in the previous section. This supposition is besides largely confirmed by mindshaping. If *i* thinks that *j* has been shaped by the same social model as her, which can be true in many circumstances, then such presumption, in first instance makes sense (unless some additional information reveals that they do not belong to any common socio-cultural category).

When the uncertainty extends to the beliefs of another player, i.e. $U_{ij}(E'), \forall E' \in \mathcal{B}_j$ with $\mathcal{B}_j$ the set of *j*'s beliefs about the beliefs of other players (i.e. *j*'s first-order beliefs, second-order beliefs, etc.), we have the following result – similar to Friedell's (1969, p. 31) intuition of the emergence of *common opinion* – (all the proofs are given in appendix):

Proposition 1. *Let E be a proposition, and suppose that $U_{ij}(E'), \forall E' \in \mathcal{B}_j$:*

$$B_i(E) \cap U_{ij}(E) \cap SIMB_i \Longrightarrow B_i\big(CB(E)\big). (2)$$

Proposition 1 means that, if *i* believes E and is uncertain about *j*'s beliefs (i.e. about *j*'s belief about E, but also about *j*'s belief about *i*'s belief about E, etc.), then *i* believes that E is common belief among them. Unless *i* has a good reason to believe that *j* does not believe E, *i* simply assumes that "*j* is in the same cognitive position as *i* himself" (Friedell, 1969, p. 31): *i* will therefore believe that they both believe E, but also that they both believe that they believe E, etc. Again, such statement rests on mindshaping. Since *i* is uncertain about *j*'s beliefs, and in particular if *j* believes that they both believe E, *i* indeed also attributes her second-order belief to *j*, etc. Friedell (1969) then shows that this 'symmetry' of cognitive positions between the players generates a structure of common belief.[104]

A corollary of proposition 1 is that, if all the players attribute their own beliefs to the others in presence of uncertainty, and if a proposition E is mutual belief among the players, then it is also common belief among them (by proposition 1, they indeed all believe that E is common belief, i.e. that the common belief of E is mutual belief, which is the same thing than the common belief of E). An interesting implication of this corollary for game theory is that mutual knowledge of the structure of the game is sufficient to ensure its common belief. This echoes the work of Bacharach (presented in chapter 3) who claims that mutual knowledge is in many cases largely sufficient. Although the assumption of common knowledge of the structure of the game seems too cognitively demanding, simulation and mutual knowledge of the structure of the game are sufficient to generate a formally equivalent epistemic structure. The insight of simulation is indeed that we do not require the *actual* belief of the players, but simply that they have the cognitive capacity to *generate* it.

Regarding the simulation of the reasoning process of the other individual, we must firstly introduce player *i*'s *choice function*. Consider a choice problem $\mathrm{P} = (A_i, C_i, O_i)$, in which *i* must choose an action $a_i \in A_i$, with a consequence $C_i(a_i)$, to satisfy her objective O_i (e.g. maximize her monetary gain). We define a choice function $\mathcal{C}_i : \mathrm{P} \mapsto A_i$ as a function that associates to a choice problem $\mathrm{P} = (A_i, C_i, O_i)$ the action to be chosen in A_i. Simulation means that, for a given choice problem P_j for player *j*:

$$\mathrm{SIMR_i}: \quad \mathcal{C}_i(\mathrm{P}_j) = f(A_j, C_j, O_j) \implies B_i\left(\mathcal{C}_j(\mathrm{P}_j) = f(A_j, C_j, O_j)\right). \quad (3)$$

$\mathrm{SIMR_i}$ means that, if the function $f : A_j \times C_j \times O_j \mapsto A_j$ corresponds to the function player *i* would apply to select an action in A_j (if she were in the position of player *j*, i.e. if she had to choose instead of *j*, given *j*'s objectives), then *i* believes that *j* would apply the same function to select an action (e.g. if my objective is to maximise my monetary gain, then f would be the argmax function applied to *j*'s monetary gain).

3.3. Reaching consistent beliefs: the massaging process

[104] Friedell does not use the expression 'common belief' but 'common opinion'. The two concepts are however mathematically identical (see Péréa, 2014, p. 11).

Note that the specificity of game theory – which was put forward by Harsanyi and then Aumann against Kadane and Larkey – is that the players expect each other to act rationally. This means that the beliefs of P1 are already the outcome of a reasoning process, because they must be revised to be consistent with the rationality of P2. To explain this process we will use Binmore's (2009, pp. 130–132) description of the 'massaging process'; thus allowing to model how P1 could reach consistent priors.[105] Binmore relies on Savage's distinction between small and large worlds to question Bayesianism – which he defines as "the philosophical position that Bayesian methods always applies to all decision problems" (Binmore, 2009, p. 96) – and in particular the claim that "rationality endows agents with prior probabilities" (Binmore, 2006, p. 3). Binmore on the contrary intends to explain the formation of prior beliefs, while remaining faithful to Savage's theory. Rather than directly using my "gut feelings" (Binmore, 2009, p. 130) at the prior stage to form my beliefs, I should imagine what would be my gut feelings after the realization of the state of nature. For each realization, I deduce a posterior belief from my hypothetical gut feelings: it is however unlikely that those posterior beliefs are consistent. I should then 'massage' my posteriors until I reach consistent posteriors (i.e. revise my posteriors, knowing that my hypothetical gut feelings were inconsistent).

Once I reached consistent massaged posteriors, Bayes' rule guarantees that they can be deduced from a prior. We endorse Bayesian rationality as underlined in the introduction of this chapter to the extent that it means that the players adopt their best reply reasoning according to their beliefs regarding the others' choice for instance. By no means we ascribe to what Binmore calls the "bayesianist" view of epistemic game theory. The formation of my prior belief thus requires a stage of introspection and self-reflection, and this is precisely what we will do with ST, by explicitly modelling the introduction of the common belief in rationality in the players' beliefs via a similar 'massaging process'.

Binmore quickly mentions how the massaging process could work in game theory. Rather interestingly, he suggests that "Alice will then not only have to massage her own probabilities until consistency is achieved, she will also have to simulate Bob's similar massaging efforts" (Binmore, 2009, p. 135), and suggests that the resulting beliefs form a subjective equilibrium, which should necessarily be a Nash equilibrium. Although our proposition 4 will confirm Binmore's intuition, we will also show that the introduction of simulation (which is simply a passing remark in Binmore's argument, and not a well-developed game theoretical analysis) could justify a much larger set of equilibria (that we will call 'subjective belief equilibrium' instead of 'subjective equilibrium'). The reason is that, even when massaging our priors until they become consistent, it is possible to form action-dependent beliefs (that Binmore explicitly rejects for reasons we will discuss in Section 4).

---

[105] By 'consistency', we mean the internal consistency between the player's beliefs about the rationality of the others and her belief about their actions. The beliefs that you will play a strictly dominated strategy and that you are a best reply reasoner are for instance inconsistent, and requires me revising my beliefs (the 'massaging process' corresponds to this phase of self-introspection during which I revise my beliefs until they become consistent).

As an illustration of the formation of action-dependent beliefs through the massaging process, consider the previous Stag Hunt:

SH	A_2	B_2
A_1	(\$3;\$3)	(\$0;\$2)
B_1	(\$2;\$0)	(\$1;\$1)

Unlike EGT – according to which P1 is endowed with prior probabilities, which should necessarily be consistent with the mutual rationality of the players – we suppose that P1 imagines what would be her gut feelings about the action of P2 after she chose her own action. In the spirit of Haruvy, Stahl and Wilson (1999), suppose that each player can either be 'optimistic' or 'pessimistic'. An optimistic type "tends to choose the strategy which can potentially give him the highest payoff for a given game" (Haruvy, Stahl and Wilson, 1999, p. 256) and a pessimistic type "is motivated by worst case scenarios and hence tends to choose a secure action" (ibid, p. 257). An optimist would therefore select the pay-off dominant equilibrium, while a pessimist would opt for a maximin strategy, and therefore the risk dominant equilibrium. P1 can thus be motivated by two conflicting reasons: either (R1) try to reach the highest pay-off (maximax) or (R2) secure the highest minimum pay-off (maximin). P1 however does not necessarily know what psychological factors make her privilege a reason over another.[106] Therefore, if P1 eventually chooses A1, she may deduce that the reason (R1) appeared as relatively more important than (R2) for some unknown psychological reason. When simulating the reasoning of P2, she will assume that the same psychological forces are driving P2's choice (since she attributes her own mental states to P2 when simulating her reasoning). From that prospect, if P1 chooses A1, she may attribute a higher probability for P2 choosing A2, because she will assume that the psychological forces that pushed her to privilege (R1) and then to choose A1 are also probably operating in P2's mind. P1 can then form an action-dependent belief, not because she thinks her choice directly influences the choice of the other, but because she is aware that her choice is driven by psychological factors that could also influence P2. Her gut feelings about the choice of the other could indeed depend on her own actions: if P1 feels optimistic, and thus chooses A1, she will more likely believe that P2 chooses A2, because she tends to attribute her own optimism to P2. On the contrary, when considering what would be her gut feelings about P2's choice if she chose B1, P1 will attribute her own pessimism to P2, and is thus more likely to believe that P2 will choose B2. P1's prior belief could therefore be of the following form:

[106] Battalio, Samuelson, and Van Huyck (2001) suggests for instance that the probability of choosing the risk-dominant equilibrium tends to increase with the optimisation premium (the difference between the pay-off of the best response and the inferior response – which is for instance larger when P_1.A; A., P_1.A; B. and P_1.B; B. are of a similar magnitude, while being significantly higher than P_1.B; A.). However, the subjects in their experiment were probably not aware that their choice was actually influenced by this premium (for the simple reason that they probably did not explicitly calculate the premium before choosing).

SH	C_2	D_2
C_1	α	0
D_1	0	$1-\alpha$

In this case, P1 believes that the outcome of the game is A1A2 or B1B2, with probability α and $(1-\alpha)$, respectively. Those priors simply state that both players are best reply reasoners (i.e. P1 plays A1 if and only if P2 plays A2). We thus obtain correlated priors, although they are consistent with the mutual rationality of the players – we even have in this example a correlated equilibrium distribution.

The massaging process can therefore offer a psychological explanation of the formation of consistent (and possibly correlated) prior beliefs. The player indeed refers to her own psychological make-up (her gut feelings) to form her prior beliefs, which she revises until they are consistent with the rationality of the other players. The formation of these prior beliefs understood as gut feelings primarily refers to the mindshaping perspective of our analysis while the belief revision until consistency is mainly explained by the simulation process.

Our account of games and of strategic reasoning for explaining how players' converge toward an equilibrium and thus are able to coordinate, as is the main point of the thesis, requires to bypass the mere mathematical and closed system account of game theory to include determinants that refer to cognitive psychology or social psychology (with respectively the simulation theory and mindshaping).

4. Subjective belief equilibrium

We now present the concept of subjective belief equilibrium, as a strategy profile that results from the joint maximization of individual expected pay-offs, when the beliefs of the players have been 'massaged' by the players. We start by characterizing a massaged belief hierarchy as the beliefs formed by Bayes rational players who simulate the reasoning of the other players and define a subjective belief equilibrium as the pay-off maximizing profile derived from massaged beliefs (Section 4.3). We then illustrate the possibility of forming consistent action-dependent beliefs (Section 4.2). We finally show that simulation by allowing action-dependent beliefs challenges the controversial assumption of ratifiability (section 4.3).

4.1. The Massaged belief hierarchy and the subjective belief equilibrium

We suggested in the previous section that simulation could lead players to form action-dependent beliefs (ADB). We now show that it is also possible for the players to rationalize those beliefs, i.e. to hold consistent ADB. We introduce the notion of massaged belief hierarchy as the beliefs of the players resulting from the simulation of each other's reasoning.

We consider finite games in normal forms, i.e. games $G = \langle N, X, \Pi \rangle$, with $N = \{1; \dots; n\}$ the set of players, $X = \prod_{i \in N} X_i$ with X_i the finite set of player *i*'s strategies, and $\Pi_i : X \longmapsto \mathbb{R}$ player *i*'s material payoff. We will see in section 5 how to extend the analysis to games in which such conditions are not secured (i.e. the set of strategies of the other players is not *a priori* finite, and their payoff unknown). As discussed above, this material payoff function is the primitive of the game, and we assume that maximizing one's material payoff is the objective the individual intends to achieve (in a non-tautological way). $\Delta(X) = \left\{ \{P(x)\}_{x \in X} \in [0;1]^{|X|} \mid \sum_{x \in X} P(x) = 1 \right\}$ denotes the set of discrete probability distributions over *X*. While *x* denotes players' actions, we use the letter *s* to denote players' beliefs about actions. $s_{i,j,k} \in X_k$ denotes for instance *i*'s belief about *j*'s belief about *k*'s strategy:

$$s_{i,j,k} = \bar{x}_k \iff B_{ij}(x_k = \bar{x}_k). \quad (4)$$

Consider a game in normal form $G = \langle N, X, \Pi \rangle$. For a given profile of conditional probability distributions $P(X) = \{P(X_i|X_{-i})\}_{i \in N}$, we denote by $\Omega(\{P(X_i|X_{-i})\}_{i \in N})$ the set of probability distributions that can represent $\{P(X_i|X_{-i})\}_{i \in N}$. $\Phi_i(X)$ denotes the set of conditional probability distribution $P(X_i|X_{-i})$ for player *i*. As illustrated below, the function $\Omega : \Phi \longmapsto \Delta(X)$ is neither injective nor surjective: a single set of conditional probability distributions can indeed be represented by several probability distributions over outcomes, and several sets of conditional probability distributions can be represented by the same probability distribution. Lastly, two conditional probabilities $P(X_i|X_{-i})$ and $P(X_{-i}|X_i)$ are not necessarily compatible, i.e. there may not exist a distribution $p \in \Delta(X)$ that can simultaneously represent those two conditional probabilities (see e.g. Arnold and Press, 1989) – in which case $\Omega(\{P(X_i|X_{-i})\}_{i \in N}) = \emptyset$.

We denote by $S_i = \left\{ \{s_{i,j}\}_{j \neq i}; \{s_{i,j,k}\}_{k \neq j;\, j \neq i}; \{s_{i,j,k,l}\}_{l \neq k;\, k \neq j;\, j \neq i}; \dots \right\}$ player *i*'s belief hierarchy, i.e. the infinite set including player *i*'s 1st-order beliefs $s_{i,j}$ (her belief about *j*'s strategy), her 2nd order beliefs $s_{i,j,k}$ (her belief about *j*'s belief about *k*'s strategy), etc. Our objective is to identify the belief hierarchies that a Bayes rational player could hold if she simulates the reasoning of the other players (we suppose throughout the rest of the section that players are initially uncertain about the beliefs of the others, and must in consequence form those beliefs by simulating their reasoning). Suppose that players $i \in N$ choose the strategy that maximizes their expected material payoff:

$$\text{PM: } \max_{x_i \in X_i} \left[\sum_{x_{-i} \in X_{-i}} s_{i,-i}(x_{-i}|x_i)\, \Pi_i(x_i, x_{-i}) \right] \quad (5)$$

with $s_{i,-i}: X_i \longmapsto \Delta(X_{-i})$ *i*'s belief about the strategy of the players in *-i*.

Definition. S_i is a massaged belief hierarchy if and only if PM, $SIMB_i$and $SIMR_i$ are true for player i.

A massaged belief hierarchy (MBH) is the belief hierarchy of player *i* when she simulates the reasoning of other players. Those beliefs must be compatible with the rationality of the other players, since *i* will assume that the others are expected payoff maximizers if she is herself an expected payoff maximizer (which is true by PM). We can now characterise a MBH:

Proposition 2. S_i *is a massaged belief hierarchy if and only if:*

i) $s^*_{i,[k],j} = s^*_{i,j}$ with $j \neq i$, for all sequences of players $[k] = k_1, k_2, \dots, k_m$.

ii) $s^*_{i,[k],i} = s^*_{i,j,i}$, $\forall j \neq i$, for all sequences $[k] = k_1, k_2, \dots, k_m$.

iii) and there exists $s^* \in \Omega\left(\left\{s^*_{i,j,k}\right\}_{i\neq j; j\neq k}\right)$ such that:

$$\sum_{x\in X} s^*(x)\Pi_k(x) \geq \sum_{x\in X} s'(x)\Pi_k(x), \qquad \forall s' \in \Omega\left(\left\{s^*_{i,j,k}\right\}_{i\neq j; j\neq k}; s'_{i,j,k}\right), \forall s'_{i,j,k} \in \Phi_k, \forall k \in N. (6)$$

Condition (i) means that *i* believes that her belief about *j*'s strategy is shared by all the other players, and that it is common belief (the *m*th order belief of *i* about *j*'s strategy is indeed always equal to her first order belief about *j*'s strategy). Condition (ii) is similar to (i), since it means that *i* believes that all the players have the same belief about her own strategy, and that it is common belief. Those two conditions ensure that the beliefs of the players converge, in the sense that *i* believes that the first order beliefs of all the players are common belief and identical. It also implies that it is sufficient to work with *i*'s 1st and 2nd order beliefs, rather than with the whole belief hierarchy S_i.

Condition (iii) means that the expected payoff of player *k* at the MBH (i.e. if the choice of the players is accurately described by the distribution s^*) should be at least equal to her expected payoff if she 'deviates' to another conditional distribution $P'(X_k|X_{-k})$. In other words, a MBH is the probability distribution $P^*(X)$ induced by a set of conditional probability distributions $\{P^*(X_i|X_{-i})\}_{i\in N}$, such that $P^*(X_i|X_{-i})$ is maximizing the expected utility of player *i*, $\forall i \in N$.

The intuition supporting condition (iii) is the following. Consider *i*'s belief about the strategy of player *j*:

$$s_{i,j}(x_{-j}) = P(X_j | X_{-j} = x_{-j}) (7)$$

This belief is rationalizable if and only if *i* can justify why *j*'s choice is accurately described by $P(X_j | X_{-j} = x_{-j})$. This means that $s_{i,j}$ must be a 'rational' choice for player *j*, given *j*'s belief about the choice of players in *–j*. Given S_i, *i*'s belief about *j*'s belief about the strategy of players $k \in -j$ is:

$$s_{i,j,-j}(x_j) = P(X_{-j}|X_j = x_j) (8)$$

If there exists a distribution $P'(X_j|X_{-j}) \neq P(X_j|X_{-j})$ such that:

$$\sum_{x \in X} s(x)\Pi_j(x) > \sum_{x \in X} s'(x)\Pi_j(x), (9)$$

with $s' \in \Omega\left(P'(X_j|X_{-j}); P(X_{-j}|X_j)\right)$, then it means that $s_{i,j}$ is not a rationalizable belief from *i*'s perspective, because *j* would be better off if her choice was described by $P'(X_j|X_{-j})$. The belief hierarchy of *i* should therefore be such that she can rationalize the beliefs she attributes to the other players (this is why the player 'massages' her prior so as to reach consistent posteriors): each conditional distribution $P(X_j|X_{-j})$ should therefore maximize the expected utility of player *j*, given *i*'s belief about *j*'s belief about the strategy of the players $k \in -j$ (and these higher order beliefs must also be rationalizable – this is ensured by parts (i) and (ii) of the proposition, because they are the same than $s_{i,j}$ and $s_{i,j,i}$).

We can show the following existence result for a MBH:

Proposition 3*. Let $G = \langle N, X, \Pi \rangle$ be a game in normal form. If G has a Nash equilibrium $p^* \in \Delta(X)$, then there exists a massaged belief hierarchy S_i^* for each player i.*

Proposition 3 ensures that the existence of a Nash equilibrium is a sufficient condition for the existence of a massaged belief hierarchy for each player. In other words, if a Nash equilibrium exists, then we know that the players can always manage to rationalize a belief hierarchy S_i^*.

We want to emphasize that the optimization condition defining a MBH is about players' *beliefs*, and not their actual choices. Condition (6) indeed suggests that player *i*'s strategy can be conditional on player *j*'s strategy (since we look for a *conditional* probability distribution that maximizes *i*'s expected material payoff): this is only because player *j* believes that her choices and those of player *i* could be correlated (and *j* can believe this if she believes that *i* believes it). *When forming their beliefs* (and not when choosing their actual strategies), players therefore test the stability of their beliefs not with regard to their own actions (which are independent), but with regard to their beliefs about the others (which may be dependent). Although all the players know that they cannot choose conditional distributions, they can use the fact that all the players can believe that beliefs are action dependent, and then form a belief hierarchy in which beliefs are correlated. Furthermore, since this massaging process only happens in *i*'s mind, nothing guarantees that all the players will reach the *same* MBH (in particular when several distributions satisfy condition iii). The mindshaping hypothesis however ensures that a certain degree of homogeneity among players' mind prevails so that players ultimately converge on a common MBH. A MBH is therefore an individual concept, and does not guarantee that the players' actual choice will be accurately described by the MBH (the methodological implications of the multiplicity of MBH will be discussed in conclusion).

We now define a solution concept representing the choice of Bayes rational players who massaged their beliefs. We define a *subjective belief equilibrium* as the strategy profile resulting from the maximization of the players' subjective expected payoff. The subjective dimension here translates the idea that it is from their own evaluation, their own perception. This equilibrium is

restricted to *pure* strategies, because – in line with Aumann (1987, p. 15) – we interpret mixed strategy equilibria in terms of beliefs. The players cannot randomize when choosing their strategy, but their belief hierarchy can include mixed strategies if they are not certain of the actions of the other players. In our model, a mixed strategy Nash equilibrium would be the (common) massaged belief hierarchy of the players, while the subjective belief equilibrium would correspond to the strategy profile resulting from the maximization of their expected payoff (if they accurately anticipate it).

Definition. *A strategy profile* $x^* \in X$ *is a subjective belief equilibrium if and only if,* $\forall i \in N$*, there exists a massaged belief hierarchy* S_i^* *such that:*

i) $\sum_{x_{-i} \in X_{-i}} s^*(x_{-i}|x_i^*)\ \Pi_i(x_i^*; x_{-i}) \geq \sum_{x_{-i} \in X_{-i}} s^*(x_{-i}|x_i')\ \Pi_i(x_i'; x_{-i}),\ \forall x_i' \in X_i$

ii) $s^*(x_{-i}|x_i^*) = x_{-i}^*,\ \ \forall i \in N, i \neq j.$

With $s^* \in \Omega\left(\left\{s_{i,j,k}^*\right\}_{i \neq j; j \neq k}\right)$ the representation of S_i^* given by proposition 2.

The first condition means that each player chooses the strategy that maximizes her expected payoff, given her beliefs about the action of the others. Those beliefs should however be rationalizable, since they are derived from a massaged belief hierarchy. The second condition means that the players accurately predict the choice of the other players. This means that the existence of a MBH is a necessary but not sufficient condition for the existence of a SBE. For instance, in the 'matching pennies' game, the two possible strategies have the same expected payoff (if players' beliefs are such that they believe it is common belief that they play each strategy with probability ½, which is a MBH), but the players cannot predict the actual choice of the other player.

Similarly to Binmore's (2009, p. 135) suggestion, we can establish a direct connection between Nash and a subjective belief equilibrium:

Proposition 4. *A Nash equilibrium in pure strategies is a subjective belief equilibrium*

The intuition of the proof is quite straightforward and follows from proposition 3 if there exists a Nash equilibrium in pure strategies, then the players can form the consistent priors that they all play this equilibrium. Given this prior belief, they maximize their expected pay-off by playing their part of the Nash equilibrium. Unlike Binmore we will however show that the set of subjective belief equilibria is much larger, since it is possible to form consistent ADB.

4.2. Illustration: Prisoner's dilemma

As an illustration, consider the following Prisoner's dilemma (the pay-offs in the matrix should be interpreted as material pay-offs and not as vNM utilities):

G	C_2	D_2
C_1	(2;2)	(0;3)
D_1	(3;0)	(1;1)

Consider the following belief hierarchies for player 1 (we only focus on 1st and 2nd order beliefs, and assume that higher order beliefs will be identical):

$\boldsymbol{S_1^1}$: $s_{1,2}(C_2|X_1) = 0,\ s_{1,2,1}(C_1|X_2) = 0,\ \forall X_i \in \{C_i, D_i\}$. P1 believes that P2 never cooperates, and believes that P2 believes that P1 never cooperates,

$\boldsymbol{S_1^2}$: $s_{1,2}(C_2|X_1) = 1,\ s_{1,2,1}(C_1|X_2) = 1,\ \forall X_i \in \{C_i, D_i\}$. P1 believes that P2 always cooperates, and believes that P2 believes that P1 always cooperates,

$\boldsymbol{S_1^3}$: $s_{1,2}(C_2|C_1) = 1,\ s_{1,2}(C_2|D_1) = 0,\ s_{1,2,1}(C_1|C_2) = 1,\ s_{1,2,1}(C_1|D_2) = 0$. P1 believes that P2 is ready to cooperate if and only if P1 cooperates, and believes that P2 believes that P1 is ready to cooperate if and only if P2 cooperates (P1 believes that P2 is a conditional cooperator, and believes that P2 believes that she is also a conditional cooperator),

$\boldsymbol{S_1^4}$: $s_{1,2}(C_2|C_1) = 0,\ s_{1,2}(C_2|D_1) = 0,\ s_{1,2,1}(C_1|C_2) = 1,\ s_{1,2,1}(C_1|D_2) = 0$. P1 believes that P2 always defect, and believes that P2 believes that P1 is a conditional cooperator,

$\boldsymbol{S_1^5}$: $s_{1,2}(C_2|C_1) = 0,\ s_{1,2}(C_2|D_1) = 1,\ s_{1,2,1}(C_1|C_2) = 1,\ s_{1,2,1}(C_1|D_2) = 0$. P1 believes that P2 cooperates if and only if P1 defects, and believes that P2 believes that P1 is a conditional cooperator.

We now represent each belief hierarchy S_1 by a probability distribution $s \in \Omega(s_{1,2}; s_{1,2,1})$ and check whether they form a MBH.

$\boldsymbol{S_1^1}$	C_2	D_2
C_1	0	0
D_1	0	1

Given that P1 believes that P2 always defect and believes that P2 believes that P1 always defect, the only possible representation s of those conditional distributions is that the probability of reaching D_1D_2 is equal to 1. Given those beliefs, the expected payoff for both players (from P1's perspective, since we only consider P1's belief hierarchy) is 1. P1 knows that, if P2 believes that P1 always defects, then P2 can only get a payoff of 0 (if she cooperates) or 1 (if she defects). This means that any other conditional probability distribution than $s_{1,2}(C_2|X_1) = 0$ cannot give a

strictly higher payoff to P2. $s_{1,2}$ is therefore a rationalizable belief, since believing $s_{1,2,1}(C_1|X_2) = 0$ implies that $s_{1,2}(C_2|X_1) = 0$ is payoff maximizing (if P1 believes $s_{1,2,1}(C_1|X_2) = 0$, then she can explain why P2 would unconditionally defects). By a similar argument (since we assumed that $s_{1,2,1,2} = s_{1,2}$) we could rationalize P1's 2nd-order belief, i.e. that P2 believes that she always defects. This is indeed payoff maximizing, given P1's belief about P2's beliefs.

Since P1 can rationalize the beliefs about P2's strategy and the beliefs she attributes to P2, S_1^1 is a massaged belief hierarchy (if players could choose conditional strategies instead of pure strategies, then no player could strictly increase her material payoff by switching to another conditional strategy).

$\boldsymbol{S_1^2}$	C_2	D_2
C_1	1	0
D_1	0	0

We now consider the second belief hierarchy, according to which both players always cooperate. If P1 believes that P2 believes that P1 always cooperates, then P1 cannot also believes that P2 always cooperates: if P2 believes that P1 always cooperates, then it would be in the interest of P2 to always defect. P1 must therefore revise her inconsistent beliefs: $\boldsymbol{S_1^2}$ is not a massaged belief hierarchy.

$\boldsymbol{S_1^3}$	C_2	D_2
C_1	α	0
D_1	0	$1-\alpha$

Unlike the two previous belief hierarchies, $\boldsymbol{S_1^3}$ can be represented by several distributions $s \in \Omega(s_{1,2}; s_{1,2,1})$. Since P1 believes that it is common belief that P1 and P2 are conditional cooperators, the two only possible outcomes of the game are mutual cooperation and mutual defection. There is however not enough information to determine the *frequency* of each outcome: we have therefore $\alpha \in [0; 1]$. We can check that the expected payoffs for both players is $\big(2\alpha + (1-\alpha)\big) = 1 + \alpha.$

Since P1 believes that it is common belief that P1 and P2 defects if the other defects, we can check that P1 cannot believe that P2 would always choose to defect rather than conditionally cooperate (since P1 would then defect, and P2 would only get a payoff of 1). Note however that, if $\alpha < 1$ (players do not necessarily coordinate on the cooperative outcome), P1 cannot believe that P2 is a conditional cooperator: it would indeed be in P2's interest to always cooperate (guaranteeing a payoff of $2 > (1+\alpha)$) – P1 would therefore revise her beliefs. Nevertheless, if

$\alpha = 1$ P1 knows that being a conditional cooperator is payoff maximising for P2 (even though being an unconditional cooperator would also be payoff maximizing – but in this case P1 could not simultaneously believe that P2 always cooperates, and that P2 believes that P1 is a conditional cooperator, because P2 should realize that being a conditional cooperator is not payoff maximizing for P1).

$\boldsymbol{S_1^2}$ is therefore a massaged belief hierarchy, because there exists a distribution $s \in \Omega(s_{1,2}; s_{1,2,1})$ such that P1 can rationalize the common belief that both players are conditional cooperators (when $\alpha = 1$). It is noticeable that, although the representations of the belief hierarchy $\boldsymbol{S_1^2}$ and $\boldsymbol{S_1^3}$ are the same (the profile $(C_1; C_2)$ occurs with probability 1), only S_1^3 is a MBH: it is indeed the conditional distributions supporting the distribution over strategy profiles (and not the distribution itself) that are relevant for characterising a MBH – mutual cooperation is rationalizable only if both players are conditional cooperators.

$\boldsymbol{S_1^4}$	C_2	D_2
C_1	0	0
D_1	0	1

In the present case, although P1 believes that P2 believes that P1 is a conditional cooperator, P1 believes that P2 always defect. The only possible outcome is therefore (D_1, D_2), with an expected payoff of 1 for both players. However, P1 cannot simultaneously believe that P2 (i) always defect and (ii) believes that P1 is a conditional cooperator. If P2 believes that P1 is a conditional cooperator, then P2 would be better off by always cooperating (the probability of (C_1, C_2) would be 1), or becoming a conditional cooperator (with a probability of (C_1, C_2) of α). $\boldsymbol{S_1^4}$ is not a massaged belief hierarchy, because P1 cannot rationalize her beliefs $s_{1,2}$.

The last belief hierarchy $\boldsymbol{S_1^5}$ is a bit peculiar, because it suggests that one player could intend to cooperate if and only if the other defects, i.e. that the objective of this player is to reach only asymmetric payoffs. The main issue is however that this 'asymmetric' behavior and being a conditional cooperator are *incompatible* probability distributions. It just means that it is not possible to represent in a matrix a probability distribution consistent with both conditional probability distributions. Since $\Omega(s_{1,2}, s_{1,2,1}) = \emptyset$, $\boldsymbol{S_1^5}$ is not a massaged belief hierarchy.

4.3. Simulation, ratifiability and action-dependent beliefs

We now highlight that the introduction of ST will not only provides a psychological explanation of belief formation, but could also undermine the implicit but problematic identification of Bayes rationality with best-reply reasoning in game theory.

The most common requirement of epistemic game theory – its 'central idea' according to Péréa (2014, p. 13) – is the common belief in rationality (CBR), i.e. that it is common belief that players choose the strategy that maximizes their expected utility. A widely accepted proposition is that CBR implies the iterated deletion of dominated strategies (see e.g. Bernheim, 1984), meaning that a rational player cannot believe that another rational player could choose an iteratively dominated strategy. This result however requires the additional (and often implicit)[107] assumption of ratifiability (Jeffrey, 1990; Levi, 1998), according to which the actions of the players j ≠ i are probabilistically independent of i's choices. Formally:

Definition. *Let* $p \in \Delta(X)$ *be a probability distribution over X. A strategy profile* $x^* \in X$ *is ratifiable if and only if,* $\forall i \in N$*:*

$$\sum_{x_{-i} \in X_{-i}} p(x_{-i}|x_i^*)\ \Pi_i(x_i^*; x_{-i}) \geq \sum_{x_{-i} \in X_{-i}} p(x_{-i}|x_i^*)\ \Pi_i(x_i'; x_{-i}), \forall x_i' \in X_i. (10)$$

A strategy is ratifiable if and only if it gives a higher expected payoff than any other strategy $x_i' \in X_i$, while the probability $p(x_{-i}|x_i^*)$ defining the expected payoff remains the same after the deviation to x_i'. Jeffrey (1990) however argues that ratifiability is not implied by Bayesian rationality, since the maximisation of one's expected payoff implies:

$$\sum_{x_{-i} \in X_{-i}} p(x_{-i}|x_i^*)\ \Pi_i(x_i^*; x_{-i}) \geq \sum_{x_{-i} \in X_{-i}} p(x_{-i}|x_i')\ \Pi_i(x_i'; x_{-i}), \forall x_i' \in X_i. (11)$$

When deviating to x_i', player *i* must consider the possible impact of her choice on the state of the world, i.e. on the strategies of the other players (the posterior distribution $p(x_{-i}|x_i')$ is therefore not necessarily equal to $p(x_{-i}|x_i^*)$, (see Mariotti, 1996, pp. 143-144, for a similar point).

A reason why ratifiability is often conflated with Bayesian rationality is that believing that the actions of the others could depend on one's own action – and therefore that $p(x_{-i}|x_i^*) \neq p(x_{-i}|x_i')$ – seems to be a fallacious mode of reasoning. If Bayes rationality is common belief among us, we should know that our decisions are independent, and therefore that our choices cannot directly influence the choices of others (e.g. Binmore, 1992, pp. 311-312). The direct implication is that players can never play strictly dominated strategies, such as cooperating in a prisoner's dilemma: whatever my belief is about your strategy, defecting is always payoff

[107] Bernheim (1984, p. 1014) explicitly states a similar (but stronger) condition of uncorrelated beliefs: 'the choices of any two agents are by definition independent events […] Consequently, I restrict players to have uncorrelated probabilistic assessments of their opponents' choices'. However, as pointed by Levi (1998, footnote 9) and discussed in footnote 11 Aumann (1987) claims that Bayes rational choices should be ratifiable: this claim then remained implicit in the subsequent literature in epistemic game theory.

maximizing. The common belief of Bayes rationality and ratifiability therefore implies the elimination of iteratively dominated strategies.

However, since our objective is to develop a *psychological* theory of belief formation in games, we should also consider the possibility that players believe that their actions are correlated,[108] and accordingly that their beliefs about the action of others may depend on their *own* actions. *Action-dependent beliefs* (ADB) could indeed explain the experimental findings of Shafir and Tversky (1992), according to which subjects cooperate more often in a prisoner's dilemma when they are not told the choice of the other player rather than when they know that the other has cooperated. While players are best reply reasoners when their belief about the action of the other is fixed (in line with the ratifiability assumption), they are not in presence of uncertainty.[109] Masel (2007) formalises the idea of ADB in a 'Bayesian model of quasi-magical thinking', and demonstrates that in public good games a positive correlation between the players' contributions and their beliefs about the strategy of others can explain cooperative behaviors. Hammond (2009) also defended ADB as a case of 'rational folly', since although players know that their actions cannot directly influence the actions of others, they could be better off if they actually hold that belief: it would therefore be rational for them to hold irrational and false beliefs (see Lecouteux, 2015, propositions 11 and 12, for an evolutionary justification of 'rationally irrational' behaviors).

Furthermore, ADB are not incompatible *per se* with Bayesian rationality. The choices of the players are actually independent (choosing the strategy that maximizes my expected payoff cannot directly influence the choice of your strategy) but the players can believe that the strategy of the other players depends on their own action. They can furthermore *rationalize* this belief – at least when they form their beliefs by simulating the reasoning of other Bayes rational players.[110]

Proposition 5*. If there exists a strategy profile* $\bar{x} \in X$ *that Pareto-dominates a Nash equilibrium* $x^* \in X$*, then* $\bar{x}$ *is a SBE.*

---

[108] Note that Aumann (1987) defends the idea that beliefs can be correlated (suggesting that the choices of the player could be dependent), but still defines Bayes rational choices as ratifiable choices. His justification is that the choice of the individual is a two-step process (Aumann, 1987, pp. 3–4): given an initial probability distribution over the set of profiles, the individual is firstly informed of her strategy (which gives her a belief about the strategy of the others – this belief can therefore be conditional on the strategy suggested to her), and then independently chooses her best reply given her beliefs. Player i can choose a new strategy without affecting her belief, because the actions of the other players are 'fixed' to the state of the world revealed to the player in the first stage. This interpretation however requires the existence of a public signal – e.g. a social norm – to inform all the players of the profile selected in the first stage: actions are therefore independent, while beliefs may be correlated by the public signal. This two-step structure was initially suggested by Harsanyi (1967–1968), but his argument was that 'nature' selects in a first stage the type of the players (their utility functions) only, and not their strategies.

[109] Similar predictions are found with Newcomb's problem by Gardner and Nozick (1974) and Shafir and Tversky (1992): around two-third of the individuals chose to take only one box, and not to play the dominant strategy of taking the two boxes (as if they believed that their action could actually affect the probability of getting a high outcome in the 'risky' box).

[110] Board (2006) for instance shows that Aumann's (1987) notion of Bayes rationality is equivalent to causal rationality (corresponding to our equation (11)) only if we add a condition of causal independence. Our point is that ST could question the players' belief in the causal independence of their actions. Indeed, if causal independence were common belief among the players, then our proposition 5 would not hold (see Hédoin, 2016, pp. 11–15, for a related point).

Proposition 5 can be seen as a generalization of Binmore's 'fallacy of the twins', since it means that the players can always rationalize the choice of a profile that Pareto dominates the Nash equilibrium (just as in a Prisoner's Dilemma). According to Binmore (1992, pp. 311-312):

> it is false that rational players can restrict their attention in the Prisoners' Dilemma to the main diagonal of the payoff table … This would only make sense if the two players did not reason independently. If player I could count on player II reasoning precisely as he reasons, then it would be as though he could force her to choose whichever strategy he found expedient simply by choosing it himself. [But if two rational players] reason in the same way in identical circumstances, it is not because they have no alternative but to think identically: it is because the rational thing to think is the same in both cases.

While we agree that two players who 'reason in the same way in identical circumstances' could play a Nash equilibrium (this is precisely our proposition 4), it is only because the beliefs in their massaged belief hierarchy are action-independent, and not because their reasoning processes are independent. Binmore's definition of rationality in games is indeed that players are best-reply reasoners, while – in line with Kadane and Larkey's initial claim – a Bayesian theory of choice in games should not restrict the definition of admissible strategies to ratifiable choices.

Proposition 5 also suggests that the set of SBE is potentially quite large, and therefore that players who do not converge on the same belief hierarchy during the massaging process are likely to miscoordinate. This translates the idea that the equilibrium in our model is not a hypothesis but a result of the model. From that prospect it is indeed perfectly natural to assume that in reality, in a descriptive account of games and of strategic reasoning, the set of possible solutions is large. We found the same idea in von Neumann and Morgenstern (see chapter 1). This also explains the importance of the mindshaping hypothesis and of the players' social background that can ultimately reduce of space of possibilities. A Prisoner's Dilemma has for instance two SBE:[111] mutual cooperation (with the underlying belief that both players cooperate if and only if the other cooperates) and mutual defection (with the belief that both players unconditionally play their Nash strategy). Player 1 could for instance believe that reciprocity is the norm when playing a Prisoner's Dilemma, and that players are more likely to converge during the massaging process on the beliefs that both players cooperate if the other cooperates. However, if player 2 believes that the norm is to coordinate on the Nash equilibrium or to play best reply strategies, then the massaging process will lead her to another belief hierarchy, and we could end up in an asymmetric profile (which cannot be a SBE, since the players did not accurately predict the choice of the other players).

5. Extending the players' choice problem in large worlds

[111] We can indeed show that all the players get at least their maximin pay-off at a SBE, which excludes the asymmetric profiles from the set of possible SBE. If player i has less than her maximin, then the underlying belief hierarchy cannot be a MBH: it would indeed be in the interest of i to unconditionally play her maximin strategy (the other players cannot therefore rationally believe that i would play her part of the SBE).

The analytical framework we develop in this section should (i) explain how the players translate a large world into a small world, (ii) how they tend to identify the same focal points through the cognitive homogenization of the environment, and (iii) characterize the beliefs that the agents may form as a way to rationalize the choice of focal points.[112] We will deal with the first part by considering that each individual is *aware* of a certain number of 'properties' of the outcomes, and have a ranking of those properties based on their salience. We will show that those two ingredients are sufficient to ensure that players choosing the most salient outcome choose as if they were maximizing a strictly increasing payoff function. We then suppose that the individuals are able to distinguish between their own perception of the game (and the individual focal point they identify), and what they think is the shared perception of the game (like Schelling, see Orléan, 2004). An important addition to standard Bayesian game theory is that we next assume that this shared perception offers a basis for defining the prior beliefs of the players. We then show that, given this prior belief, it is indeed individually rational for the player to play their part of what they identify as a social focal point. This section explains the origin of the players' priors (that in no way correspond to players' prior as standardly defined in epistemic game theory) but as form of gut feelings before the massaging process corresponding to the strategic reasoning phase of the game, described and formalized in the section 3 and 4 of this chapter.

5.1. Preliminaries

We define *GW* as the grand world associated to a strategic interaction involving a set $N = \{1; \dots; n\}$ of players. We formalize how a player $i \in N$ will represent *GW* as a small world SG_i (note that we index by *i* because it is not sure that *i*'s representation of the grand world will be the same than the other players – cf. the case of Brexit above).

In the grand world GW, each combination of GW-acts determines a possible outcome of the strategic interaction (e.g. each of us stands in a precise GPS coordinate, at a precise altitude in Paris). The exhaustive set of outcomes is assessed unconsciously according to a set of *properties P*. Each property assigns a value to a combination of GW-acts: this can either be 0 or 1 (i.e. the property is satisfied vs. it is not), or a ranking of different values. We can thus partition the set of combinations of GW-acts, depending on the properties that are satisfied (or more generally, depending on the index associated).

In the case of the meeting in Paris, a possible property is 'well-known touristic places in Paris', which ranks e.g. the Eiffel tower first, then Notre Dame, the Louvre museum, etc.; another could

[112] As emphasized by Casajus (2000, pp. 264-265), this two-step process of describing how players (i) frame the game, and (ii) choose once the problem has been framed, "is an important step towards the rationalization of salience", though "this kind of rationalization of choosing the salient is to be distinguished from the formalization of salience itself" (see also Goyal and Janssen 1996).

be 'nice places to have lunch', with the ranking of your favorite restaurants. Note that, in the 'linguistic' definition of the two properties we just introduced, it is likely that two individuals may consider different rankings for e.g. touristic places (e.g. rank Montmartre before Notre-Dame). To avoid such confusions, we will define a property analytically as a function $P(j): \mathcal{A}_i \times S \mapsto \mathbb{N}$. A property $P(j)$ associates an index from the set of natural integers[113] to each GW-act for a player *i*, given a possible state of the world (i.e. the action of the other players) – equivalently, this corresponds to a vector of GW-acts.

Each player is aware of a certain number of properties, that we will call her *awareness set* $\mathcal{P}_i = \{P(j)\}_{j \in J_i}$ with $J_i \subseteq \mathbb{N}$ a set of indices (which can be finite or not). From the perspective of player *i*, it is thus possible to partition the set of outcomes depending on the properties satisfied. If a combination of GW-acts does not satisfy any property, then the player is not aware of this specific outcome (e.g. if you are not an economist, you will probably not be aware that the campus of Paris School of Economics is a possible meeting point).

The two basic ingredients of our framework are: (i) the awareness set, $\mathcal{P}_i = \{P(j)\}_{j \in J_i}$, and (ii) a complete and transitive relation $\succcurlyeq_i$ over this awareness set. The resulting ordering of properties is called the *checklist* of player *i* and is denoted $\mathcal{C}_i = \{P(j)\}_{\succcurlyeq_i}$. The term of 'checklist' comes from Mandler et al (2012), who study the behavior of individuals who choose by following a checklist (i.e. I first compare the two alternatives according to the first property, and if they are equivalent, I use the second property, etc.). The relation $\succcurlyeq_i$ is interpreted as the ranking in terms of salience of the different properties: at noon, the property 'nice place to have lunch' may typically be more salient than the property 'touristic place'.

We also introduce *i*'s belief about *j*'s checklist, denoted $\mathcal{C}_{ij}$ (i.e. the checklist that *i* believes *j* has in mind), and *i*'s belief about the *common* checklist $\mathcal{C}_{iC}$, the checklist that *i* believes she 'shares' with the other players, i.e. a checklist that *i* believes is common belief among the players. This common checklist is not necessarily the checklist of a given individual, it can be seen as the checklist of the 'group' of players, i.e. what distinguishes them as a particular group.[114] While $\mathcal{C}_i$ and $\mathcal{C}_{ij}$ are necessary to define how *i* will represent the initial problem as a small world, $\mathcal{C}_{iC}$ will determine the players' prior beliefs about the outcome of the game.

5.2. From large to small worlds

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

[114] This capacity to be aware that her own mental states, like her beliefs, may be different than the collective beliefs and that within the collective individuals may also have different individual and personal beliefs, i.e. acknowledging at the same time the existence of these collective beliefs but not adhering to them, is conceptualized by Orléan (2004).

As discussed above, $\mathcal{C}_i$ gives a procedure to partition the set of GW-acts. The player then derives her set of strategies from the set of GW-acts she is aware of. A *strategy* is a set of GW-acts for player *i* such that, for all the GW-acts of players in *-i*, the corresponding set of combinations of GW-acts has the same properties. In other words, the partition defined above allows the player to define indifference classes over GW-acts. For instance, if I am aware of the outcome 'we both meet at the Eiffel tower', I define 'go to the Eiffel tower' as a strategy, though it corresponds to a large set of GW-acts (i.e. all the GPS coordinates and altitude corresponding to the Eiffel tower). The question of how to choose a GW-act among this indifference class is left for after, once the small world has been solved – rather than defining immediately my precise choice, I treat the problems sequentially, by first finding a good spot in Paris, and then a good spot at the Eiffel tower. This second problem is the large world that remains to be solved once I am at the Eiffel tower (and again, I will frame it as a small world, and look for salient outcomes such as 'meeting at the south leg', 'meeting at the ticket office', etc.).

Once the player has framed her own set of strategies, she will imagine how the other players frame their own set of strategies. She will *simulate* the framing of player *j* by following the exact same procedure than the one described above (*i* indeed assumes by default that her own reasoning is a good predictor of the reasoning of *j* – this constitutes the basic principle of simulation thinking), while taking $\mathcal{C}_{ij}$ as the input. If I am asked to meet a foreign tourist in Paris, there is little reason to expect that she will be aware of the outcome 'meet at Chatelet les Halles', but only of the outcome 'meet at the Eiffel tower'. If the other individual is Parisian, it seems more reasonable to assume that she will be aware of both outcomes, though she may not be aware of the outcome 'meet at Paris School of Economics'. By simulation, player *i* repeats the same operation than for her set of strategies, but by considering a partition of the GW-acts only with her belief about *j*'s checklist $\mathcal{C}_{ij}$.[115] We now have a set of strategy for player *i* and for the other player.

Given the awareness set of player *i* and the set of strategy profiles *X* defined above, we now investigate the payoff that player *i* attributes to each strategy profile. When considering all the possible strategy profiles (which, by construction, do not satisfy the same properties), the player will rank them according to whether they satisfy the properties she finds the more salient. This means that the ranking of the strategy profiles $\rhd$ induced by the checklist $\mathcal{C}_i$ represent the relative salience of the different outcomes the player is aware of.

[115] In principle, player *i* can also consider that there are properties that *j* is aware of, but not her – e.g. 'meet at Institut Marie Curie', which would be salient for some doctors but probably not for an economist. We will not discuss such cases here, and always assume that *i* reasons as if she necessarily knew at least as much as the other players. This echoes Bacharach's assumption when he assumes that the sets of frames and beliefs that a player attributes to the other are necessarily the same or susbets of her owns.

Proposition: Let X be the set of strategy profiles as defined above, and $\mathcal{C}$ a checklist. There exists a strictly increasing payoff function $\Pi: X \mapsto \mathbb{R}$ such that $\Pi(x) > \Pi(x')$ if and only if $x \rhd x'$.[116]

The proposition above means that we can define a payoff function from our basic ingredients, i.e. that given the player's checklist and her beliefs about the checklist of the other, we can define (i) a strategy space, and (ii) a set of payoff functions for each player. This means that we have derived a small world representation from the initial large world.

An interesting point of the proposition is that the payoff function is strictly increasing: in other words, we cannot find two profiles *x* and *x'* such that their payoff is the same. This is due to the way the small world representation was build: if two GW-acts satisfy the exact same properties, then they are virtually undistinguishable from the perspective of player *i*. It is therefore possible to treat those two acts as part of the same 'event' (in Savage's terminology), which implies that two identical acts should be considered identically by the decision maker. This directly echoes the principle of 'Equal treatment of payoff-equivalent strategies' of Mariotti (1997): if two strategies yield the exact same payoff for player *i*, then there is no reason for a player *j* to consider that player *i* may attach a higher probability for one of the strategies than the other. Two actions yielding the same payoff can therefore be treated as the *same* strategy.[117]

5.3. Focal points

We can now define our notion of focal points. Intuitively, a focal point refers to a strategy profile that is more 'salient' than the others. Since the payoff function defined above determines the profile that satisfies the properties that the agent finds salient, the function $\Pi_i(x)$ gives us a measure of the salience of each profile, from the perspective of a player *i*.

Definition 1. Let $G_i(\mathcal{C}_i) = \langle N, X, \Pi \rangle$ be the small world defined by player *i* representing the strategic interaction *GW*, with *X* the set of strategy profiles, and $\Pi_j, \forall j \in N,$ the payoff function of player *j*, when the set of strategies and payoff functions are defined with respect to *i*'s checklist. The strategy profile $x^* \in X$ that maximizes $\Pi_i(x)$ is called *i*'s *individual focal point*.

Definition 2. Let $G_i(\mathcal{C}_{iC}) = \langle N, X, \Pi \rangle$ be the small world defined by player *i* representing the strategic interaction *GW*, with *X* the set of strategy profiles, and $\Pi_j, \forall j \in N,$ the payoff function

---

[116] The proof of the proposition follows directly from the theorem 2 of Mandler *et al* (2012) on 'multivalued checklists', since $\rhd$ is generated by the exact same procedure. The existence of the function is guaranteed by the countable number of properties, while the strict increasing is due to the construction of the small world, and the fact that two profiles generating the same payoff are considered as the same strategy, by definition of the set of strategy profiles (which are indifference classes over GW-acts).

[117] This is also consistent with the sequential approach of dealing with large worlds: once the small world has been solved (and a strategy x_i chosen), the player will realise that she must now refine her choice (e.g. go to the south leg of the Eiffel Tower), because the initial act included too many GW-acts.

of player *j*, when the set of strategies and payoff functions are defined with respect to *i*'s belief of the common checklist. The strategy profile $x^* \in X$ that maximizes $\Pi_{iC}(x)$, the function that represents $\mathcal{C}_{iC}$ is called *i*'s *subjectively based social focal point*.

Definition 3. Let $G_i(\mathcal{C}_{iC})$ be the small world defined by player *i* representing the strategic interaction *GW*, with X^i the set of strategy profiles framed by player *i*. A strategy profile x^* is an *objective social focal point* if and only if:

$$\forall i \in N, x^* \in X^i$$

$\forall i \in N,\ x^*$ is a subjective social focal point of the game $G_i(\mathcal{C}_{iC})$

The difference between the individual and subjectively based social focal points captures the idea of strategic framing, since the individual does not only pay attention to her own payoff (i.e. the outcomes that seem salient from *her* perspective), but restricts it to what she thinks is the shared perception among the group. Note also that, since the payoff functions are *strictly* increasing by construction of the small world itself, there necessarily exists a unique individual focal point, as well as a unique subjectively based social focal point. The objective social focal point exists if and only if the strategy profile in question is conceivable by all the players (for instance, a tourist will not be aware that Châtelet les Halles is a possible profile), and that they all identify this profile as their subjectively based social focal point – which may still be different from their individual focal point.

The definition of a social focal point may look restrictive, since it could seem that we require it to be a payoff-dominant outcome (like HH in Hi-Lo or cooperation in a Stag Hunt). It must however be emphasized that the payoffs we traditionally associate in a Hi-Lo or Stag Hunt do not capture all the features of our notion of 'payoff' (which is multidimensional and captures the different properties of the checklist). This means that, when we ask players in a lab experiment to choose based on a payoff matrix, this is still a large world: we do not consider how the players will interpret what seem to be irrelevant features of the problem, such as the label of the strategies. We emphasize that the matrix representation we use in practice are not necessarily the representations used by the player – and that the labels we may arbitrarily add to the strategies for sakes of readability, cannot be used by the players we model (this is a label for *us*, the theorists, and not something useable by the player). A payoff dominant outcome in our framework is not a payoff that gives the highest monetary rewards to both players: it is the most salient profile, once we considered all the various aspects of the problem (individual monetary gains, the repartition of those gains, the label of the strategies, etc.). Our notion of 'payoff' is close to a behavioristic interpretation in the sense that it represents 'all-things-considered preferences' and is thus tightly related to the individual's choice, though it is used as an input of decision-making: we do not define the payoffs from the behavior of the players, but from their mental states.[118] There is then

[118] A difficulty that may arise is then to properly calibrate the payoff function in the model, depending on the various properties that may motivate the players. In many lab experiments, we can probably confidently assume that the main property considered by the individuals is their monetary gains, but this assumption should not be systematic, in particular in games where concerns of fairness may be decisive.

a close relation from our notion of payoff to the choice of the player – as the *only* motivating reason for the player to act – but it is not tautological as with behavioral preferences. The fact that there exists a payoff-dominant profile simply means that there is a form of alignment of the players' interests, because they both identify the same outcome as the most salient.

Another possible concern is that our definition suggests that any game can have a social focal point: even though this assumption is realistic in matching games, it becomes much less convincing in e.g. zero-sum games. This is because – to keep the mathematical presentation in its simplest form – we did not consider for now the possibility of mixed strategies. We could echo here the distinction between the 'massaged belief hierarchy' and 'subjective belief equilibrium' in section 2: the massaged belief hierarchy corresponds to the equilibrium beliefs (possibly in correlated strategies) that the players may hold, while the subjective belief equilibrium is only considered as the equilibrium in pure strategies when the players correctly anticipated the behaviors of others. A simple extension of the definitions above would be to allow correlated strategies for the definition of subjective social focal points (I can for instance believe in a matching game that we should alternate between two equilibria over time), while keeping a definition in pure strategies for objective social focal points (since they must actually occur, to be able to shape the beliefs of the other players). The implication would be that there could always exist a subjective social focal point for zero sum games (e.g. the mixed Nash equilibrium), while there is not one objective social focal point.

5.4. Mindshaping and the formation of prior beliefs

We have seen thus far how players can frame a large world into a small world, thanks to their own checklist (this part of framing is largely unconscious, as in Bacharach's variable frame theory), and to their beliefs about the checklist of the other. We have also seen that we can always define an individual and a social subjective social focal point in this framework, by construction of the payoff functions Π_i and Π_{iC}. The question that must now be solved is how players choose, once they have framed the initial problem as a small world. A recurrent point of discussion is indeed whether it is rational or not to play one's part of a focal point. The attempt to rationalize focal points play on purely individualistic grounds usually faces an infinite regress problem (as it is not clear why it should be commonly believed that we all coordinate on the focal point rather than another profile; see e.g. Sugden [1991]). We must usually add an *ad hoc* principle to justify the selection of Pareto-superior profiles, such as Luce and Raiffa (1957) 'solution in the strict sense', Harsanyi and Selten (1988) payoff-dominance principle, Bacharach (1993) 'principle of coordination', or Janssen (2006) 'principal of individual team member rationality'. An alternative way is allowing a form of collective agency, and to consider that players may think of themselves as part of a group, who should play their part of the collectively rational profile – such theories of *team reasoning* have been proposed by Sugden (1991, 1993) and Bacharach (1999,

2006).[119] Another road, which avoids possible debates about the nature of collective agency, is to consider the case of *evidential reasoning*, and that players may typically condition their beliefs about the others' actions on their own actions (but this remains a controversial position, see Weirich (2016) for a discussion).

The solution we suggest is that subjectively based social focal points offer a basis for the prior beliefs held by the players. This does not suggest that playing focal points is 'rational', but merely that, since we tend to spontaneously identify the focal point as the obvious solution of the game (by definition of the focal point, as something that spontaneously comes to your mind), it seems empirically relevant to consider that our prior gut feelings when we come to play a game are derived from this first intuition. Our argument is therefore that, if a profile – for any material, cultural, historical, or other reason – seems to be the one that 'we', the group of players involved in the game (Parisians, colleagues, everybody…), identify as the obvious one, then we should start any game-theoretic reasoning based on this prior belief. Given their prior beliefs, the players may then start to think strategically by simulating the reasoning of the other players, as in section 3 and 4. They should check that their prior belief is consistent in a Bayesian sense (I cannot believe that you play for instance a dominated strategy, nor that you could believe that I play a dominated strategy). This process of simulation will lead the players to 'massage' their prior beliefs until they become consistent – and the existence of a massaged belief hierarchy derived from this prior belief is ensured by the finiteness of the small world. Once their beliefs are consistent, the players simply maximize their expected payoff (defined thanks to their checklist $\mathcal{C}_i$). Recall that simulation thinking may be coherent with the rationalization of non-ratifiable choices, meaning that this process may lead to the formation of action-dependent beliefs, which *in fine* could for example rationalize the choice of cooperation in a prisoner's dilemma, as soon as the underlying belief is that the players are both conditional co-operators.

As soon as we recognize that players are socially-embedded agents, who necessarily (voluntarily or not) belong to certain groups, directed by various norms and rules (translated in our framework as various *common checklists*), then considering that those various social influences shape our prior beliefs seems to be a reasonable assumption. In this perspective, we could relatively easily model the role of mindshaping in a game-theoretic setting. Since the prior belief of player *i* corresponds to the subjectively based social focal point (the profile *x* that maximises the function Π_{ic}), it will be sufficient to study the evolutionary dynamics of prior beliefs, and to consider that changes in prior beliefs are the result of an underlying change in $\mathcal{C}_{ic}$, and therefore of the rules and norms that *i* believes are shared by the group of players. An objective social focal point would be reached only once the prior beliefs of the players become aligned, and all identify the same strategy profile as their subjectively based social focal point. Formally, we can easily model the evolutionary dynamics as a multi-population game, with each population corresponding to one individual with her collection of possible prior beliefs. Since stability requires the convergence of the prior beliefs to a single distribution, the underlying evolutionary game has a structure of a matching game, with several strict Nash equilibria. The only

[119] See Lecouteux (2018b) for a review, and Gold and Colman (2018) for a discussion of team reasoning and the choice of payoff-dominant profiles.

evolutionary stable states in this setting are the strict Nash equilibria (e.g. Weibull, 1997), i.e. when all the players have the same prior beliefs. Given the simple structure of the game, we can expect a rapid convergence towards a stable state, depending on the initial conditions and the model considered for the replicator dynamics. Note also that the focal point in question should be a subjective belief equilibrium.

6. Conclusion

It has been argued in this chapter that offering a Bayesian theory of choices in games should integrate a psychological theory of belief formation. Rather interestingly, Harsanyi (1982a, p. 122, emphasis in original) initially recognized that a theory of choice in which the rationality of the players is not certain would require:

> "an empirically supported *psychological* theory making at least probabilistic predictions about the strategies people are likely to use … given the nature of the game and given their own psychological make-up"

He also acknowledged that looking for a psychological or normative theory of games are "very different intellectual enterprises, using very different methodologies as a matter of logical necessity" (Harsanyi, 1982a, p. 122) – this is why Kadane and Larkey (1982b, p. 124) declared "our differences with Professor Harsanyi are not as profound as might appear". Harsanyi (1982b, p. 125) however argued that normative game theory could provide a solid foundation for a descriptive theory of games, since the actual choice of the players is "*either* … the *correct* move prescribed by normative arithmetic, *or* … a psychologically understandable deviation from it." Aumann's reply (in Aumann, 1987) to Kadane and Larkey turned out to be remarkably similar, since the prior beliefs he attributes to the players are an equilibrium of the game. We however argued that Aumann's argument that the essence of game theory is to impose CBR on the set of prior beliefs is problematic, because the identification of the choices compatible with CBR requires solving the game before actually playing it. If players are Bayes rational, the only options available to them are indeed those that maximize their expected payoff but identifying the set of rational strategies requires defining *ex ante* the beliefs of the players. If Bayesian rationality is common belief, the prior beliefs must be the equilibrium of the game (players cannot indeed simultaneously believe that the others are rational, and that they could play a non-rational strategy). A solution to escape this infinite regress is to investigate the process of belief formation, i.e. how players actually identify the beliefs they attribute to the others.

We argued that game theory did not provide the adequate tools to explain the formation of these subjective beliefs, and suggested introducing the players' capacity of mindreading in game theory. We then assumed that players form their beliefs by simulating the reasoning of the other players, and showed that the belief hierarchy of Bayes rational players, when their rationality is common belief (by proposition 1) does not necessarily rule out non-ratifiable choices, and therefore action-dependent beliefs.

In presence of uncertainty, a player who simulates the reasoning of the other can indeed take her own choice as an evidence of the choice of the other, because she tends to assume that the reasoning processes of the other are similar to her own reasoning processes.[120] We were finally able to derive a solution concept – the subjective belief equilibrium – capturing the mutual Bayes rationality of the players.

An apparent difficulty of this approach is the multiplicity of subjective belief equilibria. An acceptable solution concept should indeed select a restricted number of strategy profiles. However, proposition 5 implies that the set of SBE is quite large, because of the multiplicity of subjective beliefs to which players can converge during the massaging process. A descriptive theory of choice in games should therefore go beyond the mere mathematical representation of games, and not try to give a rational determinate solution to games. Pure deductive reasoning is insufficient to give determinate solutions in the majority of games (see e.g. Schelling, 1980[1960], p. 163; and Sugden, 1991, in the context of bargaining games). We therefore understand the crucial role played by the context of a game in its resolution and accordingly the function of mindshaping which provides the basis for common understanding and common reasoning in a given context, i.e. homogenization. Considering that the players have been shaped by the same social and cultural models, by the same interactional models, mindshaping provides a common intersubjective background among players and accordingly allows successful coordination by possibly drastically restricting the set of SBE.

Therefore, rather than discarding the social, cultural, psychological and historical background of the players, and look for solutions that rational players could not fail to find, a descriptive theory of games should investigate how those backgrounds affect the formation of individual beliefs (through the definition of focal points and social norms for instance, see e.g. Schelling, 1980[1960]). A descriptive theory of strategic choice should thus rest on works in cognitive and social psychology, so as to characterize the formation of the initial gut feelings of the players, which are determined by personal and social experience (see Bacharach, 1990; Scazzieri, 2008, 2011). Neglecting those factors "deshumanizes the decision-maker in the opposite direction to the traditional idealization of her powers: instead of exaggerating her resources, it understates them" (Bacharach and Hurley, 1991, p. 3).

Mindshaping offers a possible analytical tool to extend game theory to the integration of those socio-cultural factors ultimately explaining the occurrence of a particular outcome among all the potential ones (i.e. among the possibly large set of SBE). Colman and Bacharach (1997, pp. 8–9) for instance assume the transparency of deliberation and not merely the transparency of reason, which is captured through CBR. Various works in social psychology, like Zawidski's conceptual

---

[120] We can mention Segal and Hershberger's (1999) experiment on cooperation between twins to support our hypothesis that 'similar' individuals are more likely to form ADB (they indeed find that monozygotic (identical) twins are more likely to cooperate in a prisoner's dilemma compared to dizygotic twins). Furthermore, several studies on social identity theory (Tajfel and Turner, 1979) suggest that players who think of themselves as members of a common group are more likely to cooperate in social dilemmas (e.g. Kramer and Brewer, 1984). A possible explanation of those results is that 'similar' individuals (whether it be socially or genetically) are more likely to believe that their actions are correlated, and therefore to form ADB (which could then explain the higher rates of cooperation in the prisoner's dilemma).

work on mindshaping, "have revealed a remarkable degree of consensus in people's understanding of their social environment" (Colman and Bacharach, 1997, p. 9), and research in attribution theory – a theory in social psychology explaining how people attribute second-order beliefs, which is tightly related to the Theory of Mind (see Bacharach, 1986) – has "shown that the same basic cognitive processes underlie people's predictions and explanations of their own behaviour and that of others" (ibidem). Different works in cognitive sciences, adopting different methodologies or even epistemolies, indeed agree with this role of homogenization provided by social. We can no longer discard players from this undeniable and necessary social dimension in interactional decision problem. Beyond understanding how the others reason (thanks to the transparency of reason), the individuals can understand how the others discriminate between different 'rational' alternatives (thanks to the transparency of deliberation), and therefore can better predict their actual choices.

The aim of the extension of the analysis of the formation of players' beliefs to cases of incomplete information is therefore to offer a theory of coordination that considers the role of the context within which the individuals interact and in which the role of mindshaping in players' reasoning is decisive. Those two dimensions strongly impinge on players' decision. This required attacking the problem of choosing in large worlds, i.e. of explicitly dealing with decision problems as open rather than closed systems. This indeed allowed us to highlight the role of the processes of mindshaping in the formation of the preferences and beliefs of the individuals, both as influencing the way the players conceptualize their decision problems (when framing the initial large world as a self-contained small world), and as influencing their gut feelings, and hence their prior beliefs in the small world representation of the decision problem. Unlike within the standard approach to coordination games, the players can identify *ex ante* a simple solution to the game, and they need not form complex belief hierarchies prior to their choice.

One limit of our approach could be the hypothesis of common rationality even if it is justified by the simulation theory that players attribute to the other their own mode of reasoning and accordingly that in some way we impose a kind of common rationality like in standard game theory.

We remain anchored to an eductive mechanism which still impose some depth of reasoning even if they are far less restrictive than standard and epistemic game theory. We thus leave less space to the evolutionary approach than Bacharach in our formalism while admitting that the explanation of the players' prior beliefs must be determined by an evolutionary model that selection the beliefs that the players have about each other's beliefs and choices. We assume instrumental rationality but we could however rely on different form of rationalities. Our model is broad enough to encompass many hypotheses of mode of reasoning: other regarding reasoning, altruism, commitment, etc. We could have assumed different type of intentions rather than maximizing her expected payoffs. Such possible level of generality is permitted by the integration of the simulation theory in our model to determine the players' beliefs.

However the kind of interaction assumed in our model remains strategic and founded on the rationalization of players' beliefs even if we get rid of the standard Bayesian justification of the existence of such beliefs. Our approach is compatible with modes of strategic interactions that

accept the existence of rules of behavior and accept that collectively following rules of behavior can be rational, i.e. in everybody best interest.

We also assume implicitly, like Schelling, that players want to coordinate and that in reality it is always in their interest to coordinate

General Conclusion

Our thesis supports an ontological thinking on strategic reasoning in game theory. By claiming that game theory and games should no longer be appraised as mathematical objects, we induce a change of perspective and a change of meaning of strategic reasoning which should be understood as a real process of reasoning in which players' mental states must be incorporated, as a process of coordination explaining how an why a specific equilibrium of individual beliefs occurs. The main claim of the thesis is that such progressive convergence of individual beliefs with respect to other's choice or behavior is the subject of analysis of game theory. In that perspective we gather contributions that comprise the kind of critical analysis allowing assessing the type of intersubjectivity involved in standard game theory. Our thesis thus offers a new understanding of this strategic reasoning process and shows that it necessitates having a new epistemology of game theory founded on a new appraisal of the type of intersubjectivity involved in strategic reasoning, i.e. the intersubjectivity involved in the progressive congruence of individual's beliefs. Such methodology conduced in the thesis founded on a new ontological approach of game theory and of games entails first the understanding of strategic interaction as a real process of reasoning which implies the incorporation of players' mental states leads in consequence to rethink the concept of intersubjectivity and the kind of intersubjectivity involved in strategic reasoning. The thesis ultimately proposes some paths of thinking, some solutions to answer the question of how and why an equilibrium occurs and thus provides some possibilities to overcome the indeterminacy problem. The change of epistemology emphasized must be considered alongside interdisciplinarity as the incorporation of players' mental states require considering the investigations of cognitive sciences, of philosophy, of history, and of sociology. To understand intersubjectivity and circumscribe the type of intersubjectivity involved in strategic interactions understood as a real reasoning and interactional process, it necessitates investigating the interplay between individual, i.e. personal, and collective, i.e. social dimensions, which again involve the incursion of cognitive sciences of philosophy, of history, of sociology, etc. in the analysis of games and game theory.

We claim in this thesis that considerable progress can be made regarding the analysis of coordination in game theory only by resorting to an ontological thinking on strategic reasoning and coordination. We disagree on this point with Sugden (1995, p. 534) who assert that those progresses regarding the problem of equilibrium selection have been accomplished by evolutionary game theory and by the introduction of a preplay communication before the game (see respectively Grawford, 1991; Ellison, 1993; Kandori et al., 1993; and Young, 1993; and Farrell, 1987; or Van Damme, 1989). We argue that neither of these approaches introduces thinking on the meaning of coordination. Neither of these approaches resorts to an ontological thinking on strategic reasoning and coordination, as the players' strategic reasoning process is not explicitly integrated. The solution proposed in these approaches remained focused on the mathematical constraint of equilibrium selection. The progresses made are therefore from our point of view rather limited. The explanation of coordination as a process of convergence of

individual perceptions starting from an explanation of the elicitation of players' beliefs is inexistent in these approaches. Some other approaches entail thinking the meaning of strategic rationality; some others integrate players' reasoning and their mental states (their beliefs or perceptions), or justify the beliefs they hold regarding others' players, etc. these are respectively the focal point literature and labeling games, the psychological game theory and the level-k theory. However, neither of these other approaches rely on a real form of interdisciplinary as it is defended in this thesis and thus remain, again, limited in their capacity to enhance our comprehension of coordination in game theory and thus to answer to the issue of how and why an equilibrium occur. This underscores how promising can be the integration of the theory of mind in game theory and the enrichments it can provide. In line with the achievement of both highly original and challenging scholars: Schelling and Bacharach, we attempted to show that contrary to some of the folk belief in economics, some enrichments of game theory drawing on psychological and cognitive aspects of players' reasoning grounded on the recent of development of neurosciences can be included in the game theoretic framework, without endangering its validity.

However bringing a theory of mind understood as a mechanism of interpersonal understanding into game theory needs the acknowledgement that game theory and economics cannot be conceived independently from the other social sciences since the former is at the core of the interdisciplinary cognitive sciences. It requires a conception of economics as an open science (see chapter 2 for an explanation of economics seen as an open science). It also implies that the players' cognitive process must not be considered as mere black boxes, contrary to standard and epistemic game theory. Game theory should not be immune to an inquiry of such black boxes. Game matrices must not be considered as mere representations of players' choices. The behavioristic account of games has to be challenged.

To understand the results of the thesis with respect to a new approach of intersubjectivity and the contribution of the thesis to the understanding of coordination as the process of strategic reasoning explaining how and why a specific solution emerges, let us recall the respective contributions of the different chapters of the thesis and how they progressively contribute to such renewal of the type of intersubjectivity involved in strategic reasoning in game theory.

The first chapter showed that despite the many improvements intended in the history of non-cooperative game theory regarding the existence of a solution and its properties of optimality and stability mainly, the search for *the* solution concept understood in such a mathematical way remained quite disappointing. It showed that the mathematical concern regarding the solution concept entailed that each intended improvements of non-cooperative game theory remained confined to the proposition of new mathematical constraints; to the adding of more mathematical rigor in the properties of the solution of games. Nevertheless the problem of indeterminacy, i.e. of the explanation of coordination remains. This leads to the conclusion that both from a normative, in case of indeterminacy, and from a positive perspective, as in reality people are generally perfectly able to coordinate; game theory faces a serious stalemate.

Such difficulty must be considered alongside the attempt of game theorists all along its history to get rid of psychology and of mental variables. We argue in this first chapter that bypassing such stalemate necessitates investigating the conditions under which a solution can occur; to explain

how and why a solution can occur. For that purpose a change of epistemology is required. We argued that a game must be considered as a process in which the reasoning of the player, also understood as a process must be considered. It ultimately necessitates integrating the players' mental states (perceptions, beliefs and eventually intentions in games). Indeed, games must no longer be considered as mere representations of the players' choices in which, therefore, all is settled: the choices like the players' beliefs.

We raise in this manner, the inconsistency that prevails between the discourse of epistemic game theorists who claim to integrate the player's reasoning and epistemic states, i.e. some mental states in games to explain how a specific equilibrium (a Nash equilibrium, a correlated equilibrium, etc.) and the reality of its formalism (that remain behavioristic due to the representation theorem derived from VN/M and Savage). We thus claim for a mentalistic vision of game theory which entails a change of definition of the game payoffs, from vNM utilities into monetary payoffs, and ultimately a change of ontological perspective with respect to games, i.e. on what is a game. From that perspective, the integration of player's reasoning process towards the equilibrium and the solution, the inclusion of their mental states for that purpose, become possible.

The chapter intends to lay the premise of the thesis which is that the explanation of how and why a specific solution occur requires to see games as coordination processes, i.e. processes of interaction that progressively lead to the convergence of players' intentions, beliefs and then choices. The contribution of Schelling, exposed in the chapter 2, is particularly important in that perspective. It requires an ontological thinking on strategic reasoning in which strategic reasoning must be considered as a coordination process leading to the consistency of choices. But the consistency of choices must not be assumed like in standard game theory, i.e. in both classical and epistemic game theory, but explained. And such explanation of the progressive consistency of intentions, beliefs and choices is about coordination.

One of the conclusions of the chapter is thus that bypassing the mere mathematical conception of game theory is needed. The inclusion of players' mental states and reasoning process requires the inclusion of cognitive psychology, and the ontological thinking on coordination based on philosophy, sociology and psychology. Such ontological thinking on the condition leading to the players' capacity of coordination is more specifically tackled in the second chapter with the work of Schelling. In the same way Schelling expressed the will to systematically integrate interdisciplinarity in game theory to understand what is a game and how the players reason toward the equilibrium.

Chapter 2 thus presents in detail the work of Schelling on game theory and how it helps from a conceptual and methodological point of view to open the way to the integration of interdisciplinarity in game theory, to the integration of a real strategic reasoning process in games, to a reflection on coordination understood as the result of the reasoning process of the players towards a state of consistency among their perceptions, intentions, beliefs and then choices. Chapter 2 shows that he proposed very fruitful amendments of game theory in such a perspective. The main improvements that drive Schelling's contribution are about the players' epistemic states so that it intuited that the best conceptual and methodological framework to understand coordination is epistemic game theory. However as already mentioned and as

explained in the chapter 1 many insufficiencies exist on the epistemic states as understood as mental states in epistemic game theory; many inconsistencies exist regarding the status of the players' beliefs in such formal framework. Chapter 2 thus insists on the way Schelling proposes a psychological explanation of the elicitation of player's beliefs. Schelling's contribution provides a methodological ground to endogenize the players' beliefs.

One dimension that will be of particular importance that appears in Schelling work's is that when subjective and psychological elements of players' reasoning are integrated in games (i.e. when perceptions, intentions, or beliefs, are understood as real mental states contrary to epistemic game theory) and that common knowledge of instrumental or Bayesian rationality is not assumed, he resorts to focal points to explain how and why a specific solution occur. He introduced an intersubjective dimension in place of common knowledge rationality that traditionally ensure this role. Two dimensions play this role the player's capacity of empathy and social determinants: institutional facts. Both these dimensions are at the basis of his account of focal points. Focal point is multidimensional in Schelling's contribution and reveals all the many determinants that impact on player's reasoning in the coordination process and that lead to the convergence of player's beliefs, intentions, perceptions and then behaviors, i.e. to what he defines as the meeting of minds.

In particular the interest with respect to Schelling's account of focal point, is to exhibit the social philosophy behind this contribution, and that will be shared, as developed in the chapter, 3 by Bacharach one of his contemporary. Namely, such social philosophy is that the more a collective is structured by regular interactions and by institutions, or conventions that organize these interactions the more the intersubjective dimensions required in strategic interactions will be developed, and the easier will be coordination. Recall that conventions and institutions organize individual interactions and induce the sharing of perceptions, intentions, and induce regular pattern of behaviors that thus ultimately facilitate coordination. This is the ground of the principle of mindshaping that together with the simulation theory, as a mindreading capacity, provides the cognitive ground for the intersubjective dimension of strategic reasoning that we develop in the thesis.

Chapter 3 showed that Bacharach not only shares the same social philosophy but also formally integrates many dimensions of Schelling's reorientation of game theory in games. So that for instance, as Bacharach himself declares on one of his formal and theoretical contributions presented in the thesis: the VFT, was "an alternative foundation for and … a refinement of Nash equilibria." (Bacharach and Stahl, 2000, p. 220) We hope to have shown that the VFT in fact offer more than a mere refinement of Nash equilibrium when compared to the many refinements solely driven by mathematical concern exposed in the chapter 1. We also hope to have shown that even more can be done with the integration of the ST in games combine with a deep ontological thinking with respect to the content of game theory and of games.

In a comparative way to Schelling, Bacharach opened the frontiers of the theory of games but also expanded these frontiers, showing that many enrichments proposed by Schelling both psychologically and sociologically grounded can be formalized. He indeed incorporated social and cultural determinants in players' reasoning process which are, like for Schelling, the ground of the intersubjective dimension required in strategic reasoning when the players' perception of

the game they face is integrated. When frame intervenes, the rules of the games no longer determine the space of strategies and payoff of the players, as their apprehension of the game is personal and no longer determined by the objective characteristics of the game that are summed up by the rules of the game. Bacharach had thus to find a methodological device to justify how from free and subjective perceptions of the game they face players come to agree on the same equilibrium, i.e. on the same beliefs with respect to each other's actions, i.e. how they ultimately coordinate. He resorts to Schelling's concept of focal point which is explained by psychological and social determinants. Bacharach in particular intuits the role of mindshaping in his understanding of focal points and coordination in the same way as he intuits the role of simulation in the way players' attribute a set of frames and beliefs to the others.

Bacharach, like Schelling, incorporates in game theory a real interdisciplinary approach, more theoretically driven, and imports from cognitive psychology, social psychology, philosophy and logic some methodologies and theoretical results relevant for strategic reasoning. He succeeded in integrating them albeit not without difficulties; the major one, as emphasized in the chapter 3, being to justify the intersubjective dimension on which he relies in his model of games. The purpose of the chapter 4 was to bypass such difficulty, with the integration of mindreading while showing that there is a compatibility with his methodology, i.e. with the use of framing. Thus Bacharach offers, through his methodology and his conceptual and formal contribution in the VFT and then the TR, a theory of games in which the mental states of the players, their psychology is finally integrated contrary, again, to epistemic game theory. Therefore however some grounds are missing in his theory to bypass some of the problems he faces exposed in the chapter and that can be overcome by the introduction of a theory of mind.

Chapter 4 examines the different contributions of the theory of mind within the philosophy of mind and which encompass approaches from cognitive sciences, cognitive psychology, philosophy, and social psychology. It has always been assumed that common knowledge plays the role of the intersubjective dimension necessary for players to form beliefs about the other players' reasoning, beliefs, and choices, without any reflexive thinking on this methodological device that, as argued in the chapter 1 leads to serious difficulties and even inconsistencies in epistemic game theory for instance. There is indeed incompatibility between common beliefs of rationality and Bayesian rationality. The use of a theory mind and therefore the integration of a theory of mind in game theory can provide a new methodological device for the intersubjective dimension of strategic reasoning and an explanation the elicitation of players' beliefs which is, again, one of the major difficulties of the epistemic program of game theory. Recall that many critics are made with respect to the status of players' beliefs and in particular of their prior beliefs.

Recall that the theory of mind has for purpose to explain how individuals attribute to themselves and others some mental states like some beliefs, intentions, perceptions, but also mode of reasoning and behaviors. Different theory of mind exists and the chapter stresses the different premises and assumptions on this mechanism of attribution. In this manner the chapter emphasizes the pertinence of the simulation of theory and in particular the simulationist account of Goldman. The chapter demonstrates that Goldman's ST version is compatible with the methodologies of Schelling and Bacharach; it explains how from their perspective, from their own perceptions of the situation, of the game, and no more information, the player's come to form beliefs about others' perceptions, intentions, beliefs or choices. This is thus compatible with

framing and with the assumptions both made by Schelling and Bacharach that players are human with limited cognitive abilities and with partial information. It is of particular importance to mention the cognitive parsimony of the ST which is particularly relevant nowadays in the tendency in game theory to more and more assume limited rationality and no longer eductive mechanisms that require unlimited depth of reasoning that largely bypass human cognitive capacities. This ST account provides a methodological ground to justify the intersubjective dimension of games and to explain the elicitation of players' beliefs in a framework in which the player's perceptions is taken into account. It therefore allows to properly account for players' reasoning process and to integrate their mental states, and in particular to treat players' beliefs as real mental states contrary to epistemic game theory. The mindshaping hypothesis on the other side ensures that a sufficient degree of homogeneity among people brings sufficient accuracy in the simulation process of other's mental states or reasoning process. This ultimately brings accurate predictions and therefore coordination.

The ST also justifies the relevance of focal points (both subjective and social) in players' strategic reasoning and in the theory of games to explain coordination, i.e. to explain how and why players finally agree on an equilibrium. Endogenous or subjective focal points, which are salient perceptions of a situation indeed tend to be attributed to the others like are the social focal points. The latter ones are understood as a social norm or a convention, which entails common inductive inferences, and are grounded on common perceptions, beliefs and behaviors. This is again supported by the mindshaping hypothesis. Recall that mindshaping practices serve both as interpretive and regulative frameworks so that assuming that the other perceives or wants the same thing as you and then acts like you makes perfectly sense (this is embraced by the concept of subjectively based social focal point presented in the chapter 5).

The last chapter provides an example of the integration of the ST and the mindshaping hypothesis in game theory and the results it engender: the explanation of beliefs formation, of how and why an equilibrium occur, and from a theoretical point of view the possibility to get rid of a controversial assumption of game theory: the action-independence of beliefs.

The chapter indeed relies on the premise that a psychological theory of belief formation is required in game theory to offer a Bayesian theory of choices in games. This is not the case in epistemic game theory as beliefs are at the equilibrium so that they describe the choices of the players and the beliefs that they should possibly have held to make this specific choice. However in any case this mathematical construction can explain how a player chooses according to her beliefs which is supposedly the meaning of Bayesian rationality, which theoretically entails that the players intend to maximize their expected utility according to their beliefs. The ST on the contrary explains the formation of players' beliefs and explains how from their perspective and from the apprehension of the game only, players' can form these beliefs by attributing thanks to simulation, perceptions and beliefs to the others. We indeed state that the players form their beliefs by simulating the reasoning of the other players. One result of this assumption, when common belief of rationality is assumed (which is justified by simulation plus mindshaping) does not necessarily rule out non-ratifiable choices and action-dependent beliefs. The action-independence of the players' beliefs means that the players cannot believe that their choice impacts on the others, their beliefs about other's choice are independent of their own action). Ontologically this is hard to believe in strategic interactions, i.e. in game theory which means that

whatever a player choose she believes that this will not have an impact on the other choice, that what she does is not taken into account to a certain extent. A player who simulates the reasoning of the other can thus take her own choice as an evidence of the choice of the other, because she tends to assume that the reasoning processes of the other are similar to her own.

Our approach however leads to cases of multiplicity of subjective belief equilibria (SBE). This is explained by the multiplicity of subjective beliefs to which players can converge during the massaging process. The massaging process symbolize the reasoning process of the players: their strategic reasoning process, which test the consistency of their beliefs. This massaging process explains the way players check the validity of their beliefs with respect to other's reasoning process and beliefs. They check whether they will ultimately reach what Schelling's called the meeting of mind, and thus coordinate. This confirms Schelling's intuition stating that a purely mathematical representation of games cannot provide a rational determinate solution to games; pure deductive reasoning is insufficient to give determinate solutions. We thus see the crucial role of mindshaping which, by providing common understanding and common reasoning in a given context, i.e. homogenization, can allow successful coordination by drastically restricting the set of possible SBE.

The approach of the chapter by the inclusion of both the ST and mindshaping confirms that the intersubjective dimension in game that permits successful coordination cannot discard the social, cultural, and historical background of the players. It shows that a descriptive theory of games necessitates to reflect on how these backgrounds affect the players' strategic reasoning and in particular how deeply they affect the formation of individual beliefs.

Thus the chapter shows that a theory of games understood as a descriptive theory of strategic choices and not only as a normative theory, should rests on an interdisciplinary approach by the inclusion of works in cognitive and social psychology, or philosophy, or history, etc. Mindshaping by crossing these different approaches also highlights how the integration of these socio-cultural factors is necessary in strategic interactions to bring coordination.

Game theory by assuming strategic reasoning and rational choices necessitates theories that explain the homogenization needed for coordination instead of resorting to simple pattern of behavior like mimetism which would also bring coordination. Eductive reasoning when game theory is extending to cases where players' payoffs are not understood as representations of their choices, and to cases were we go beyond the imposition of common belief rationality or common knowledge rationality prior to the game, requires to a certain degree homogenization and a common background. This is one the main messages delivered by the thesis.

Though, we would like to draw the attention on the justification that the ST provides with respect to the imposition of common knowledge. As already mentioned in the thesis common knowledge of rationality is a methodological device imposed to justify the capacity of players to form beliefs on others' beliefs, reasoning and choices. This is the standard intersubjective dimension of strategic reasoning in game contexts. However this assumption is never justified as standardly accepted and is furthermore, in some case inconsistent with other standard assumptions of game theory like the hypothesis of common aligned beliefs prior to the game in epistemic game theory for instance.

As stressed by Janssen (2001, p. 230)

> "it is important to realize that even the weakest solution concepts that are employed in non-cooperative game theory, like iterative elimination of dominated strategies (IEDS) or rationalizability, have individual players impose restrictions on the likely behavior of others (see Bernheim, 1984 and Pearce, 1984). The common knowledge of rationality assumption on which IEDS is based assumes that all players conjecture (or even: know) other players to be as rational as they are. It is clear that this assumption imposes restrictions beyond the simple notion of rational individual behavior. Nevertheless, the assumption is well accepted in non-cooperative game theory, so much so that it is hard to encounter justifications for it. One way to interpret the assumption is that while thinking about what the other may choose, players impose their own kind of rationality on their opponents. According to this interpretation the assumption only imposes restrictions on the thought processes of individual players"

In this respect, while the ST at the same time offers a possible justification of the CKR by the fact that the players tend to attribute to the other, in first instance, their own perceptions, reasoning process, intentions, beliefs and then choices, this inconsistency does not prevail. Besides ST is at the same time consistent with individualism understood in a broad sense as will be argued below and with a non-individualistic mode of reasoning.

One implication of the thesis is in this perspective to show that the intersubjective dimension of strategic reasoning, even from an individualistic basis, thanks to the ST, requires at least a minimal form of social of collective reasoning, so that it appeals to a form of broad individualism. To be accurate enough the attribution of mental states and reasoning process to the others, as demonstrated by mindshaping, requires the experience of sociality; it requires the existence of conventions, or institutions that entail common inductive reasoning. Mindshaping induce the sharing of perceptions, of beliefs, of reasoning processes, of behaviors, etc. Therefore CKR of rationality when justified by psychological and social grounds with the ST and mindshaping resort to a form a collective reasoning and not only a hypothesis imposed by mathematical constraints.

As Binmore (1994, p.142) declares "Game theorists of the strict school believe that their prescriptions for rational play in games can be deduced, in principle, from one-person rationality considerations without the need to invent collective rationality criteria – provided that sufficient information is assumed to be common knowledge." This thesis shows that even common knowledge rationality or more generally common belief of rationality in a sense supposes a form of collective rationality.

Thus the ST, in addition to providing a non-tautological explanation of the elicitation of players' belief, i.e. of prior beliefs in games, could therefore constitute the basis of a more general theory of social interactions, based on a genuinely intersubjective theory of behavior in strategic environments. ST is compatible with many forms of reasoning and can accommodate many form of reasoning. It can be compatible with TR, with self-centered self-interest, with other-regarding concern, with commitment (Sen, 1985, 2007).

And finally mindreading, i.e. the introduction of a theory of mind as the ground for interpersonal understanding, i.e. as the ground for players' epistemic state about each other is related to a real form of interdisciplinarity compared to other approaches like the focal point literature and labeling games, the psychological game theory and the level-k theory. We argue in the thesis, by relying on Schelling and Bacharach in particular, that widening non-cooperative game theory and solving the problem of coordination in game theory requires interdisciplinarity. It cannot be reduced to a problem of equilibrium selection solely based on mathematical constraints, like was the case of the refinement program. Only interdisciplinarity can avoid the *ad-hoc* adds of determinants in player's reasoning that tend to rationalize empirical regularities with respect to coordination but without questioning the analysis of coordination, as a process, as the substance of strategic situations.

We have shown that the ontological thinking with respect to the content of game theory and the concept of games led to the conclusion that seeing game theory as a framework of analysis having for purpose to determine the nature of strategic reasoning and seeing games as a real and active process of interaction and reasoning contribute to provide solutions for the indeterminacy problem that game theory face in particular and which drive many contributions all along the history of game theory. It shows that the type of intersubjectivity involved in standard game theory without conceiving player's beliefs as real mental states is ill defined and leads methodologically to many difficulties such as the logical contradiction between assuming Bayesian rationality and prior beliefs at the equilibrium (as it is the case with the behavioral interpretation of games).

The change of ontology and of methodology involved in the thesis reveal that players' mental states, their intentions, perceptions and beliefs must be integrated in games and in the framework of analysis that is game theory. It thus necessitates an incursion of cognitive sciences in game theory. It necessitates an interdisciplinary approach of game theory.

Our thesis thus offers a new approach of intersubjectivity and a renewal of apprehension of coordination as a process of strategic reasoning explaining how and why a specific solution emerges. It shows that the understanding of intersubjectivity in strategic reasoning and coordination is at the core of a research program in game theory. In that perspective, we see that the integration of framing in game theory can be seen as the beginning of a research program assessing the apprehension of strategic reasoning and the type of intersubjectivity involved in strategic reasoning. Our thesis shows in particular how an ontological thinking on strategic reasoning and intersubjectivity in game theory is at the core of research program ultimately enabling game theory to overcome its main methodological difficulties (the inappropriate definition of strategic rationality, its positive and normative difficulties, and so on).

Within this research program, we identify some works that are in progress. They contribute to extend the results of the thesis and to deepen the understanding of strategic reasoning, to deepen the apprehension of intersubjectivity in game theory and the status of game theoretic modeling with respect to the new approach of games and of game theory we propose in this thesis.

With respect to the intersubjective dimension of strategic reasoning and the interplay between individual and collective and historical determinants in players' capacity to form beliefs about

others' beliefs and choice, three new contributions are planned. A model of games in incomplete information with repeated interactions is in progress with Guilhem Lecouteux. It intends to explain how social focal point can be selected from subjective focal point in interaction. This work intends to deepen the contribution of the chapter 5 in incomplete information, and in dynamic situations, with repeated interactions. One contribution on the permeability of the philosophy of mind during the 20th century and the reflexive literature on economics and on economic methodology with respect to the distinction underlined by Schumpeter (1942) between the observer and the observed for instance is in project. It is an attempt to determine how the philosophy of mind and the interdisciplinary approach it involved offers new ontological thinking on the intersubjective dimension involved in interpersonal interactions and how our own psychological make up affect our apprehension of the other. Again it helps to underline the interplay between individual and collective dimensions in the intersubjective capacity of mental state attribution. In the same perspective the other contribution is a thinking on the roots of intersubjectivity understood as "forms of life" and on the phenomenological side of mindreading with Richard Arena. These two last contributions thus offer more specifically two extensions of chapter 4.

One of the main dimensions of the renewal of understanding on intersubjectivity in games entails a reassessment of the way players are defined in games and in game. If reassessing strategic reasoning, and coordination as we defend in the thesis necessitates incorporating players' mental states and defining players' beliefs no longer as mathematical artifact but as mental states it implies reconsidering the way players' are portrayed. We intend to provide an historical and methodological work on the representation of the players in the history of game theory; the evolutions that this representation undergoes with the recent developments of new game models and new theories of games and the impact that a new ongoing form of interdisciplinary has on it.

And finally as a deepening on the epistemological dimension implied by the new approach of games and of game theory we defend in the thesis three contributions are planned.

A work is in progress with Cléo Chassonery-Zaïgouche on the conception of modeling as an open science in Schelling's work and its roots on the methodology and conception of science inherited from the RAND Corporation. It relies on a thinking on the articulation of the type of modeling proposed in standard game theory, with the faith that mathematical tools as an universal language provided an interdisciplinary ground. We stress on the contrary that at the same time there were at RAND Corporation an appeal for real interdisciplinarity and a form of incompatibility between mathematical modeling understood as a closed system and coherent and self-sufficient system and such interdisciplinarity. We attempt to emphasize that the ontological thinking on strategic reasoning and coordination expected by Schelling necessitates an open science perspective and a specific methodology founded on such open system apprehension of games and of game theory.

A contribution is ongoing with Cyril Hédoin on the foundations of an empirical social ontology with the use of mindreading and simulation theory in epistemic game theory. As ST induces the introduction of real mental states in games, it thus make possible to test epistemic game theory and the mechanism of elicitation of players' beliefs in games. We propose to make some suggestion for building an empirical social ontology of gaming and for testing this mechanism of

elicitation. It again participates of the ontological thinking with respect to intersubjectivity and coordination in games and the improvements that can be made by crossing different methods and theories from social sciences. Lastly, a reflection on the application of epistemic game theory to decision situations in industrial economics and in business sciences with Jamal-Eddine Azzam is in conception. The contribution intends to scrutinize the epistemology of epistemic game theory and its relation with real decision situations in strategic contexts. We propose to investigate the interactions between the theory and its applicability from concrete case studies. This again offers an epistemological thinking on the positive and normative aspects of epistemic game theory and to confront them with the new approach we defend with the inclusion of mindreading.

Appendix

Proof of proposition 1

Since *i* is uncertain about *j*'s belief, and that *i* attributes her own beliefs to others, we have $B_i(E) \cap U_{ij}(E) \Rightarrow B_{ij}(E)$. Since *i* believes E and believes that *j* also believes E, *i* believes that E is mutual belief: $B_i(E) \cap B_{ij}(E) = B_i\big(MB(E)\big)$. *i* however does not know whether *j* believes that E is mutual belief or not. By $SIMB_i$ we have therefore:

$$B_i\big(MB(E)\big) \cap U_{ij}\big(MB(E)\big) \Longrightarrow B_{ij}\big(MB(E)\big) \tag{12}$$

Since *i* believes that E is mutual belief, and that *j* also believes it, *i* believes that it is mutual belief that E is mutual belief: $B_i\big(MB(E)\big) \cap B_{ij}\big(MB(E)\big) = B_i\big(MB^2(E)\big)$. By continuing the iteration, we find that *i* believes that E is common belief.

Proof of proposition 2.

We start by proving $s_{i,j} = s_{i,[k],j}$ for any sequence $[k]$ of players. We consider three players *i*, *j*, and *k* and their respective choice problems:

P_j: *j* must choose a strategy $x_j \in X_j$ so as to maximize her expected payoff

P_k: *k* must form a first order belief $s_{k,j} \in X_j$

P_i: *i* must form a second order belief $s_{i,k,j} \in X_j$

By PM, *j*'s choice function is $f_j(s_{j,-j}) = \underset{x_j \in X_j}{\operatorname{argmax}}\, E\Pi_j(x_j; s_{j,-j}(x_j))$, with:

$$E\Pi_j\left(x_j; s_{j,-j}(x_j)\right) = \sum_{x_i \in X_i} s_{j,-j}\left(x_{-j} \middle| x_j\right) \Pi_j\left(x_{-j}, x_j\right). \tag{13}$$

k's choice function is given by $SIMR_k$: since PM is true for *k*, *k* would also choose the strategy that maximizes *j*'s payoff, if she had to choose at her place. We have therefore:

$$\mathcal{C}_k(\mathrm{P}_j) = f_k\left(s_{k,j,-j}\right), \tag{14}$$

$$\mathcal{C}_k(\mathrm{P}_j) = \underset{x_j \in X_j}{\operatorname{argmax}}\, E\Pi_j\left(x_j; s_{k,j,-j}(x_j)\right). \tag{15}$$

Consider now the case of player *i*, who must form a belief about *k*'s belief about *j*'s strategy. If *i* had to form a belief about *j*'s strategy, she would simulate her reasoning (just as *k* in the previous case):

$$\mathcal{C}_i(\mathrm{P}_j) = \underset{x_j \in X_j}{\operatorname{argmax}}\, E\Pi_j\left(x_j; s_{i,j,-j}(x_j)\right). \tag{16}$$

i therefore assumes that *k* also simulates *j*'s reasoning (she indeed assumes that *k* has the same reasoning process than hers, i.e. attributing her own reasoning process (PM) to simulate the reasoning of another player). We have therefore:

$$\mathcal{C}_i(\mathrm{P}_k) = \underset{x_j \in X_j}{\operatorname{argmax}}\, E\Pi_j\left(x_j; s_{i,k,j,-j}(x_j)\right). \tag{17}$$

Here we can clearly see that $s_{i,j} = s_{i,k,j}$ (condition (i) of proposition 2) if $s_{i,j,-j} = s_{i,k,j,-j}$. In the absence of a prior belief about the beliefs of player *k*, *i* simply attributes her own beliefs to *k* by $SIMB_i$. We thus have $s_{i,j,-j} = s_{i,k,j,-j}$ (*i* assumes that, if she believes $s_{i,j,-j} = x'_{-j}$, then *k* also believes it, i.e. $s_{i,k,j,-j} = x'_{-j}$), and therefore $s_{i,j} = s_{i,k,j}$. We could reproduce the exact same reasoning for higher-order beliefs: player *i* will indeed simulate the reasoning of the succession of players k_1, k_2, etc. which leads her *in fine* to choose the strategy that maximizes the expected payoff of *j* ($SIMB_i$ then allows player *i* to attribute her belief about k_1, and then k_2, etc.). This proves part (i) of proposition 2.

The proof of part (ii) is similar to the proof of part (i): *i* indeed considers the beliefs of other players *k* about her own choice, and simulating the maximization of $E\Pi_i$ in player *k*'s mind, or simulating player *j* simulating the maximization of $E\Pi_i$ in player *k*'s mind leads to the same outcome if they all share the same higher-order beliefs, which is ensured by $SIMB_i$.

We now turn to the proof of part (iii). For notational convenience, we give the proof for two players: the generalization to *n* players is conceptually similar but less tractable (we must indeed determine the beliefs of player 1 about player 2's belief about player 3's belief … about player *i*'s strategy, which makes the notations quite heavy although the resolution is based on a fixed point argument which is conceptually similar for *2* and *n>2* players). PM means that player *i* chooses the strategy that maximizes her expected payoff given her beliefs about *j*'s strategy (that may depend on *i*'s strategy). By simulation, *i* assumes that PM is true for *j* (it indeed corresponds to her choice function): *i* therefore believes that *j* chooses her strategy so as to maximize her expected payoff, given *j*'s beliefs about *i*'s strategy (that may also depend on j's strategy). If player *i* plays a strategy $\hat{x}_i$, she believes that *j* plays:

$$s_{i,j}(\hat{x}_i) = \underset{x_j \in X_j}{argmax}\left[\sum_{x_i \in X_i} s_{i,j,i}(x_i|x_j)\, \Pi_j(x_i, x_j)\right] \tag{18}$$

with $s_{i,j,i}: X_j \longmapsto \Delta(X_i)$ *i*'s belief about *j*'s belief about *i*'s strategy. *i* must therefore anticipate how *j* forms her belief about her actions. Since she attributes her own reasoning process to *j*, she assumes that *j* forms her belief by simulating *i*'s reasoning process. Since *i* may believe that her choice could be correlated with *j*'s choice, she believes that *j* also believes that their choices could be correlated (by $SIMB_i$). If player *j* plays a strategy $\hat{x}_j$, then *i* believes that *j* believes that *i* plays:

$$s_{i,j,i}(\hat{x}_j) = \underset{x_i \in X_i}{argmax} \left[\sum_{x_j \in X_j} s_{i,j,i,j}(x_j | x_i)\, \Pi_i(x_i, x_j) \right] \tag{19}$$

with $s_{i,j,i,j}: X_i \longmapsto \Delta(X_j)$ *i*'s belief about *j*'s belief about *i*'s belief about *j*'s strategy. By simulation, this function is also defined as the best reply to a higher order belief, and so on.

Note that we have only investigated *i*'s belief about *j*'s action for a given strategy $\hat{x}_i$: the same operation should then be done for each possible strategy in X_i (and at each step, for all the strategies $\hat{x}_j \in X_j$ of player j). A more concise notation of the problem would be the following:

$$s_{i,j}(x_i) = \underset{f_j \in F_j}{argmax} \left[\sum_{x_i \in X_i} s_{i,j,i}\left(f_j(x_i)\right)\, \Pi_j(x_i, x_j) \right], \tag{20}$$

$$s_{i,j,i}(x_j) = \underset{f_i \in F_i}{argmax} \left[\sum_{x_j \in X_j} s_{i,j,i,j}\left(f_i(x_j)\right)\, \Pi_i(x_i, x_j) \right] \tag{21}$$

with $F_j = \{f_j: X_i \longmapsto \Delta(X_j)\}$ the set of functions associating a probability distribution $p_j \in \Delta(X_j)$ for a strategy $\hat{x}_i \in X_i$. We however know that $s_{i,j,i,j} = s_{i,j}$ at the MBH: this means that $s_{i,j}$ and $s_{i,j,i}$ must be a mutual best reply. When massaging her beliefs, *i* is therefore looking for a vector of conditional probability distributions $\{f_j^*, f_i^*\}$ that simultaneously maximizes the expected payoff of both players, i.e. such that:

$$\sum_{x \in X} s^*(x)\Pi_i(x) \geq \sum_{x \in X} s'^{(x)\Pi_i(x)}, \qquad \forall s' \in \Omega\left(f_j', f_i^*\right), \forall f_j' \in \Phi_i, \forall i \in N, \tag{22}$$

With $s^* \in \Omega\left(f_j^*, f_i^*\right)$. The probability distribution s^* representing the vector of conditional probability distributions $\{f_j^*, f_i^*\}$ verifies condition (iii).

Proof of proposition 3.

Consider the unconditional beliefs $s_{i,j} = p_j^*$ and $s_{i,j,i} = p_i^*$ $\forall i, j \in N$ (and suppose that the higher-order beliefs of *i* correspond to her 1st and 2nd order beliefs). Player *i* believes that all the other players unconditionally play their Nash (mixed) strategy, and believes that all the players *j* believe that she unconditionally plays her Nash strategy. Even if player *j* adopts another conditional distribution, she knows that the strategy of the others will remain the same: the highest payoff player *j* can get is therefore the one induced by her best reply to p_{-j}^*, i.e. her Nash strategy p_j^*. The belief hierarchy S_i^* such that all the players unconditionally play their Nash strategy is therefore a massaged belief hierarchy.

Proof of proposition 4.

The proposition is a corollary of proposition 3: if $s^*_{i,j} = x^*_j$ and $s^*_{i,j,i} = x^*_i,\ \forall i,j \in N$, with $x^* \in X$ the Nash equilibrium of G, then the belief hierarchy generated by s^* is a MBH. The optimal strategy for each player (the strategy that maximizes their expected payoff) is their Nash strategy by construction, and – since the equilibrium is in pure strategies – they accurately predict the choice of the other players.

Proof of proposition 5.

Let $x^* \in X$ denote a Nash equilibrium of *G* and $\bar{x} \in X$ a strategy profile that is Pareto superior to p^*, i.e. a profile such that:

$$\Pi_i(\bar{x}) \geq \Pi_i(x^*),\ \forall i \in N,$$

with at least one strict inequality. Consider the following beliefs:

$$\begin{cases} s^*_{i,j}(x_i \neq \bar{x}_i) = x^*_j,\ \ s^*_{i,j,i}\left(x_j \neq \bar{x}_j\right) = x^*_i, & \forall i,j \in N, i \neq j \quad (23) \\ s^*_{i,j}(x_i = \bar{x}_i) = \bar{x}_j,\ \ s^*_{i,j,i}\left(x_j = \bar{x}_j\right) = \bar{x}_i, & \forall i,j \in N, i \neq j \quad (24) \end{cases}$$

Condition (23) means that, when the players choose a strategy different from $\bar{x}_i$, they believe that all the other players unconditionally play their Nash strategy. However, condition (24) means that, when they play $\bar{x}_i$, they all believe that all the others play their part of the profile $\bar{x}$. Player *i* should therefore play $\bar{x}_i$ to maximize her expected payoff (condition (i)): if *i* chooses another strategy, she indeed believes that all the other plays their Nash strategy, and the highest payoff for *i* in this case is her Nash payoff. Since they all reach a higher payoff at $\bar{x}$, they should all play their part of the profile. Furthermore, by construction of their beliefs, they predict well the choice of the other players (condition (ii)). Lastly, the belief hierarchy generated by s^* is a MBH: changing one's conditional probability can either lead *i* to play $\bar{x}_i$ (in which case the payoff is the same) or a different strategy $x_i \neq \bar{x}_i$, which induces all the other players to play their Nash strategy (in which case player *i*'s payoff is bounded by her Nash payoff, which is lower than the initial payoff). $\bar{x}$ is well a subjective belief equilibrium.

References

Adolphs, R. and Tranel, D. (2000). Emotion Recognition and the human amydala. In The amydala: a functional analysis. Aggleton J. P. (ed.), Oxford: Oxford University Press, 587-630.

Adolphs, R., Tranel, D., Damasio, H., and Damasio A. R. (1995). Fear and the human amygdala. *The Journal of Neurosciences*, 15(9), 5879–5892.

Alexander, J. McK. (2009). Evolutionary game theory. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.), Cambridge: Cambridge University Press.

Amadae, S. (2003). Rationalizing capitalist democracy: the cold war origins of rational choice liberalism. Chicago: University of Chicago Press.

Anderlini, L. (1990). Communication, computability and common interest games. *Games and Economic Behavior*, 27(1), 1-37.

Andreoni, J. (1989). Giving with impure altruism: applications to charity and ricardian equivalence. *Journal of Political Economy*, 97(6), 1447-1458.

Andreoni, J. (1990). Impure altruism and donations in public good: a theory of warm-glow giving. *The Economic Journal*, 100(401), 464-477.

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion?. *The American Economic Review*, 85(4), 891-904.

Andreoni, J., Miller, J. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. *The Economic Journal*, 103 (418), 570-585.

Andreoni, J., Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2), 737-753.

Anscombe, F. J., Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics*, *34*(1), 199-205.

Aoki, M. (2001). *Toward a comparative social analysis*. Cambridge, MA: The MIT Press.

Apperly, I. (2011). *Mindreaders: The cognitive basis of "theory of mind"*. Psychology Press.

Arena, R., Larrouy, L. (2016). Subjectivity and coordination in economic analysis. *Oeconomia*, 6(2), 201-233.

Armbruster, W., Böge W. (1979). Bayesian Game Theory. In *Game Theory and Related Topics*, Moeschlin O. and D. Pallaschke (eds.), Amsterdam: North-Holland.

Armstrong, D. M. (1968). *A Materialist theory of the mind*. New York: Humanities Press.

Arnold, B. C., Press, J. S. (1989). Bayesian estimation and prediction for Pareto data. *Journal of the American Statistical Association*, 84, 1079–1084.

Astington, J., Gopnik, A. (1988). Knowing you've changed your mind: Children's understanding of representational change. In J. Astington, P. Harris, and D. Olson, eds., Developing Theories of Mind. Cambridge: Cambridge University Press.

Augenstein, B. W. (1993). A brief history of RAND's mathematics department and some of its accomplishments. RAND, DRU-218-RC. Santa Monica, CA: RAND Corporation.

Aumann, R. J. (1976). Agreeing to Disagree. *Annals of Statistics*, 4, 1236-1239.

Aumann, R. J. (1981). Survey of repeated games. In *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern*, Aumann R. J. et *al* (eds.), Mannheim, Zurich: Bibliographisches Institut, 11-42.

Aumann, R. J. (1985). An axiomatization of the non-transferable utility. *Econometrica*, 53(3), 599-612.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1), 1-18.

Aumann, R. J. (1998). On the Centipede game. *Games and Economic Behavior*, 23, 97-105.

Aumann, R. J. (2000). Collected papers. Volume I. Cambridge, MA: The MIT Press.

Aumann R. J., (2010). In *Epistemic Logic: Five Questions*. Hendricks, V., and O. Roy (eds.), Copenhagen: Automatic Press/VIP, 21–33.

Aumann, R. J., Sorin, S. (1989). Cooperation and bounded recall. *Games and Economic Behavior*, 1, 5-39.

Aumann, R. J., Hart, S. (1992). *Handbook of Game Theory with Economic Applications*. Amsterdam, North-Holland: Elsevier Science Publishers.

Aumann, R., Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 63, 1161–1180.

Aumann, R. J., Dreze, J. H. (2008). Rational expectations in games. *American Economic Review*, 98, 72–86.

Aumann, R. J., Dreze, J. H. (2009). Assessing strategic risk. *American Economic Journal: Microeconomics*, 1(1), 1-16.

Axelrod, R., Hamilton, W. D. (1981). The evolution of cooperation. *Science, New Series*, 211(4489), 1390-1396.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Aydinonat, N. E. (2005). An interview with Thomas C. Schelling: interpretation of game theory and the checkboard model. *Economics Bulletin*, 2(2), 1-17.

Aydinonat, N. E. (2007). Models, conjectures and exploration: an analysis of Schelling's checkboard model of residential segregation. *Journal of Economic Methodology*, 14(4), 429-454.

Ayson, R. (2004). *Thomas Schelling and the nuclear age: strategy as a social science*. London: Frank Cass.

Bach, C.W., Tsakas, E. (2012). Pairwise interactive knowledge and Nash equilibrium. Working paper, Maastricht University.

Bacharach, M. (1976). *Economics and the Theory of Games*. London, UK: The Macmillan Press LTD.

Bacharach, M. (1985). Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory*, 37, 167-90.

Bacharach, M. (1986). The problem of agents' beliefs in economic theory. In *Foundations of Economics*, Baranzini, M. and R. Scazzieri (eds.), Oxford: Blackwell, 175-203.

Bacharach, M. (1992). Mutual Knowledge and Human Reason. In *Knowledge, Belief and Strategic Interaction*, Bicchieri, C. and M. L. Dalla Chiara (eds.), New York: Cambridge University Press, 285-316.

Bacharach, M. (1987). A theory of rational decision in games, *Erkenntnis*, 27, 17-55.

Bacharach, M. (1989). Expecting and affecting. *Oxford Economic Papers*, 41(2), 339-355.

Bacharach M. (1990). Commodities, Language, and Desire. *The Journal of Philosophy*, 87(7), 346-368.

Bacharach, M. (1991). Games with concept sensitive strategy spaces, Working Paper, *Institute of Economics and Statistics*, University of Oxford.

Bacharach, M. (1992). Backward induction and beliefs about oneself. *Synthese*, 91, 247-284.

Bacharach, M. (1993). Variable Universe Game. In *Frontiers of Game Theory*, Binmore K., Kirman A., and P. Tami (eds.), Cambridge, MA: The MIT Press.

Bacharach, M. (1994). The epistemic structure of a theory of game. *Theory and Decision*, 37, 7-48

Bacharach, M. (1995). Cooperating without communicating", *ESRC* Economics, Beliefs and Behaviour Programme, Discussion Paper No.1

Bacharach, M. (1997a). Common Knowledge. In *The New Palgrave Dictionary of Law and Economics*, London: Macmillan.

Bacharach, M. (1997b). We equilibria: a variable frame theory of cooperation. *Institute of Economics and Statistics*, University of Oxford.

Bacharach, M. (1998). Agents with multiple identities in the theory of cooperation. *Institute of Economics and Statistics*, University of Oxford.

Bacharach, M. (1999). Interactive team reasoning: a contribution to the theory of cooperation. *Research in Economics*, 53(2), 117-147.

Bacharach, M. (2001a). Framing and cognition in economics: the bad news and the goods. Lecture notes, *ISER* Workshop XIV, july 2001: Cognitive Processes in Economics.

Bacharach, M. (2001b). Superagency: beyond individualistic game theory. Invited Lecture presented at *TARK VIII*: Eight Conference on Theoretical Aspects of Rationality and Knowledge, Certosa di Pontignano, University of Siena.

Bacharach, M. (2006). *Beyond Individual Choice: Team and Frame in Game Theory*. Gold, N. and R. Sugden (eds.), Princeton, NJ: Princeton University Press.

Bacharach, M., Hurley, S. (1991). Issues and advances in the foundations of decision theory. In *Foundations of Decision Theory*, Oxford, UK: Blackwell publishers.

Bacharach M., Mongin, P. (1994). Epistemic logic and the foundations of game theory. *Theory and Decision*, 36, 1-6.

Bacharach, M., Bernasconi, M. (1997). The variable frame theory of focal points: an experimental study, *Games and economic Behavior*, 19(1), 1-45.

Bacharach, M., Stahl, O. (2000). Variable-Frame Level-*n* Theory. *Games and Economic Behavior*, 32 (1), 220-246.

Backhouse, R., Cherrier, B. (2016). 'It's computerization, Stupid!' The spread of computers and the changing role of theoretical and applied economics. Available from SSRN: https://ssrn.com/abstractD2781253, https://doi.org/10.2139/ssrn.2781253

Banks, J., Sobel, J. (1987). Equilibrium Selection in Signaling Games. *Econometrica*, 55, 647-661.

Bargh, J. A., Chartrand, T. (1999). The unbearable automaticity of being. *American Psychologist*, 54(7), 462–479.

Baron-Cohen, S. (1989). Perceptual role-taking and protodeclarative pointing in autism. *British Journal of Developmental Psychology*, 7, 113–127.

Baron-Cohen, S. (1995). Mindblindness: An Essay on Autism and Theory of Mind. Cambridge, MA: MIT Press.

Baron-Cohen, S. (1999). The extreme male brain theory of autism. In *Neurodevelopmental Disorders*, Tager-Flusberg, H. (ed.), Cambridge, MA: MIT Press.

Baron-Cohen, S., Leslie, A., Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21, 37–46.

Baron-Cohen, S., Leslie, A., Frith, U. (1986). Mechanical, behavioral, and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, 4, 113–125.

Bardsley, N., Mehta, J., Starmer, C., Sugden, R. (2010). Explaining focal points: Cognitive hierarchy theory versus team reasoning. *Economic Journal*, 120, 40–79.

Bartsch, K., Wellman, H. (1995). *Children talk about the mind*. New York: Oxford University Press.

Battalio, R., Samuelson, L., Van Huyck, J. (2001). Optimization incentives and coordination failure in laboratory stag hunt games. *Econometrica*, 69, 749–764.

Battigali, P. (1988). Implementable strategie, prior information and the problem of credibility in extensive games. *International Review of Economics and Business*, 35, 705-733.

Battigali, P. Bonnano G. (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53, 149-225.

Battigalli, P., Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144, 1-35.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22, 725–730. https://doi.org/10.1121/1.1906679

Becker, G. S. ([1957] 1971). *The economics of discrimination*. Chicago, IL: University of Chicago Press.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, 1, 557–560.

Bergstrom, T. C. (2002). Evolution of social behavior: individual and group selection. *The Journal of Economic Perspectives*, 16(2), 67-88.

Bergstrom, T. C., Stark, O. (1993). How altruism can prevail in an evolutionary environment. *The American Economic Review*, 83(2), 149-155.

Bernheim, D. (1984). Rationalizable strategic behavior. *Econometrica*, 52, 1007–1028.

Bernheim, D. (1986). Axiomatic characterizations of rational choice in strategic environments. *The Scandinavian Journal of Economics*, 88(3), 473-488.

Bermúdez, J. L. (2005). Philosophy of psychology: a contemporary introduction. New York: Routledge.

Bermúdez, (2006a). Cognitive science: an introduction to the science of the mind (2nd ed.). New York: Cambridge University Press.

Bermúdez, 2006b. Self-consciousness. Encyclopedia of Cognitive Sciences, John Wiley & Sons

Bhatt, M. A., Camerer, C. F. (2005). Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior*, 52(2), 424-459

Bicchieri, C. (1993). *Rationality and coordination*. New York, NY: CUP Archive.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Binmore, K. G. (1987). Modeling rational players: Part I. Economics and Philosophy, 3, 179-214.

Binmore, K. G. (1988). Modeling rational players: Part II. *Economics & Philosophy*, 4(1), 9-55.

Binmore, K. G. (1990). *Essays on the foundations of game theory*. Cambridge, MA: Basil Blackwell.

Binmore, K. G. (1992). *Fun and games: A text on game theory*. Lexington, MA: Heath.

Binmore, K. G. (1993). De-Bayesing Game Theory. In *Frontiers of game theory*, Binmore, K. G., Kirman, A. P., and P. Tani (eds.). Cambridge, MA: The MIT Press.

Binmore, K. G. (1994). *Game theory and the social contract: Playing Fair Vol I*. Cambridge, MA: The MIT Press.

Binmore, K. G. (1998). *Game theory and the social contract: Just playing Vol II*. Cambridge, MA: The MIT Press.

Binmore, K. G. (2006). Rational decisions in large worlds. Lecture given at 2006 ADRES Conference, Marseille, France. Retrieved from http://discovery.ucl.ac.uk/14433/1/14433.pdf

Binmore, K. G. (2007).*Playing for Real*. New York: Oxford University Press.

Binmore, K. G. (2009). *Rational decisions*. Princeton, NJ: Princeton University Press.

Binmore, K. G., Brandenburger, A. (1988). Common knowledge and game theory. London School of Economics, MIMEO.

Binmore, K. G., Osborne, M. J., Rubinstein, A. (1992). Noncooperative models of bargaining. In *Handbook of Game Theory with Economic Applications (Vol. 1)*, Aumann, R., and S. Hart (eds.), North-Holland: Elsevier, 179-225.

Binmore, K. G., Shaked, A. (2010). Experimental Economics: Where next? *Journal of Economic Behavior & Organization*, On the Methodology of Experimental Economics, 73(1), 87-100.

Birch, S. A. J., Bloom, P. (2003). Children are cursed: An asymmetric bias in mental-state attribution. *Psychological Science*, 14, 283–286.

Birch, S. A. J., Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, 8, 255–260.

Blackledge, J. T. (2003). An introduction to relational frame theory: basics and applications. *The Behavior Analyst Today*, 3(4), 421-433.

Blake R., Mouton J., 1986. From theory to practice in inter-face problem solving. In *Psychology of Intergroup Relations*. Worchel, S., and Austin W. (eds.), Chicago: Nelson Hall.

Bloom, P. (2000). How Children Learn the Meaning of Words. Cambridge, MA: MIT Press.

Blume, L., Brandenburger, A., and Dekel, E. 1991. Lexicographic Probabilities and Choice Under Uncertainty. *Econometrica*, 59, 61-79.

Blume, L., Brandenburger, A., and Dekel, E. 1991. Lexicographic Probabilities and Equilibrium Refinements. *Econometrica*, 59, 81-98.

Board, O. (2006). The equivalence of Bayes and causal rationality in games. *Theory and Decision*, 61(1), 1–19.

Böge, W., Eisele, T. (1979). On solutions of Bayesian games. *International Journal of Game Theory*, 8, 193–215.

Bolton, G., Ockenfels A. (2000). ERC: A theory of equity, reciprocity and competition. *American Economic Review,* 90(1), 166-193.

Bonanno G., Nehring K. (1999). How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory*, 28, 409-43.

Boudon, R. (2004). Théorie du choix rationnel ou individualisme méthodologique?. *Revue du MAUSS*, 24, 281–309.

Boulding, K. (1957). A New Look at Institutionalism. *The American Economic Review*, 47(2), 1-12.

Bardsley, N., Mehta, J., Starmer C., Sugden R. (2010). Explaining focal points: cognitive hierarchy theory *versus* team reasoning. *The Economic Journal*, 120, 40-79.

Bardsley, N., Ule, A. (2017). Focal points revisited: Team reasoning, the principle of insufficient reason and cognitive hierarchy theory. *Journal of Economic Behavior & Organization*, 133, 74-86.

Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In *Economic analysis of markets and games*, Dasgupta, P. Gale, D. Hart, O., and E. Maskin (eds), Cambridge: MIT Press, 282–290.

Brandenburger, A. (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35, 465–492.

Brandenburger, A. (2010). Origins of epistemic game theory. In *Epistemic logic: Five questions*, Hendricks, V., and O. Roy (eds.), 59–69.

Brandenburger, A., Dekel, E. (1987). Rationalizability and correlated equilibria. *Econometrica*, 55, 1391–1402.

Brandenburger, A. Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, 59, 189-198.

Bratman, M. E. (1993). Shared intention. Ethics, 104(1), 97-113.

Breslau, D. (1997). The political power of research methods: knowledge regimes in U.S. labor-market policy. *Theory and Society*, 26(6), 869–902.

Brewer, M., Gardner, W. (1996). Who is this "we"? levels of collective identity and self representations. *Journal of Personality and Social Psychology*, 71(1), 83-93.

Brewer, M., Kramer, R., (1986). Choice behavior in social dilemmas: effects of social identity, group size, and decision framing, *Journal of Personality and Social Psychology*, 50(3), 543-549.

Brewer M., Miller N. (1996) *Intergroup Relationships*. Buckingham, UK: Open University Press.

Bruch, E. E., Mare, R. D. (2003). Neighborhood choice and neighborhood change. *American Journal of Sociology*, 112(3), 667-709.

Brueckner, A. (1998). Deductive closure principle. *Routledge Encyclopedia of Philosophy*, Taylor and Francis. https://www.rep.routledge.com/articles/thematic/deductive-closure-principle/v-1., doi:10.4324/9780415249126-P011-1.

Brueckner, A. (1998). Closure and context. *Ratio*, 11(1), 78-82.

Bruner, J. (1957). On perceptual Readiness. *Psychological Review*, 64(2), 123-152.

Bruni, L., Sugden, R. (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. *The Economic Journal,* 117(516), 146-173.

Buccino, G. et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience*, 13(2), 400–404.

Butler, D. J. (2012). A choice for 'me' or for 'us'? Using we-reasoning to predict cooperation and coordination in games. *Theory and Decision*, 73(1), 53-76.

Calder, A. J., Keane, J., Manes, F., Antoun, N., Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience*, 3, 1077–1078.

Camerer, C., Loewenstein, G., Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, 97, 1232–1254.

Camerer, C., Issacharoff, S., Loewenstein, G., O'donoghue, T., Rabin, M. (2003). Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism". *University of Pennsylvania law review*, 151(3), 1211-1254.

Campbell, D. (1958). Common fate, similarity an other indices of the status of aggregates of persons as social entities, *Behavioral Science*, 3, 14-25.

Carruthers, P. (1996). Autism as mind-blindness: An elaboration and partial defence. In *Theories of Theories of Mind*, Carruthers, P. and P. K. Smith (eds.), Cambridge: Cambridge University Press, 257–273.

Casajus, A. (2000). Focal points in framed strategic forms. *Games and Economic Behavior*, 32(2), 263-291.

Casati, R. (2003). Representational advantages. *Proceedings of the Aristotelian Society*, 2003, 281–298.

Charness, G., Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, 74(6), 1579-1601.

Charness, G., Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817-869.

Charness, G. Rabin, M. (2004). Expressed preferences and behavior in experimental games. *Working Paper, Department of Economics*, UC Berkeley.

Chartrand, T. L., Bargh, J. A. (1999). The chameleon effect: The perception behavior link and social interaction. *Journal of Personality and Social Psychology*, 76, 893–910.

Chassonnery-Zaïgouche, C., Larrouy, L. (2017). "From warfare to welfare": Contextualising Arrow and Schelling's models of racial inequalities (1968–1972). *The European Journal of the History of Economic Thought*, 24(6), 1355-1387.

Chemero, A. (2009). Radical embodied cognitive science. Cambridge, MA: MIT Press.

Cherniak, C. (1986). *Minimal Rationality*. Cambridge, MA: MIT Press.

Chick, V. (2004). On open systems. *Brazilian Journal of Political Economy*, 24(1), 3-16

Chick, V., Dow, S. (2005). The meaning of open systems. *Journal of Economic Methodology*, 12(3), 363–381

Cho, I., and Kreps, D. (1987). Signaling Games and Stable Equilibria. *Quarterly Journal of Economics*, 102, 179-221.

Churland, P. S. (1991). Our brains our selves: reflection on neuroethical questions. In *Bioscience and society*, Roy, D. Wynne, B. W. Old R. W. (eds.), West Suseex: Wiley & Sons, 77-96.

Clark, A. (1997). Being there: Putting brain, body and world together again. Cambridge, MA: MIT Press.

Clark, A. (2008). *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford, New York: Oxford university Press.

Clark, H. H., Marshall, C. R. (1981). Definite reference and mutual knowledge. In *Elements of Discourse Understanding*, Joshi, A. K., Webber, B. L., and I. A. Sag (eds.), Cambridge: Cambridge University Press.

Cohen, S. (2004). The lasting legacy of an american dilemma. New York: Carnegie Corporation. Available from: https://www.carnegie.org/media/filer_public/98/65/9865c794-39d9-4659-862e-aae1583278a8/ccny_cresults_2004_americandilemma.pdf [Accessed 3 Aug 2016].

Colman, A. M. (1997). Salience and Focusing in Pure Coordination Games. *Journal of Economic Methodology*, 4, 61-81.

Colman, A. M. (2003). Cooperation, Psychological Game Theory and Limitations of Rationality in Social Interaction. *Behavioral and Brain Sciences*, 26, 139-198.

Colman, A. M. (2004). Reasoning about strategic interaction: Solution concepts in game theory. In *Psychology of reasoning: Theoretical and historical perspectives*, Manktelow, K., and M. C. Chung (eds.), London: Psychology Press, 287-308.

Colman, A. M. (2006). Thomas C. Schelling's psychological decision theory: introduction to a special issue. *Journal of Economic Psychology*, 27, 603-608.

Colman, A. M., Bacharach, M. (1997). Payoff dominance and the Stackelberg heuristic. *Theory and Decision*, 43 (1), 1-19.

Colman, A. M., Pulford, B. D., Rose, J. (2008). Collective rationality in interactive decisions: evidence for team reasoning. *Acta Psychologica*, 128(2), 387-397.

Colman, A. M., Pulford, B. D., Lawrence, C. L. (2014). Explaining strategic coordination: cognitive hierarchy theory, strong stackelberg reasoning, and team reasoning. *Decision*, 1(1), 35–58.

Cookson, R. (2000). Framing effects in public goods games. *Experimental Economics*, 3, 55-79.

Cooper, R. W., DeJong, D. V., Forsythe, R., Ross, T. W. (1990). Selection Criteria in Coordination Games: Some Experimental Results. *American Economic Review*, 80(1), 218-233.

Crawford, V. P. (1991). Thomas Schelling and the analysis of strategic behavior. In *Strategy and choice*, Zeckauser R. J. (ed.), Cambridge, MA: MIT Press, 265-269

Crawford, V. P., Haller, H. (1990). Learning how to cooperate: optimal play in repeated coordination games. *Econometrica*, 58, 571-595.

Cruz, J., Gordon, R. (2005). Simulation theory. in (2005), *Encyclopedia of Cognitive Science*, Nadel L. (ed), Hoboken: John Wiley and Sons.

Cubitt, R. P., Sugden, R. (2003). Common knowledge, salience and convention: a reconstruction of david lewis' game theory. *Economics and Philosophy*, 19(2), 175-210.

Cussins, A. (1988). Dennett's realisation theory of the relation between folk and scientific psychology. *Behavioral and Brain Science*, 11(3), 495-546.

Damasio, A. (1994). *Descartes' Error*. New York: Grosset/Putnam Press.

Damasio, A. R., Tranel, D., Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience*, 13, 89–109.

Dantzig, G.B. (1963). Linear programming and extensions. RAND/R-366-PR. Santa Monica, CA: RAND Corporation.

Dantzig, G.B. (2002). Linear programming. *Operations research*, 50 (1), 42–7.

Davidson, D. (1984). Inquiries into truth and interpretation. Oxford: Oxford University Press.

Davis, J. B. (2003), *The Theory of the Individual in Economics: Identity and Value*, London: Routledge.

Davis, J. B. (2004). Collective intentionality, complex economic behavior, and valuation. In *The Elgar Companion to Economics and Philosophy*, Davis, J. B., Marciano A., Runde J., (eds.) Edward Elgar Publishing, 386-402.

Davis J. B., (2011), *Individuals and Identity in Economic*s, New York: Cambridge University Press.

Dawes, R., van de Kragt, A., Orbell J. (1988). Not me or they but we: the importance of group identity in eliciting cooperation in dilemma situations: a transformation of motives. *Acta Psychologica*, 68(1-3), 83-97.

De Bruin, L. (2009). Overmathematisation in game theory: pitting the Nash Equilibrium Refinement Programme against the Epistemic Programme. *Studies in History and Philosophy of Science*, 40(3), 290-300.

Decety, J., and Lamm, C. (2006). Human empathy through the lens of social neuroscience. *The Scientific World Journal*, 6, 1146-1163.

De Cremer, D., Van Vugt, M. (1999). Social identification effects in social dilemmas: a transformation of motives. *European Journal of Social Psychology*, 29, 497-713.

De Jongh, D., Liu, F. (2009). Preference, priorities and belief. In *Preference Change*, Grüne-Yanoff T., and S. O. Hansson (eds.), Dordrecht: Springer, 85-107.

Dekel, E., Gul, F. (1997). Rationality and knowledge in game theory. In *Advances in economics and econometrics*, Kreps, D., and K. Wallis (eds.), Cambridge, UK: Cambridge University Press.

Dekel, E. and Siniscalchi (2015). Epistemic game theory. *Handbook of Game Theory with Economic Applications Vol. 4*, Young, H. P., and S. Zamir, Amsterdam, North-Holland: Elsevier Science Publishers, 619-702.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, 68, 87–106.

Dennett, D. (1978a). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1, 568–570.

Dennett, D. C. (1978b). Intentional systems. In *Brainstorms*, Dennett, D. (ed.), Montgomery, VT: Bradford, 3–22.

Dennett, D. C. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

Dennett, D. C. (1988). Out of the armchair and into the field. *Poetics Today*, 9, 205–221.

Dennett, D. C. (2008). Fun and Games in Fantasyland. *Mind & Language*, 23(1), 25-31.

Dennett, D. C. (2009). Intentional systems theory. In *The Oxford handbook of philosophy of mind*, Beckermann, A., McLaughlin B. P., and S. Walte (eds.), Oxford, New York: Oxford University Press. DOI: 10.1093/oxfordhb/9780199262618.003.0020

Devaine, M., Hollard, G., Daunizeau, J. (2014.) The social bayesian brain: does mentalizing make a difference when we learn? *PLoS Computational Biology*, 10(12). e1003992. https://doi.org/10.1371/journal.pcbi.1003992

De Waal, F. B. M. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology*, 59, 279-300.

Dietrich, F., List, C. (2013). Where do preferences come from?.*International Journal of Game Theory*, 42(3), 613-637.

Dietrich, F., List, C. (2016). Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics & Philosophy*, *32*(2), 249-281.

Dijksterhuis, A., van Knippenberg, A. (2000). Behavioral indecision: Effects of self focus on automatic behavior. *Social Cognition*, 18, 55–74.

Dilthey, W. (1977). *Descriptive psychology and historical understanding*. The Hague: Martinus Nijhoff.

Dion, K. (1973). Cohesiveness as a determinant of ingroup-outgroup bias. *Journal of Personality and Social Psychology*, 28(2), 163-171.

Dixit, A. (2006). Thomas Schelling's Contributions to Game Theory. *The Scandinavian Journal of Economics*, 108(2), 213-229.

Don Vitto, P. A. (1969). The essentials of a planning-programming-budgeting system. RAND/P4124. Santa Monica, CA: RAND Corporation.

Dufwenberg, M., Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47, 268-298.

Dunn, S. P. (2001). Bounded rationality is not fundamental uncertainty: a post Keynesian perspective. *Journal of Post Keynesian Economics*, 23(4), 567-587.

Düppe, T., Weintraub, E. R. (2014). *Finding equilibrium: Arrow, Debreu, McKenzie and the problem of scientific credit*. Princeton, NJ: Princeton University Press.

Epley, N., Keysar, B., Van Boven, L., Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, 87, 329–339.

Epley, N., Morewedge, C. K., Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, 40, 760–768.

Epstein, J.M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.

Epstein, J. M., Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Washington, DC: Brookings Institution Press.

Erickson, P., et al., (2013). *How reason almost lost its mind. The strange career of cold war rationality*. Chicago: University of Chicago Press.

Faillo, M., Smerilli, A., Sugden R. (2017). Bounded best-response and collective-optimality reasoning in coordination games. *Journal of Economic Behavior & Organization*, 140, 317-335.

Falk, A., Fehr, E., Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, 41(1), 20-26.

Falk, A., Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54, 293-315

Falk, A., Fehr, E., Fischbacher, U. (2008). Testing theories of fairness - intentions matter. *Games and Economic Behavior*, 62, 287-303.

Farrell, J., (1988). Communication, coordination and Nash equilibrium. *Economic Letters*, 27, 209-214.

Fehr, E., Schmidt, K. M. (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, 114(3), 817-868.

Fehr E., Schmidt, K. M. (2006). The economic of fairness, reciprocity and altruism: experimental evidence and new theories. In Handbook of the Economics of giving, altruism and reciprocity (Vol. 1), Kolm, S.C., Ythier, J. M. (eds.), North-Holland: Elsevier, 615-691.

Ferber, J. (2007). Multi-agent concepts and methodologies. In *Agent based modeling and simulations in the human and social sciences*, A. Phan (ed.), Oxford: The Bardwell Press.

Fiske, S. T., Taylor, S. E. (1991). *Social cognition*. McGraw-Hill series in social psychology (2nd ed.), Mcgraw-Hill Book Company.

Fletcher, P. C. et al. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, 57(2), 109-128.

Fleury, J.-B. (2012). Wandering through the borderlands of the social sciences: Gary Becker's economics of discrimination. *History of political economy*, 44 (1), 1–40.

Flood, M. M. (1952). On game-learning theory and some decision-making experiments. RAND/ AD0604157, Santa Monica, CA: RAND Corporation.

Fodor, J. A. (1975). *The Language of Thought*. Scranton, PA: Crowell.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fontaine, P. (1997). Identification and economic behavior sympathy and empathy in historical perspective. *Economics and Philosophy*, 13(2), 261-280.

Fontaine, P. (2000). Making use of the past: theorists and historians on the economics of altruism. *The European Journal of the History of Economic Thought*, 7(3): 407-422

Forget, E. L. (2011). A tale of two communities: fighting poverty in the great society (1964–68). *History of political economy*, 43(1), 199–223.

Fortun, M. and Schweber, S.S., 1993. Scientists and the legacy of world war ii: the case of operations research (OR). *Social Studies of Science*, 23(4), 595-642.

Foss, N. (1999). Austrian economics and game theory: a stocktaking and an evaluation. *The Review of Austrian Economics*, 13, 41–58.

Foster, D., Young, P. (1990). Stochastic evolutionary game dynamics. *Theoretical Population Biology*, 38, 219–232.

Friedell, M. F., 1967. Organizations as semilattices. *American Sociological Review*, 32, 46–54.

Friedell, M. F. (1969). On the structure of shared awareness. *Behavioral Science*, 14, 28–39.

Fudenberg, D., Kreps, D., Levine, D. 1988. On the robustness of equilibrium refinements. *Journal of Economic Theory*, 44, 351-380.

Fudenberg, D., and Tirole, J. 1991. Perfect Bayesian Equilibrium And Sequential Equilibrium. *Journal of Economic Theory*, 53, 236-260.

Fullbrook, E., (2001). Conceptual displacement: from the natural to the social", *Review of Political Economy*, 54(3), 285–96.

Fullbrook, E. (2002). An intersubjective theory of value. In *Intersubjectivity in Economics: Agents and Structures*, Fullbrook, E. (ed.), London: Routledge, 273–299.

Fullbrook, E. (2004). Descartes' legacy: Intersubjective reality, intrasubjective theory. In *The Elgar Companion to Economics and Philosophy*, Davis, J. B., Marciano A., Runde J. (ed.), Cheltenham: Edward Elgar Publishing, 403-422.

Frith, U., Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society, series B*, 358, 459–473.

Gallagher, S. (2004) Understanding interpersonal problems in autism: interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry, & Psychology*, 11(3), 199-217.

Gallagher, S. (2008). Direct perception in the intersubjective context. Consciousness and Cognition, 17(2), 535-543.

Gallagher, H. L., Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, 7(2), 77-83.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, 38(1), 11-21.

Gallagher, H. L., Jack, A. I., Roepstorff, A., Frith, C. D. (2002). Imaging the Intentional Stance in a Competitive Game. NeuroImage, 16(3), 814-821.

Gallese, V. (2003). The manifold nature of interpersonal relations: The quest for a common mechanism. *Philosophical Transactions of the Royal Society of London B, Biological Science*, 358, 517–528.

Gallese,V., Fadiga, L., Fogassi, L., Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 593–609.

Gallese, V., Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences*, 2, 493–501.

Gallese, V., Keysers, C., Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8, 396–403.

Galison, P., (1998). The Americanization of unity. *Daedalus*, 127 (1), 45–71.

Garcia, S. M., Weaver, K., Moskowitz, G. B., Darley, J. M. (2002). Crowded minds: The implicit bystander effect. *Journal of Personality and Social Psychology*, 83, 843–853.

Gardner, M., Nozick, R. (1974). Mathematical games. *Scientific American*, 230, 102–109.

Gauthier, D. (1975). Coordination, *Dialogue*, 14, 195-221.

Geanakoplos, J. (1992). Common Knowledge. *Journal of Economic Perspectives*, 6(4), 53-82.

Geary, D., (2011). Racial liberalism, the Moynihan Report & the Daedalus Project on "The Negro American." *Daedalus*, 140 (1), 53–66.

German, T. P., Leslie, A. M. (2000). Attending to and learning about mental states. In *Children's reasoning and the mind*, Mitchell, P., and K. Riggs (eds.), Hove, UK: Psychology Press, 229–252.

Gilbert, M. (1981). Game theory and "Convention". *Synthese*, 46(1), 41-93

Gilbert, M. (1989). *On social Facts*, London: Routledge

Gilbert, M. (1990). Rationality, coordination, and convention. *Synthese*, 84, 1–21.

Gilbert, M. (1996). *Living together: Rationality, sociality, and obligation*. Lanham, MD: Rowman & Littlefield.

Gilbert, M. (2000). *Sociality and responsibility, new essays in plural subject theory*. Lanham, MD: Rowman and Littlefield Publishers.

Gilbert, M. (2003). The structure of the social atom: joint commitment as the foundation of human social behavior', in *Socializing metaphysics*, Frederick, F.S. (ed.), Lanham, MD: Rowman & Littlefield.

Gilbert, M. (2006). *A theory of political obligation: membership, commitment, and the bonds of society*. Oxford, New York: Oxford University Press.

Gilbert, M. (2013). *Joint commitment: How we make the social world*. Oxford, New York: Oxford University Press.

Gilbert, D. T., Gill, M. J., Wilson, T. D. (2002). The future is now: temporal correction in affective forecasting. *Organizational Behavior and Human Decision Processes*, 88(1), 430-444.

Gilboa, I. (2011). Why the empty shells were not fired: A semi-bibliographical note. Episteme, 8(3), 301 – 308.

Gilovich, T., Griffin, D., Kahneman, D. (2002). *Heuristics and biases: the psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Gilovich, T., Savitsky, K., Medvec, Y. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, 75, 332–346.

Gioccoli, N. (2003). *Modeling Rational Agents: From the Interwar Economics to Early Modern Game Theory*. Massachusetts: Edward Elgar.

Godfrey-Smith, P. (2002). On the evolution of representational and interpretive capacities. Monist, 85(1), 50–69.

Govindan, S., Wilson, R. B. (2008). Refinements of Nash equilibrium. In *The new palgrave dictionary of economics* (2nd ed.), Durlauf, S., andL. Blume (eds.), New York: Palgrave Macmillan.

Goel, Grafman, Sadato and Hallett, (1995). Modeling other minds. *Cognitive Neuroscience and* Neuropsychology, Neuroreport, 6, 1741-1747.

Gold, N. (2018). Team reasoning and spontaneous collective intentions. *Revue d'économie politique*, 128(3), 333-353.

Gold, N., List, C. (2004). Framing as path dependence. *Economics & Philosophy*, 20(2), 253-277.

Gold, N., and Colman, A. M. (2018). Team Reasoning and the Rational Choice of Payoff-Dominant Outcomes in Games. *Topoi*, 1-12.

Gold, N., Sugden, R. (2007), Theories of team agency. In *Rationality and Commitment*, Ed. by Peter and Schmidt, Oxford University Press, 281-312

Goldman, A. (1986). *Epistemology and cognition*. Cambridge, MA: Harvard University Press.

Goldman, A. (1989). Interpretation psychologized. *Mind and Language*, 4, 161–185.

Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences*, 16, 15-28.

Goldman, A. (1995). Simulation and interpersonal utility. *Ethics*, 105, 709–726.

Goldman, A. (2002). Simulation theory and mental concepts. In *Simulation and knowledge of action*, Dokic, J., and J. Proust (eds.), Amsterdam: John Benjamins, 1–20.

Goldman, A. (2003). Conceptual clarification and empirical defense of the simulation theory of mindreading. In *Persons: An Interdisciplinary Approach*, Kanzian, C., Quitterer, J., and E. Runggaldier (eds.),Vienna: Obvahaupt.

Goldman, A. (2004). Epistemology and the evidential status of introspective reports. Journal of Consciousness Studies, 11(7–8), 1–16.

Goldman, A. (2005). Imitation, mind reading, and simulation. In *Perspectives on imitation: from neuroscience to social science*, Vol. 2, Hurley, S., and N. Chater (eds.), Cambridge, MA: MIT Press, 79–93.

Goldman, A., (2006). Simulating minds: the philosophy, psychology, and neuroscience of mindreading, Oxford, NY: Oxford University Press.

Goldman, A. (2010). Why social epistemology is real epistemology. In *Social Epistemology*, Haddock, A., Millar, A., and D.Pritchard (eds.), Oxford, NY: Oxford University Press, 1-29.

Goldman, A. (2012). Theory of mind. In *Oxford Handbook of Philosophy and Cognitive Science*, Margolis, E., Samuels, R., and S. Stich (eds.), Oxford, NY: Oxford University Press, 402–424.

Goldman, A., Sripada, C. (2005). Simulationist models of face-based emotion recognition. *Cognition*, 94, 193–213.

Goldman, A., Shanton, K. (2012). The case for simulation theory. In *Handbook of 'Theory of Mind'*, Leslie, A., and T. German (eds.), Mahwah, NJ: Erlbaum, 202–204.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16, 1–14.

Gopnik, A., Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In The Cognitive basis of science, Carruthers, P., Stich, S., and M. Siegel (eds.), Cambridge: Cambridge University Press, 117–132.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 3–32.

Gopnik, A., Sobel, D. M., Schulz, L. E., Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37, 620–629.

Gordon, R. (1986). Folk Psychology as simulation. *Mind and Language*, 1, 158-171.

Gordon, R. (1995a). Sympathy, simulation, and the impartial spectator. *Ethics*, 105, 727–742.

Gordon, R. (1995b). Simulation without introspection or inference from me to you. In *Mental Simulation*, Stone, T., and M. Davies (eds.), Oxford: Blackwell, 53–67.

Gordon, R. (1996). "Radical" simulationism. In *Theories of theories of mind*, Carruthers, P., and P. Smith (eds.), Cambridge: Cambridge University Press, 11–21.

Goyal, S., Janssen, M. (1996). Can we rationally learn to coordinate?. *Theory and Decision*, 40(1), 29-49.

Grüne-Yanoff, T., Lehtinen, A. (2012). Philosophy of game theory. In *Handbook of the Philosophy of Economics*, Mäki, U. (ed.), Oxford: North-Holland, 531-576.

Guala, F. (2016). Understanding institutions: the science and philosophy of living together. Princeton: Princeton University Press.

Guala, F. (2018). Coordination, team reasoning, and solution thinking. *Revue d'économie politique*, 128(3), 355-372.

Guala, F. (forthcoming), Solving the Hi-lo Paradox: Equilibria, Beliefs, and Coordination, in *Minimal Cooperation and Shared Agency*, edited by A. Fiebich. Dordrecht: Springer.

Gul, F. (1998). A comment on Aumann's Bayesian view. *Econometrica*, 66, 923–927.

Gul, F., Pesendorfer, W. (2005). The Canonical Type Space for Interdependent Preferences. NajEcon Working Paper Reviews, Princeton University from www.najecon.org

Gul, F., Pesendorfer, W. (2016). Interdependent Preference Models as a Theory of Intentions. *Journal of Economic Theory*, 165, 179-208.

Gurin, P., Markus, H. (1988). Group identity: The psychological mechanisms of durable salience. *Revue Internationale de Psychologie Sociale*, 1(2), 257-274.

Güth, W. (1995). An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory*, 24, 323-344.

Hahn, F. (1973). On the notion of equilibrium in economics. Inaugural lecture, working paper *Cambridge University*.

Hammond, P. J. (2009). Isolation, assurance and rules: can rational folly supplant foolish rationality? In *Arguments for a better world: Essays in honor of Amartya Sen. Volume I: Ethics, welfare, and measurement*, Basu, K., and R. Kanbur (eds.), Oxford, NY: Oxford University Press.

Hargreaves Heap, S. P. (2004). Economic Rationality. In *The Elgar Companion to Economics and Philosophy*, Davis, J. B., Marciano, A., and J. Runde (eds.), Cheltenham: Edward Elgar Publishing, 42-60.

Hargreaves Heap, S. P., Varoufakis, Y. (2004[1995]). *Game theory: A critical text*, 2nd Ed., London: Routledge.

Harman, G. (1973). *Thought*. Princeton, NJ: Princeton University Press.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, 1, 576–577.

Harsanyi, J. C. (1956). Approaches to the bargaining problem before and after the theory of games: a critical discussion of Zeuthen's, Hicks', and Nash's theories. *Econometrica*, 24(2), 144-157.

Harsanyi J. C., (1961), "Theoretical Analysis in Social Science and the Model of Rational Behavior", *Australian Journal of Politics and History*, 7, 60-74.

Harsanyi, J. C. (1962). Bargaining and international relations. *The Journal of Conflict Resolution*, 6(1), 29-38.

Harsanyi, J. C. (1967/68). Games with Incomplete Information Played by 'Bayesian' Players. *Management Science* 14(3-5-7), 159-182, 320-334, 486-502.

Harsanyi, J. C. (1975). The tracing procedure: A Bayesian approach to defining a solution for n-person non cooperative games. *International Journal of Game Theory*, 4, 61–94.

Harsanyi, J. C. (1976). A solution concept for n-person non cooperative games. *International Journal of Game Theory*, 5, 211–225.

Harsanyi, J. C. (1977). *Rational behavior and bargaining equilibrium in games and social situations*, Cambridge: Cambridge University Press.

Harsanyi, J. C. (1982a). Comment – Subjective probability and the theory of games: Comments on Kadane and Larkey's paper. *Management Science*, 28, 120–124.

Harsanyi, J. C. (1982b). Rejoinder to Professors Kadane and Larkey. *Management Science*, 28, 124–125.

Harsanyi, J. C. (1987). Review of Gauthier's "Morals by Agreement", *Economics and Philosophy*, 3, 339-343.

Harsanyi, J. C. (1995). Games with Incomplete Information. The American Economic Review, 85(3), 291-303.

Harsanyi, J. C. (2004). Games with Incomplete Information Played by "Bayesian" Players, I-III. *Management Science*, 50(12), 1804-1824.

Harsanyi, J. C., Selten, R. (1988). A General Theory of Equilibrium Selection in Games, Cambridge, MA: MIT Press.

Hart, S. (2006). Robert Aumann's Game and Economic Theory. The Scandinavian Journal of Economics, 108(2), 185-211.

Haruvy, E., Stahl, D. E., Wilson, P. W. (1999). Evidence for optimistic and pessimistic behavior in normal-form games. *Economics Letters*, 63, 255–259.

Hausman, D. M. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy*, 16, 99–115.

Hausman, D. M. (2012). Preference, choice, value and welfare. NY: Cambridge University Press.

Heal, J. (1986). Replication and functionalism. In *Language, mind, and logic*, J. Butterfield (ed.), Cambridge, NY: Cambridge University Press.

Heal, J. (1996). Simulation and cognitive penetrability. *Mind and Language*, 11, 44–67.

Heal, J. (1998). Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language*, 13, 477–98.

Heath, C., Ho, B., Berger, J., (2006). Focal points in coordinated divergence. *Journal of Economic Psychology*, 27, 635-647.

Hédoin, C. (2012) Linking institutions to economic performance: the role of macro-structures in micro-explanations. *Journal of Institutional Economics*, 8 (3), 327-349.

Hédoin, C. (2014). A Framework for community-based salience: common knowledge, common understanding and community membership. *Economics and Philosophy*, 30(3), 365-395.

Hédoin, C. (2015). Accounting for constitutive rules in game theory. *Journal of Economic Methodology*, 22, 439–61.

Hédoin, C. (2016). Community-based reasoning in games: Salience, rule-following, and counterfactuals', Games, 7(4), 36.

Hédoin, C. (2017). Institutions, rule-following and game theory. *Economics and Philosophy*, 33, 43–72.

Hédoin, C. (2018). Institutions, rule-following and conditional reasoning. *Journal of Institutional Economics*, 15(1), 1-25.

Hédoin, C., Larrouy, L. (2016). Game Theory, institutions and the Schelling-Bacharach principle: Toward an empirical social ontology. GREDEG Working Papers 2016-21 (GREDEG CNRS), Université Côte d'Azur, France.

Heidl, S. (2016). Philosophical problems of behavioural economics. Abingdon: Routledge.

Heifetz, A. (2018). Epistemic game theory: Incomplete information. In *The New Palgrave Dictionary of Economics*:, Macmillan Publishers Ltd (eds.), London Palgrave: Macmillan.

Hegselmann, R. (2012). Thomas C. Schelling and the computer: some notes on Schelling's essay "On letting a computer help with the work." *Journal of artificial societies and social simulation*, 15(4), 1–6.

Heider, F. (1958). *The psychology of interpersonal relations*, NY: Wiley.

Hewstone, M. (1983). *Attribution Theory*, Oxford: Basil Blackwell.

Hichri, W., Kirman, A. (2007). The Emergence of coordination in public good games. *The European Physical Journal B*, 55, 149-159.

Hillas, J., Kohlberg, E. (2002). The foundations of strategic equilibrium. In *Handbook of Game Theory III*, Aumann, R. and S. Hart (eds.), Amsterdam, North-Holland: Elsevier Science Publishers.

Hodgson, G., (1988). *Economics and Institutions*, Cambridge: Polity Press.

Hodgson, G. (2002). Reconstitutive Downward causation: Social structure and then development of individual agency. In *Intersubjectivity in Economics: Agents and Structures*, Fullbrook, E. (ed.), London: Routledge, 159–80.

Hollis, M. (1998). *Trust within reasons*. Cambridge: Cambridge University Press.

Hollis, M., Sugden, R. (1993). Rationality in action. *Mind New Series*, 102(405), 1-35.

Hounshell, D. A. (1997). The Cold War, RAND, and the generation of knowledge. *Historical Studies in the Physical and Biological Sciences*, 27(2), 237–67.

Hume, D. (1958[1739]). A Treatise of Human Nature, 1st ed., L. A. Selby-Bigge, ed. New York: Oxford University Press.

Hurley , S. (1989). *Natural reasons*. New York: Oxford University Press.

Hurley, S. (1991). Newcomb's Problem, Prisoners' Dilemma, and collective action. *Synthese*, 86, 173–196.

Hurley, S. (1998). *Consciousness in action*. Cambridge, MA: Harvard University Press.

Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Science*, 31(1), 1-22.

Hutchison, W. D., Davis, K. D., Lozano, A. M., Tasker R. R., Dostrovsky, J. O. (1999). Pain-related neurons in the human cingulate cortex. *Nature Neuroscience*, 2, 403-405.

Hutto, D. D. (2008). *Folk psychological narratives. The sociocultural basis of understanding reasons*. A Bradford Book, Cambridge, MA: The MIT Press.

Hutto, D. D. (2009). Mental representation and consciousness. In Encyclopedia of Consciousness, Banks W. P. (eds.). Oxford, UK: Academic Press, 19-32.

Iacobini, M. (2007). Face to face: the neural basis of social mirroring and empathy. *Psychiatric Annals*, 37(4), 236-241.

Iacoboni, M. (2009) Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60, 653-70.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, 3, 529–535.

Innocenti, A. (2007). Player heterogeneity and empiricism in Schelling. *Journal of Economic Methodology*, 14(4), 409-428.

Isaac, J. (2010). Tool shock: technique and epistemology in the postwar social sciences. *History of political economy*, 42, 133–164.

Jackson, P. L., Meltzoff, A. N., Decety, J. (2004). How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage*, 24, 771–779.

Janssen, M. C. (2001). Rationalizing focal points. *Theory and Decision*, 50(2), 119-148.

Janssen, M. C. (2006). On the strategic use of focal points in bargaining situations. *Journal of Economic Psychology*, 27(5), 622-634.

Jardini, D. R. (1996) 'Out of the blue yonder: the RAND Corporation's diversification into social welfare research, 1946–1968', PhD dissertation, Carnegie Mellon University, Pittsburgh.

Jardini, D.R., 2000. Out of the blue yonder: the transfer of systems thinking from the Pentagon to the Great Society, 1961–1965. In: T.P. Hughes and A.C. Hughes, eds. Systems, experts, and computers. The systems approach in management and engineering, World War II and after. Cambridge, MA: The MIT Press, 311–57.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, 14, 103–109.

Jeannerod, M., Pacherie, E. (2004). Agency, simulation, and self-identification. *Mind and Language*, 19(2), 113–146.

Jeffrey, R. C. (1990). *The logic of decision*. Chicago, IL: University of Chicago Press.

Kadane, J. B., Larkey, P. D. (1982a). Subjective probability and the theory of games. *Management Science*, 28, 113–120.

Kadane, J. B., Larkey, P. D. (1982b). Reply to Professor Harsanyi. *Management Science*, 28, 124.

Kahneman, D., Slovic, P., Tversky, A., eds. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.

Kahneman, D. and A. Tversky, eds. (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

Kaneko, M. (2013). Symposium: Logic and economics-interactions between subjective thinking and objective worlds. *Economic Theory*, 53(1), 1-8.

Kanwisher, N., McDermott, J., Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17, 4302–4311.

Kelly, Y. (2009). Mises, Morgenstern, Hoselitz and Nash: The Austrian connection to early game, *The Quarterly Journal of Austrian Economics*, 12(3), 37-42.

Kelly, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, 14, 785–794.

Keren, G. (2011). On the definition and possible underpinnings of framing effects: A brief review and a critical evaluation. In *Perspectives on Framing*, Keren, G. (ed.), New York: The Society for judgment and decision making series, Psychology Press.

Keysar, B. (1994). The illusion of transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology*, 26, 165–208.

Keysar, B., Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language*, 34, 89–109.

Keysar, B., Lin, S., Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25–41.

Keysers, C., Wicker, B., Gazzola, V., Anton, J.-L., Fogassi, L., Gallese, V. (2004). A touching sight: SII/PV activation during the observation of touch. *Neuron*, 42, 335–346.

Kirman, A. (2011) Complex economics: Individual and collective rationality, Abingdon: Routledge.

Kirman, A., Vinkovic, C. (2006). A Physical analogue of the Schelling model. *Proceedings of the National Academy of Sciences*, 103, 19261-19265.

Kirman, A., Teschl, M. (2010). Selfish of selfless? The role of empathy in economics. *Philosophical Transactions of the Royal Society B*, 365, 303–317.

Klaes, M. (2008). Rationality and its bounds: Re-framing and social framing. In *Reasoning, Rationality and Probability*, Galavotti, Scazzieri, Suppes, (eds.) CSLI Publication, Chicago University Press.

Kohlberg, E. (1975a). Optimal strategies in repeated games with incomplete information. *International Journal of Game Theory*, 4(1), 7-24.

Kohlberg, E. (1975b). The information revealed in infinitely-repeated games of incomplete information. *International Journal of Game Theory*, 4(2), 57-59.

Kohlberg, E. 1990. Refinement of Nash Equilibrium: The Main Ideas. Game Theory and Applications, edited by T. Ichiishi, A. Neyman, and Y. Tauman. San Diego: Academic Press.

Kohlberg, E., and Mertens, J-F. 1986. On the Strategic Stability of Equilibria. Econometrica 54, 1003-1038.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297, 846–848.

Krauss, R. M., Glucksberg, S. (1969). The development of communication: Competence as a function of age. *Child Development*, 40, 255–266.

Kramer, R., Brewer, M. (1984). Effects of group identity on resource use in a simulated commons dilemmas. *Journal of Personality and Social Psychology*, 46(5), 1044-1057.

Kreps, D. 1990. *Game theory and economic modeling*. New York: Oxford University Press.

Kreps, D., Wilson, R. 1982. Sequential Equilibria. *Econometrica*, 50, 863-894.

Kreps, D., Milgrom, P., Roberts, J., Wilson, R. (1982). Rational cooperation in the finitely repeated prisoner's' dilemma. *Journal of Economic Theory*, 27, 245-252.

Kripke, S. A. (1963). Semantical analysis of modal logic I Normal modal propositional calculi. Mathematical Logic Quarterly, 9: 67-96.

Kuhn, H. W., Tucker, A. W. (1958). John von Neumann's work in the theory of games and mathematical economics. *Bulletin of the American Mathematical Society*, 64, 100-122.

Larrouy, L., Lecouteux, G. (2017). Mindreading and endogenous beliefs in games. Journal of Economic Methodology, 24(3): 318-343.

Latzko, D. (1998). Thomas Schelling's dissent from the narrow scope of economics. In *Economics and its discontents: Twentieth century dissenting economists*, Holt, R.P.F., and S. Pressman (eds.), Cheltenham: Edward Elgar Publishing.

Lawson T., (2003). Theorizing ontology. *Feminist Economics*, 9(1), 161-169.

Lecouteux, G. (2015). Reconciling Normative and Behavioural Economics. PhD thesis, École Polytechnique.

Lecouteux, G. (2018a). What does "we" want? Team reasoning, game theory, and unselfish behaviours. *Revue d'économie politique*, 128(3), 311-332.

Lecouteux, G. (2018b). Bayesian game theorists and non-Bayesian players. *European Journal of the History of Economic Thought*,25(6), 1420-1454.

Lehtinen, A. (2011). The revealed-preference interpretation of payoffs in game theory. *Homo Oeconomicus*, 28, 265–296.

Lempert, D. (2018). On evolutionary game theory and team reasoning. *Revue d'Economie Politique*, 128(3), 423-446.

Léonard, R. J. (1992). Creating a context for game theory. In *Toward a history of game theory*, Weintraub, E. R. (ed.), Durham: Duke University Press, 29-76.

Léonard, R. J. (2010). *Von Neumann, Morgenstern, and the creation of game theory: From chess to social science, 1900–1960*. New York: Cambridge University Press.

Leslie, A. (1987). Pretence and representation: The origins of "theory of mind." *Psychological Review*, 94, 412–426.

Leslie, A. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition*, 50, 211–238.

Leslie, A. (2000). How to acquire a representational theory of mind. In *Metarepresentations: A multidisciplinary perspective*, Sperber, D. (ed.), New York: Oxford University Press, 197–223.

Leslie, A., Friedman, O., German, T. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences*, 8, 528–533.

Leslie, A., German, T. (1995). Knowledge and ability in "theory of mind": One eyed overview of a debate. In *Mental Simulation*, Davies, M., and T. Stone (eds.), Oxford: Blackwell, 123–150.

Leslie, A., German, T., Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50, 45–85.

Leslie, A., Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science*, 1, 247–253.

Leslie, A., Roth, D. (1993). What autism teaches us about metarepresentation. In *Understanding Other Minds: Perspectives from Autism*, Baron-Cohen, S., Tager-Flusberg, H., and D. Cohen (eds.), Oxford: Oxford University Press, 83–111.

Leslie, S.W. (1993). *The Cold War and American science: the military–industrial–academic complex at MIT and Stanford*. New York: Columbia University Press.

Lesourne J., Orléan A., Walliser B., (2006), *Evolutionary Microeconomics*, Berlin: Springer.

Levi, I. (1995). The common prior assumption in economic theory. *Economics and Philosophy*, 11(2), 227-253.

Levi, I. (1998). Prediction, Bayesian deliberation and correlated equilibrium. In *Game theory, experience, rationality*, Leinfellner, W., and E. Köhler (eds.), Dordrecht: Springer, 173–185.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3), 593-622.

Lewis, D., (2002 [1969]), *Convention*, Oxford: Blackwell Publisher.

Lewis, D. (1980[1972]) Psychophysical and theoretical identifications. In Readings in philosophy of psychology, Vol. 1, Block, N. (ed.), Cambridge, MA: Harvard University Press.

Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63, 17–25.

Lieberman, M. (2007). Social Cognitive Neuroscience: A Review of Core Processes. *Annual Review of Psychology*, 58(1), 259-289.

Light, J.S. (2005). *From warfare to welfare. Defense intellectuals and urban problems in Cold War America.* Baltimore, MD: Johns Hopkins University Press.

Lipman, M. (1995). Good thinking. *Inquiry: Critical Thinking Across the Disciplines*, 15(2), 37-41.

Lipps, T. (1907a). Das Wissen von fremden Ichen. In *Psychologische Untersuchungen I*, Lipps, T. (ed.), Leipzig: Engelmann, 694–722.

Lipps, T. (1907b). Ästhetik. In *Systematische Philosophie*, Hinneberg, P. (ed.), Berlin: Verlag von B. G. Teubner, 351–390.

Lipps, T. (1909). *Leitfaden der Psychologie*. Leipzig: Verlag von Wilhelm Engelmann.

Livet, P. (2007). Towards an epistemology of multi-agent simulations in social sciences. In *Agent based modeling and simulations in the human and social sciences*, Phan, D., and F. Amblard (eds.), Oxford: The Bardwell Press.

Livet, P., Phan, D., Sanders, L. (2008). Why do we need ontology for agent-based models? In *Complexity and artificial markets*, Schredelseker, K., and F. Hauser (eds.), Berlin: Springer.

Loewenstein, G., Prelec, D., Shatto, C. (1998). Hot/cold intrapersonal empathy gaps and the prediction of curiosity. Working paper, Carnegie Mellon University.

Luce, R. D., & Raiffa, H. (1957). Games and decisions: Introduction and critical survey. New York, NY: Wiley.

Maibom, H. L. (2003). The mindreader and the scientist. *Mind and Language*, 18(3), 296-315.

Maibom, H. L. (2007). Social systems. *Philosophical Psychology*, 20(5), 257-578.

Malle, B. F. (2011). Attribution Theories: How People Make Sense of Behavior. *Theories in social psychology*, Chadee, D. (ed.), Wiley Blackwell, 72-95.

Mandler, M., Manzini, P., Mariotti, M. (2012). A million answers to twenty questions: Choosing by checklist. *Journal of Economic Theory*, 147(1), 71-92.

Maraffa, M. (2015). Mindreading and Introspection. *Rivista Internazionale di Filosofia e Psicologia*, 6(2), 249-260.

Mariotti, M. (1995). Is Bayesian rationality compatible with strategic rationality? *The Economic Journal*, 105, 1099-1109.

Mariotti, M. (1996). The decision-theoretic foundations of game theory. In *The rational foundations of economic behavior*, Arrow, K., Colombatto, E., Perlman, M., and C. Schmidt (eds.), Basingstoke: Macmillan Press, (pp. 133–148).

Mariotti, M. (1997). Decisions in games: why there should be a special exemption from Bayesian rationality. *Journal of Economic Methodology*, 4(1), 43-60.

Masel, J. (2007). A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. *Journal of Economic Behavior & Organization*, 64, 216–231.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, 98, 11832–11835.

McCall, J.J. (1968). Economics of information and job search. RAND/RM-5745-OEO. Santa Monica, CA: RAND Corporation.

McCall, J.J. (1970). Racial discrimination in the job market: the role of information and search. RAND/RM-6162-OEO. Santa Monica, CA: RAND Corporation.

Mearman, A. (2003). 'Open systems' and economic methodology. Paper presented to the conference of the Association of Heterodox Economics, Nottingham.

Mehta, J. (2013). The discourse of bounded rationality in academic and policy arenas: pathologising the errant consumer. *Cambridge journal of economics*, 37(6), 1243-1261.

Mehta, J., Starmer, C., Sugden, R. (1994a). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, 84(3), 658-673.

Mehta, J., Starmer, C., Sugden, R. (1994b). Focal points in pure coordination games: An experimental investigation. *Theory and Decision*, 36(2), 163-185.

Mertens, J-F (1971). The value of two-person zero-sum repeated games: the extensive case. *International Journal of Game Theory*, 1, 217–227.

Mertens, J-F. (1989). Stable Equilibria—A Reformulation, Part I: Definition and Basic Properties. *Mathematics of Operations Research*, 14, 575-624.

Mertens, J-F. (1992). The Small Worlds Axiom for Stable Equilibria. *Games and Economic Behavior*, 4, 553-564.

Mertens, J.F., Zamir, S. (1971). The value of two-person zero-sum repeated games with lack of information on both sides. *International Journal of Game Theory*, 1, 39–64 https://doi.org/10.1007/BF01753433

Mertens, J-F., Zamir, S. (1985). Formulation of Bayesian Analysis for Games with Incomplete Information. *International Journal of Game Theory*, 14, 1-29.

Meltzoff, A. N. Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London, Series B*, 358, 491–500.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2), 380-391.

Miller, S. (1997). Social Norms. In *Contemporary Action Theory, Vol. 2: Social Action*, Holmström-Hintikka, G. and R. Tuomela (eds.), Dordrecht, Netherland: Kluwer Academic Publishers.

Mirowski, P. (1992). What were von Neumann and Morgenstern trying to accomplish? In *Toward a history of game theory*, Weintraub, E. R. (ed.), Durham: Duke University Press, 113-148.

Mitchell, J. P., Banaji, M. R., Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 17, 1306–1315.

Monderer, D., Samet, D. (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior*, 1(2), 170-190.

Morgenstern, O. (1928). *Wirtschaftprognose: Enie untersuchung ihrer voraussetzungen und möglichkeiten*, Vienna: Julius Springer, pp. iv and 129

Morgenstern, O. (1935). Vollkommene voraussicht und wirtschaftliches gleichgewicht. *Zeitschrift für Nationalökonomie*, 6(3), pp. 337–57.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy*, 11, 227–253.

Morton, A. (2012[2005]), The importance of being understood: Folk psychology as ethics. New York: Routledge.

Myerson, R. B. (1941). An early paper on the refinement of Nash equilibrium. *Duke Mathematical Journal*, 81(1), 67-75.

Myerson, R. B. (1978). Refinement of the Nash Equilibrium concept. *International Journal of Game Theory*, 7, 73-80.

Myerson, R. B. (1999). Nash Equilibrium and the history of economic theory", *Journal of Economic Literature*, 37(3), 1067-1082.

Myerson, R. B. (2004). Comments on "Games with Incomplete Information Played by 'Bayesian'Players, I–III Harsanyi's Games with Incoplete Information". *Management Science*, 50, 1818-1824.

Myrdal, G. (1944). An American dilemma: the negro problem and American democracy. New York: Evanston.

Nash, J. F. (1950a). The bargaining problem, *Econometrica*, 18,155–162.

Nash, J. F. (1950b). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36, 48-49.

Nash, J. F. (1951). Non-Cooperative Games. *Annals of Mathematics*, 54, 286-295.

Nash, J. F. (1953). Two-Person Cooperative Games. *Econometrica*, 21, 128-140.

Nash, J. F. (1996). *Essays on Game Theory*. Cheltenham, UK: Edward Elgar.

Nash, J. F., Shapley, L. (1950). A simple three-person poker game. In *Contributions to the Theory of Games*, Kuhn, H. W., and A. W. Tucker (eds.), 105–116.

Neyman, A. (1985). Bounded complexity justifies cooperation in the finitely repeated Prisoners' Dilemma, *Economics Letters*, 19, 227-229.

Newton, E. (1990). Overconfidence in the Communication of Intent: Heard and Unheard Melodies. Unpublished doctoral dissertation, Stanford University.

Nichols, S., Stich, S. P., Leslie, A., Klein, D. (1996). Varieties of off-line simulation. In *Theories of Theories of Mind*, Carruthers, P., and P. Smith (eds.), Cambridge: Cambridge University Press, 39–74.

Nichols, S., Stich, S. (2003). Mindreading: An integrated account of pretence, self-awareness, and understanding of other minds. Oxford: Oxford University Press.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125, 737–759.

Nickerson, R. S. (2001). The projective way of knowing. *Current Directions in Psychological Science*, 10, 168–172.

Oakes P., Haslam A., Turner J., 1994: *Stereotyping and Social Reality*, Oxford: Blackwell.

Okasha, S. (2009). Biological Altruism. In *The Stanford Encyclopedia of Philosophy*, Zalta, E. N. (ed.), Cambridge: Cambridge University Press.

O'Connor, A. (2001). *Poverty knowledge: social science, social policy, and the poor in twentieth centrury U.S. history*. Princeton, NJ: Princeton University Press.

O'Craven, K. M., Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, 12, 1013–1023.

Orléan, A. (2004). What is a collective belief? In *Cognitive economics: An interdisciplinary approach*, Bourgine, P., and J. P. Nadal (eds.), Berlin: Springer.

Pascal, A.H. (1965). Reconnaissance for the war on poverty. RAND/P-3092. Santa Monica, CA: RAND Corporation.

Pascal, A.H. (1971). Enhancing opportunities in job markets: summary of research and recommendations for policy. R-580-OEO. Santa Monica, CA: RAND Corporation.

Pascal, A.H. (1972). *Racial discrimination in economic life*. New York: Lexington Books.

Pascal, A.H., McCall, J.J. (1967). Rand studies for OEO. RAND/D-16142-OEO. Santa Monica, CA: RAND Corporation.

Patterson, J.T. (2000). *America' s struggle Against Poverty in the Twentieth Century*. US: Harvard University Press

Pearl, J. (2000). *Causality*. New York: Oxford University Press.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. Econometrica, *Journal of the Econometric Society*, 1029-1050.

Perdue C. W., Dovidio, J. F., Gurtman, M. B., Tyler, R. B. (1990) Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, 59(3), 475-486.

Perea, A. (2007). A one-person doxastic characterization of Nash strategies, Synthese 158, 251-271.

Perea, A. (2012). *Epistemic game theory: Reasoning and choice*, Cambridge: Cambridge University Press.

Perea, A. (2014). From classical to epistemic game theory. *International Game Theory Review, 16*, 1440001-1–1440001-22.

Perner, J. (1991). *Understanding the Representational Mind.* Cambridge, MA: MIT Press.

Phan, D. (2004). From agent-based computational economics to cognitive economics. In *Cognitive economics: An interdisciplinary approach*, Bourgine, P., and J. P. Nadal (eds.), Berlin: Springer.

Phan, D., Schmid, A.F., Varenne, F. (2007). Epistemology in a nutshell: Theory, model, simulation and experiment. In *Agent based modeling and simulations in the human and social sciences*, Phan, D., and F. Amblard (eds.), Oxford: The Bardwell Press.

Polak, B. (1999). Epistemic conditions for Nash equilibrium, and common knowledge of rationality. *Econometrica*, 67(3), 673-676.

Ponssard, J. P. (1975a). A note on the LP formulation of zero-sum sequential games with incomplete information. *International Journal of Game Theory*, 4(1), 1-5.

Ponssard, J. P. (1975b). Zero-sum games with "almost" perfect information. *Management Science*, 21(7), 794-805.

Ponssard, J. P., Zamir, S. (1973). Zero-sum sequential games with incomplete information. *International Journal of Game Theo*ry, 2(1), 99-107.

Postema, G. J. (2008). Salience reasoning. *Topoi*, 27, 41–55.

Premack, D., Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1, 515–526.

Prentice, D., Miller, D. (1992). The psychology of ingroup attachment. Paper presented at conference on The Self and the Collective, Princeton University.

Pulford, B. D., Colman, A. M., Lawrence, C. L., Krockow, E. M. (2017). Reasons for cooperating in repeated interactions: Social value orientations, fuzzy traces, reciprocity, and activity bias. *Decision*, 4(2), 102-122.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Rabbie, J., Horwitz, M. (1969). Arousal of ingroup-outgroup bias by a chance win or loss. *Journal of Personality and Social Psychology*, 13(3), 269-277.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83, 1281-1302.

Raiffa, H. (1992). Game theory at the University of Michigan, 1949-1952. In *Toward a history of game theory*, Weintraub, E. R. (ed.), Durham: Duke University Press, 165-176.

Ramsey, F. P. (1926). The foundations of mathematics. *Proceedings of the London Mathematical Society*, 2(1), 338-384.

Ramsey, F. P. (1931). The Foundations of Mathematics. Braithwaite, R. B. (ed.), London: Routledge and Kegan Paul.

Ratcliffe, M. (2007). *Rethinking commonsense psychology: A critique of folk psychology, theory of mind and simulation*. New York: Palgrave Macmillan.

Rawls, J. (1955). Two concepts of rules. *Philosophical Review*, 64, 3–32.

Read, D., van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, 78, 189–205.

Reddy, V. (2008). *How infants know minds*. Cambridge, MA: Harvard University Press.

Regan, D. (1980). Utilitarianism and cooperation. Oxford: Clarendon Press.

Rellstab, U. (1992). New insights into the collaboration between John von Neumann and Oskar Morgenstern on the *Theory of Games and Economic Behavior*. In Toward a history of game theory, Weintraub, E. R. (ed.), Durham: Duke University Press, 77-94.

Repacholi, B., Gopnik, A. (1997). Early understanding of desires: Evidence from 14 and 18 month olds. *Developmental Psychology*, 33, 12–21.

Riker, W.H. (1997). The entry of game theory into political science. In *Toward a history of game theory*, Weintraub, E.R. (ed.), Durham, NC: Duke University Press, 207-224.

Rivzi, S. A. T. (2007). Introduction : Thomas Schelling's distinctive approach, *Journal of Economic Methodology*, 14(4), 403-408.

Rizzolatti, G., Fadiga, L., Gallese, V., Fogasi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3, 131–141.

Rol, M., (2008). Idealization, abstraction, and the policy relevance of economic theories. *Journal of economic methodology*, 15 (1), 69–97.

Rosch, E. (1978). Principles of Categorization. In *Cognition and categorization*, Rosch, E., and B. Lloyd (eds.), Hillsdale, NJ: Erlbaum.

Ross, D. (2007). H. sapiens as ecologically special: what does language contribute?. *Language sciences*, 29(5), 710-731.

Ross, D., (2011). Game Theory. In *The Stanford Encyclopedia of Philosophy*, Zalta, E. N. (ed.), Cambridge: Cambridge University Press.

Ross, L., Greene, D., House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Personality and Social Psychology*, 13, 279–301.

Rotemberg, J. J. (2008). Minimally acceptable altruism and the ultimatum game. *Journal of Economic Behavior and Organization*, 66, 457-476.

Roth, A., Schoumaker, F. (1983). Expectations and Reputations in Bargaining: An Experimental Study. *American Economic Review*, 73(3), 362-72

Royal Swedish Academy of Sciences (2005) 'The Prize in Economic Sciences 2005' (Supplementary Information to Press Release, 10 October 2005), available at www.kva.se (accessed 25 May 2006).

Rubinstein, A. (1989). The electronic mail game: strategic behavior under "almost common knowledge." *The American Economic Review*, 79, (3), 385-391.

Rubinstein, A. (1991). Comments on the interpretation of game theory. *Econometrica*, 59(4), 909-924.

Rubinstein, A. (2001). A theorist's view of experiments. *European Economic Review*, 45(4–6), 615-628.

Sally, D. (1995). Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992, *Rationality and Society*, 58–92.

Samuelson, P. (1947). *Foundations of Economics Analysis*. Cambridge, MA: Harvard University Press.

Sanghvi, A. P., & Sobel, M. J. (1976). Bayesian games as stochastic processes. International Journal of Game Theory, 5(1), 1–22.

Savage, L. J. (1954). *The foundations of statistics*. New York, NY: Wiley.

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences*, 9, 174–179.

Saxe, R., Kanwisher, N. (2003). People thinking about people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, 19, 1835–1842.

Saxe, R., Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, 43, 1391–1399.

Scazzieri, R. (2008). Coordination, context and patterns of Reasoning. In *Reasoning, rationality and probability*, Galavotti, M. C., Scazzieri, R. and P. Suppes (eds.), Chicago, IL: CSLI Publication, Chicago University Press, 187–207.

Scazzieri, R. (2011). Similarity and uncertainty. In *Fundamental uncertainty. Rationality and plausible reasoning*, M. Dall'aste Brandolini, S. and R. Scazzieri (eds.), Basingstoke: Palgrave Macmillan, 73–103.

Schelling, T. C. (1956). An essay on bargaining. *American Economic Review*, 46, 281–306.

Schelling, T. C. (1959). For the abandonment of symmetry in game theory. *Review of Economics and Statistics*, 41, 213–14.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Schelling, T. C. (1966). Arms and influence. New Haven, CT: Yale University Press.

Schelling, T. C. (1967). Economics and criminal enterprise. *Public Interest*, 7(Spring), 61–78.

Schelling, T. C. (1968). The life you save may be your own. In *Problems in Public Expenditure analysis*, Chase, S. B. (ed.), Washington, DC: Brookings Institution, 127–62.

Schelling, T. C. (1969). Models of segregation. *American Economic Review*, 59(2), 488–93.

Schelling, T. C. (1971a). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143–86.

Schelling, T. C. (1971b). On the ecology of micromotives. *The Public Interest*, 25(Fall), 61–98.

Schelling, T. C. (1972). The process of residential segregation: neighborhood tipping. In *Racial discrimination in economic life*, Pascal, A. H. (ed.), Lexington, MA: D.C. Heath, 157–84.

Schelling, T. C. (1973). Hockey helmets, concealed weapons, and daylight saving: a study of binary choices with externalities. *The Journal of Conflict Resolution*, 17(3), 381-428.

Schelling, T. C. (1979). *Thinking through the energy problem*. New York: Committee for Economic Development.

Schelling, T. C. (1984a). *Choice and consequence*. Cambridge, MA: Harvard University Press.

Schelling, T. C. (1980). The intimate contest for self-command. *Public Interest*, 60(Summer), 94–118.

Schelling, T. C. (1983). Prices as regulatory instruments. In *Incentives for Environmental Regulation*, Schelling, T. C. (ed.) Cambridge, MA: MIT Press, 1–40.

Schelling, T. C. (1984). *Choice and consequence*. Cambridge, MA and London: Harvard University Press.

Schelling, T. C. (2006 [1978]) *Micromotives and Macrobehavior*. London and New York: W.W. Norton. Cambridge, MA: Harvard University Press

Schelling, T. C. (2006) *Strategies of commitment and other essays*, London and New York: W.W. Norton. Cambridge, MA: Harvard University Press

Schelling, T. C., Halperin, M. H. (1961). Strategy and arms control. New York: Twentieth Century Fund.

Schiffer, S. (1972). *Meaning*. Oxford: Clarendon Press.

Scholl, B., Leslie, A. M. (1999). Modularity, development and "theory of mind." *Mind and Language*, 14, 131–153.

Schotter, A. (1976). *Selected writings of Oskar Morgenstern*. New York: New York University Press.

Schotter, A. (1992). Oskar Morgenstern's contraibution to the development of the theory of games. In Toward a history of game theory, Weintraub, E. R. (ed.), Durham: Duke University Press, 95-112.

Schmidt, C., Livet, P. (2014). *Comprendre nos interactions sociales, une perspective neuroéconomique*. Paris: Odile Jacob.

Searle, J.R. (1995) The construction of social reality. New York: Simon and Schuster.

Searle, J.R. (1998). *Mind, language and society*. New York: Basic Books.

Searle, J.R. (2005). What is institution? *Journal of Institutional Economics*, 1(1), p. 1-22.

Searle, J.R. (2010). *Making the social world*. New York: Oxford University Press.

Segal, N. L., Hershberger, S. L. (1999). Cooperation and competition between twins. *Evolution and Human Behavior*, 20, 29–51.

Sellars, W. (1997[1955]). *Empiricism and the philosophy of mind*. Cambridge, MA: Harvard University Press.

Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift fur die gesamte Staatswissenschaft*, 121, 301-324, 667-689.

Selten, R. (1975). Reëxamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4, 25-55.

Sent, E. M. (2007). Some like it cold: Thomas Schelling as a cold warrior. *Journal of Economic Methodology*, 14(4), 455-471.

Sent, E. M. (2006). Esther-Mirjam. Pluralisms in Economics. In *Scientific Pluralism*, Kellert, S., Longino, H. and K.Waters. (eds.),Minneapolis: Minnesota Studies in Philosophy of Science.

Setterfield, M. (2016). Heterodox economics, social ontology, and the use of mathematics. *The New School for Social Research*, Department of Economics, Working Paper.

Shafir, E., Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24, 449–474.

Shanton, K., Goldman, A. (2010). Simulation theory. WIREs *Cognitive Sciences*, 1, 527–538. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/wcs.33/full

Shapiro, L. (2011). *Embodied cognition*. New York: Routledge.

Sherif, M. (1958). Superordinate goals in the reduction of intergroup conflict. American Journal of Sociology, 63(4), 349-356.

Sherif, M. et al., (1961). Intergroup conflict and cooperation: the robbers cave experiments. Norman: University of Oklahoma Book Exchange.

Shin, H.S. (1992). Counterfactuals and a theory of equilibrium in games. In Knowledge, Belief and Strategic Interaction, Bicchieri, C., and M.L. Dalla Chiara (eds.), Cambridge: Cambridge University Press, 397–413.

Shubik, M. (1952). Information, theories of competition, and the theory of games. *Journal of Political Economy*, 60, 145–150.

Shubik, M. (1992). Game theory at Princeton, 1949-1955: A personal reminiscence. In Toward a history of game theory, Weintraub E.R. (ed.), Durham: Duke University Press, 151-164.

Simon, H. (1945). Review of theory of games and economic behavior. *American Journal of Sociology*, 50(6), 558–60

Singer, (2006). The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. Neuroscience and Behavioral Reviews, 30, 855-863.

Singer, T., Fehr, E. (2005). The neuroeconomics of mind reading and empathy. *American Economic Review*, 95, 340–345.

Singer, T., and Lamm, C. (2009). Social neuroscience of empathy. Annual NY Academic Science, 1156, 81-96.

Singer, T., Seymour, B., O'Doherty, J., Stephan, K.E., Dolan, R., Frith, C. (2006). Empathetic neural responses are modulated by perceived fairness of others. *Nature*, 439, 466-469.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R., Frith, C. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303, 1157–1162.

Singer et al., (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303(5661), 1157-62.

Sirigu, A., Duhamel, J., Pillon, B., Cohen, L., Dubois, B., Agid, Y. (1996). The mental representation of hand movements after parietal cortex damage. *Science*, 273, 1564–1568.

Smerilli, A., (2012). We-thinking and vacillation between frames: Filling a gap in Bacharach's theory. *Theory and Decision*, 73(4), 539–60.

Smith, A. (1976[1759]). *A theory of moral sentiments*. Raphael, D. D. and A. L. Macfie (eds.), Oxford: Clarendon.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43(2), 392-436.

Sober, E., Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*, Cambridge, Massachusetts: Harvard University Press.

Soja, N., Carey, S., Spelke, E. (1991). Ontological categories guide inductions of word meaning: Object terms and substance terms. *Cognition*, 38, 179–211.

Sorin, S. (1979). A note on the value of zero-sum sequential repeated games with incomplete information. *International Journal of Game Theory*, 8(4), 217-223.

Spelke, E. (1990). Origins of visual knowledge. In *Visual cognition and action*, Osherson, D. N., Kosslyn, S. M., and J. M. Hollerbach (eds.), Cambridge, MA: MIT Press(99–127)..

Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition*, 50, 431–445.

Spirtes, P., Glymour, C., Scheines, R. (2001). *Causation, prediction, and search* (Springer Lecture Notes in Statistics, 2nd ed., rev.). Cambridge, MA: MIT Press.

Spohn, W. (1977). Where Luce and Krantz do really generalize Savage's decision model. *Erkenntnis*, 11, 123–134.

Sprengelmeyer R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., Przuntek, H. (1999). Knowing no fear. *Proceedings of the Royal Society, series B: Biology*, 266, 2451–2456.

Sripada, C., Stich, S. (2006). A framework for the psychology of norms. In *The innate mind: Culture and cognition*, Carruthers, P., Laurence, S. and S. Stich (eds.), New York: Oxford University Press, 280–301.

Stahl, D. O., Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25(3), 309-327.

Stalnaker, R. (1991). The problem of logical omniscience, I. *Synthese*, 89(3), 425-440

Stalnaker, R. (1999). *Context and content*. Oxford: Oxford University Press.

Stalnaker, R. On Logics of Knowledge and Belief. Philos. Stud. 2006, 128, 169–199./ Stalnaker, R. Belief revision in games: Forward and backward induction1. Math. Soc. Sci. 1998, 36, 31–56.

Steelman, A., 2005. Interview: Thomas Schelling. Region focus, Spring. Available from: www.richmondfed.org/publications/economic_research/region_focus/spring_2005/ interview.cfm

Stein, E. (1996). *Without good reason*. Oxford: Oxford University Press.

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Oxford: Wiley-Blackwell.

Stich, S. P. (1981). Dennett on intentional systems. *Philosophical Topics*, 12, 38–62.

Stich, S., and Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind and Language*, 7, 35–71.

Stich, S., Nichols, S. (1995). Second Thoughts on Simulation. In Mental Simulation: Evaluations and Applications, Davies, M. and T. Stone (eds.), Oxford: Basil Blackwell, 87-108.

Stich, S., Nichols, S. (1997). Cognitive penetrability, rationality and restricted simulation. *Mind and Language*, 12(3–4), 297–326.

Stirling, W. C., Tummolini, L. (2018). Coordinated reasoning and augmented individualism. *Revue d'Economie Politique*, 128(3), 469-492.

Eisenberg, N., Strayer, J. (1987). Critical issues in the study of empathy. In Empathy and its development, Eisenberg, N. and J. Strayer (eds.), Cambridge studies in social and emotional development, Cambridge: Cambridge University Press, 3–13.

Strijbos, D. W., De Bruin, L. (2015). Self-interpretation as first-person mindshaping: implications for confabulation research. *Ethical Theory and Moral* Practice, 18(2), 297-307.

Sugden, R. (1991). Rational choice: a survey of contributions from economics and philosophy. *The economic journal*, *101*(407), 751-785.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, 10(1), 69-89

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, 105, 533-550

Sugden, R. (2000a). Team preferences. *Economics and Philosophy*, 16, 175–204.

Sugden, R. (2000b). Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology*, 7(1), 1–31.

Sugden, R. (2001), The evolutionary turn in game theory", *Journal of Economic Methodology*, 8(1), 113-130

Sugden R., (2002). Beyond sympathy and empathy: Adam Smith's concept of fellow-feeling. Economics and Philosophy, 18(1), 63-87.

Sugden, R. (2005). The logic of team reasoning. In *Teamwork – Multidisciplinary perspectives*, Gold, N. (ed.), New-York: Palgrave MacMillan, 181-199.

Sugden, R. (2007). Collective intentions and team agency. *Journal of Philosophy*, 104(3), 109-137.

Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*, 70, 3-27.

Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology*, 1(1), 143-166.

Sugden, R., Zamarrón, I. E. (2006). Finding the key: The riddle of focal points. *Journal of Economic Psychology*, 27, 609-621.

Swedberg, R. (1990). *Economics and sociology: redefining their boundaries. Conversations with economists and sociologists*. Princeton, NJ: Princeton University Press.

Tajfel, H. (1969). Cognitive Aspects of Prejudice. *Journal of Social Issues*, 25(4), 79-97.

Tajfel, H. (1970). Experiments in intergroup discrimination", *Scientific American Journal*, 223, 96-102.

Tajfel, H. (1972). Social categorization. *Introduction à la psychologie sociale*, *Vol. 1*, Moscovici, S. (ed.), Paris: Larousse, 272-302.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge: Cambridge University Press.

Tajfel, H., Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The social psychology of intergroup relations*, Austin, W. G., and S. Worchel (eds.), Belmont, CA: Brooks/Cole, 33–47.

Tajfel, H., Turner J. C. (1985). An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, Austin W. G., and S. Workel (ed.), Oxford: Blackwell Publishers.

Tan, T .C. C., Werlang, S. R. D. C. (1988). The Bayesian foundations of solution concepts of games. *Journal of Economic Theory*, 45(2), 370–391.

Tan, T .C. C., Werlang, S. R. D. C. (1992). On Aumann's notion of common knowledge: an alternative approach. *Revista Brasileira de Economia*, 46(2), 151-166.

Tan, J. H., Zizzo, D.J. (2008). Groups, cooperation and conflict in games. *The Journal of Socio-Economics*, 37, 1-17.

Taylor, M., (1987). *The Possibility of Cooperation*. Cambridge, MA: Cambridge University Press,

Tesfatsion, L. (2002). Agent-based computational economics: Growing economies from the bottom up. *Artificial Life*, 8, 55–82.

Townsend, D. J., Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: MIT Press.

Thomas, K.A., De Scioli, P., Haque, O. S., Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology*, 107(4), 657–676.

Thompson, E. (2001). Empathy and consciousness. *Journal of Consciousness Studies*, 8(5–7), 1–32.

Thompson, E. (2005). Empathy and human experience. In Science, religion, and the human experience, Proctor, J. D. (ed.), New York: Oxford University Press, 261–285.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Tsakas, E. (2012). Rational belief hierarchies. METEOR, Maastricht University School of Business and Economics. METEOR Research Memorandum, No. 004

Tubaro, P. (2011). Computational economics. In *The Elgar companion to recent economic methodology*, Hands, D. W., and J. B. Davis (eds.), Cheltenham: Edward Elgar.

Tuomela, R. (1995). The importance of us : a philosophical study of basic social notions. Stanford, California : Stanford University Press.

Tuomela, R. (2000). *Cooperation: a philosophical study*. Dordrecht, The Nertherlands: Philosophical Studies Series, Kluwer Academic Publishers.

Tuomela, R. (2002). The philosophy of social practices: a collective acceptance view. Cambridge: Cambridge University Press.

Tuomela, R. (2007). *The philosophy of sociality: the shared point of view*. New York: Oxford University Press.

Tuomela,R. (2013). *Social ontology: collective intentionality and group agents*. New York : Oxford University Press.

Tuomela, R., Miller, K. (1988). We-intentions. *Philosophical Studies*, 53(3), 367–389. https://doi.org/10.1007/BF00353512.

Turner, J. (1985). Social categorization and the self-concept: A social cognitive theory of a group behavior. In *Advanced in Group Processes: Theory and Research Vol. 2*, Lawler, E.J. (ed.), Greenwich, CT: JAI Press, 77-122.

Turner et al. (1987), *Rediscovering the social group: A self-categorization theory*. Oxford: Blackwell.

Tversky A., 1977: "Features of Similarity", *Psychological Review*, 84(4), 327-352.

Tversky, A., Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Sciences, New Series*, 211(4481), 453-458.

Tversky, A., Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59(4), 251-278.

Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., and Rizzolatti, G. (2001). I know what you are doing: A neurophysiological study. *Neuron*, 31(1), 155–165.

Vanberg, V. J. (2008). On the economics of moral preferences. *American Journal of Economics and Sociology*, 67, 605–628.

Van Boven, L., Dunning, D., Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of Personality and Social Psychology*, 79, 66–76.

Van Boven, L., Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, 29(9), 1159–1168.

van Damme, E. (1984). A relation between perfect equilibria in extensive form games and proper equilibria in normal form games. *International Journal of Game Theory*, 13, 1-13.

van Damme, E. (1989). Stable equilibria and forward induction. *Journal of Economic Theory*, 48, 476-496.

van Damme, E. (1991). *Stability and perfection of Nash Equilibria*. Berlin: Springer-Verlag.

Varela, F.J., Thompson, E., Rosch, E. (1991). *The embodied mind. Cognitive science and human experience*. Cambridge, MA: MIT Press.

von Neumann, J. (1928a). Calcul des probabilités – Sur la théorie des jeux. Presented by E. Borel, *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 186(25), 1689–91

von Neumann, J. (1928b). Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100, 295–320. Translated in "On the Theory of Games of Strategy" in Tucker, A.W. and R. D. Luce (eds.) (1959), 13–42.

von Neumann, J. (1984 [1931]), The formalist foundations of mathematics, in Benacerraf, P. and Putnam, H. (eds), *Philosophy of Mathematics. Selected Readings,* Benacerraf, P., and H. Putnam (eds.), Cambridge: Cambridge University Press, 61-65.

von Neumann J., A (1953). Certain Two-person Game Equivalent to the Optimal Assignment Problem. In Contributions to the Theory of Games, Vol. II, Kuhn, H.W. and A. W. Tucker, *Annals of Mathematics Studies*, 28, 5–12

von Neumann J., Morgenstern O., (1944), *The Theory of Games and Economic Behaviour*, Princeton, NJ: Princeton University Press

Walliser, B. (1989). Instrumental rationality and cognitive rationality. *Theory and Decision*, 27, 7-36.

Walliser, B. (2004). Topics of cognitive economics. In *Cognitive economics: An interdisciplinary approach*, Bourgine, P., and J. P. Nadal (eds.), Berlin: Springer.

Ware, W.H. (2008). RAND and the information evolution: a history in essays and vignettes. Santa Monica, CA: RAND Corporation.

Weibull, J. W. (1997). *Evolutionary game theory*. Cambridge, MA: MIT press.

Weintraub, R. (1991). *Stabilizing dynamics: constructing economic knowledge*. Cambridge, MA: Cambridge University Press.

Weintraub, R., (1992). Introduction. In *Toward a history of game theory*, Weintraub, E. R. (ed.), Durham: Duke University Press, 3.

Weintraub, R., (2005). Autobiographical memory and the historiography of economics. *Journal of the history of economic thought*, 27(1), 1–11.

Weirich P (2016) Causal Decision Theory. In: Zalta NE (ed) The Stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/win2016/entries/decision-causal/

Wertheimer, M. (1923). Untersuchungen zur lehre von der Gestalt II. *Psycologische Forschung*, 4, 301-350. Translated in *A source book of Gestalt psychology*, Ellis, W. (ed.) (1938) London: Routledge & Kegan Paul, 71-88.

White, H.C., 1995. Social networks can resolve actor paradoxes in economics and in psychology. Journal of institutional and theoretical economics, 151 (1), 58–74.

Wimmer, H., Hogrefe, G., Perner, J. (1988). Children's understanding of informational access as a source of knowledge. *Child Development*, 59, 386–396.

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron*, 40(3), 655–664.

Wohlstetter, A. and Coleman, S., 1970. Race difference in incomes. R-578-OEO. Santa Monica, CA: RAND Corporation.

Young, A. W., Humphreys, G. W., Riddoch, M. J., Hellawell, D. J, de Haans, E. H. F. (1994). Recognition impairments and face imagery. *Neuropsychologia*, 32, 693–702.

Zahavi, D. (2001). Beyond empathy: phenomenological approaches to intersubjectivity. Journal of Consciousness Studies 8(5–7), 151–167.

Zahavi, D. (2011). Empathy and direct social perception: A phenomenological proposal. Review of Philosophy and Psychology, 2(3), 541-558.

Zahavi, D. (2014). *Self and other. Exploring subjectivity, empathy, and shame*. Oxford: Oxford University Press.

Zaitchik, D. (1991). Is only seeing really believing? Sources of the true belief in the false belief task. *Cognitive Development*, 6, 91–103.

Zawidzki, T. W. (2013). Mindshaping: A new framework for understanding human social cognition. MIT Press.

Zawidzki, T. W. (2018). Mindshaping. In *The Oxford Handbook of 4E Cognition*.

Zeckhauser, R. (1989). Distinguished fellow: Reflections of Thomas Schelling. Journal of economic perspectives, 3(2), 153-164.

Zizzo D. J., Tan J. H. (2007). Perceived harmony, similarity and cooperation in 2 × 2 games: an experimental study. *Journal of Economic Psychology*, 28(3), 365-386.

Zizzo D. J., Tan J. H. (2011). Game harmony: a behavioral approach to predicting cooperation in games. *American Behavioral Scientist*, 55(8), 987-1013.

Résumé de la thèse

L'objectif de la thèse est d'examiner les conditions à partir desquelles les joueurs peuvent définir une « solution » à un jeu non coopératif – c'est-à-dire les conditions à partir desquelles un équilibre existe, d'identifier le processus de raisonnement qui conduit les joueurs à cette solution, et de déterminer comment ces joueurs convergent vers l'identification de la même solution.

La contribution de notre thèse au domaine de la théorie des jeux est principalement un changement épistémologique et ontologique en relation à la fois au contenu de la théorie des jeux et au concept de jeu. La théorie des jeux est comprise dans cette thèse comme un cadre d'analyse, dont le cœur est l'analyse des interactions stratégiques entre les agents actifs et non une simple théorie mathématique décrivant les choix individuels à l'équilibre. Les jeux sont considérés comme des situations d'interdépendance stratégique qui impliquent un processus de raisonnement vers une solution et non une simple représentation mathématique d'un choix individuel.

La façon dont sont définis un équilibre et une solution dans l'approche mathématique des jeux – en théorie des jeux classique et épistémique – suppose l'existence d'une solution. Cette approche des jeux se concentre sur les conditions mathématiques de l'existence d'une solution sans aucune explication possible du processus conduisant à cette solution. L'existence de l'équilibre est supposée mais non expliquée. Un jeu n'est pas conçu comme un processus mais comme une simple représentation mathématique d'un choix stratégique.

La thèse propose donc d'examiner les conditions à partir desquelles une solution peut émerger. Cela nécessite d'étudier la capacité de coordination des joueurs entendue comme le processus de convergence des intentions et des croyances des joueurs. Nous affirmons que la compréhension du processus de coordination permet d'appréhender la nature d'un raisonnement stratégique et d'apporter de nouvelles réponses au problème d'indétermination de la théorie des jeux qui est l'une des impasses auxquelles la théorie des jeux est confrontée et qui souligne ses difficultés positives et normatives.

Si la théorie des jeux comportementale a étendu l'analyse et la formalisation de certains des mécanismes conduisant les acteurs à coopérer, et a considérablement amélioré notre compréhension de la coopération à partir de données empiriques, l'étude de la coordination reste cependant sous-développée. Certaines des propositions pour résoudre la difficulté d'expliquer et de rationaliser la coordination en théorie des jeux, comme les fameux points focaux, sont bien connues, néanmoins leur utilisation et leur formalisation ont fait assez peu de progrès. Qui plus est, nous soutenons dans cette thèse que la coordination est un phénomène plus général que celui décrit dans les jeux de coordination tels que les jeux du « Hi-Lo » ou du « Stag-Hunt » dans lesquels deux équilibres existent. Comme cela sera exposé dans le deuxième chapitre, la coordination est inhérente à chaque type de jeu, à chaque type d'interaction stratégique ; même aux situations pouvant présenter un conflit d'intérêts. C'est pourquoi la question de la coordination est d'une telle importance. Si la coordination est comprise comme la convergence des perceptions et des croyances des joueurs, de sorte que leurs actions soient finalement

cohérentes, la psychologie des joueurs et donc la manière dont ils sont modélisés dans les jeux prend à nouveau une importance particulière.

Examiner les conditions dans lesquelles les croyances des joueurs peuvent converger nécessite de se concentrer sur le raisonnement des joueurs, c'est-à-dire de comprendre le raisonnement stratégique comme un véritable processus de raisonnement dans lequel les joueurs doivent s'ajuster mutuellement. Il faut donc, contrairement à ce qui se fait en théorie des jeux, intégrer les «états mentaux» des joueurs dans les jeux. Les états mentaux des joueurs se réfèrent plus spécifiquement à leurs croyances sur les choix et croyances des autres joueurs, leurs intentions, ou leurs perceptions. L'intérêt de la thèse est de rassembler des contributions qui mettent l'accent sur la dimension intersubjective des jeux, c'est-à-dire la dimension dans le raisonnement des joueurs inhérente à l'appréhension des choix, croyances ou intentions des autres joueurs ; et de rassembler des contributions provenant à la fois de l'intérieur et de l'extérieur de la théorie des jeux, comme des sciences cognitives, de la philosophie de l'esprit et de la sociologie. La dimension intersubjective dans le processus de raisonnement du joueur est insuffisamment évaluée en théorie des jeux du fait d'une approche très particulière de la rationalité individuelle et des croyances individuelles. Notre contribution est basée sur une approche interdisciplinaire qui suggère un changement d'épistémologie et pas seulement un changement de méthodologie pour surmonter cette lacune. Il s'agit en effet de penser une nouvelle forme d'intersubjectivité en théorie des jeux.

Comme l'explique Giocoli (2003), la sous-détermination du processus conduisant à une solution pour un jeu est liée à une conception spécifique de l'équilibre et une conception spécifique de l'économie en tant que science. A ces deux types d'équilibre correspondent également deux conceptions spécifiques de la notion de rationalité individuelle. Les conséquences méthodologiques et plus généralement épistémologiques de ses différentes conceptions sont importantes pour la théorie des jeux, comme nous le verrons dans la thèse.

Ces deux conceptions de l'équilibre sont successivement : i) l'équilibre comme « un attracteur de mouvements arbitraires d'un processus dynamique sous-jacent » et ii) l'équilibre comme « un état de non mouvement » (Weintraub, 1991, p. 18). Giocoli (2003, p. 138 ; se référant à Weintraub, 1991, p. 102) ajoute que le premier type d'équilibre est « associé à l'image mécanique de la réalisation d'un équilibre des forces ... cela nécessite l'existence d'un processus d'équilibrage, en vertu duquel l'équilibre est effectivement atteint » tandis que le second type d'équilibre « est caractérisé par la réalisation d'un ensemble de conditions statiques ; il n'y a pas de mécanisme par lequel l'équilibre est établi ». La théorie des jeux classique et la théorie des jeux épistémique s'inscrivent toutes deux dans cette deuxième conception de l'équilibre.

Dans ce cas de figure, l'économie est fondée sur un système de relations qui est défini par Giocoli (2003, p. 139) comme « une condition de cohérence mutuelle entre un ensemble de relations économiques » et dans laquelle l'existence de l'équilibre et les propriétés de l'équilibre sont au centre de l'analyse. Les choix des agents économiques sont cohérents, « en harmonie ». Ainsi, il n'existe pas de processus hors équilibre, il n'y a pas de recherche des forces qui poussent le système économique vers l'équilibre : l'analyse se concentre sur l'existence de l'équilibre et les propriétés de cet équilibre (comme l'optimalité par exemple). L'analyse économique est basée sur une analyse statique de l'équilibre : sur une simple représentation de cet équilibre. Cette

conception de l'économie est fondée sur une vision de la rationalité en terme de cohérence des choix individuels, selon laquelle le plan de chaque agent économique doit être conforme, c'est-à-dire cohérent, à celui de l'autre. Mais là encore, « le prix à payer … est l'impossibilité d'expliquer ou de justifier comment et pourquoi l'équilibre se produit en premier lieu » (Giocoli, 2003, p. 208). Nous proposons un changement épistémologique pour remettre au centre de l'analyse des jeux ce double questionnement : « comment » et « pourquoi » un équilibre survient.

Les théoriciens des jeux modernes se sont donc concentrés sur la recherche des conditions mathématiques assurant l'existence, l'unicité et la stabilité des solutions. Ils ont essentiellement développé et affiné les outils mathématiques de la théorie des jeux afin de proposer des solutions définies pour les jeux (voir par exemple Schelling, 1960 ; Bacharach et Hurley, 1991 ; Bacharach, 1986, 2006 ; Hausman, 2000 ; Grüne-Yanoff et Lehtinen, 2012). Par exemple, la théorie des jeux épistémiques - la version contemporaine de la théorie des jeux non-coopératifs à information incomplète (Péréa, 2014) - fournit les conditions épistémiques des joueurs qui sont compatibles avec des concepts de solution prédéfinis. Cependant, comme on le verra, la théorie des jeux épistémique s'appuie sur une définition très spécifique des croyances des joueurs. Elle ne fournit pas les outils permettant d'expliquer d'où viennent les croyances des joueurs et pourquoi elles peuvent converger vers une solution. La théorie des jeux épistémiques se contente de décrire les croyances et choix des joueurs compatibles avec l'équilibre, avec le concept de solution défini a priori.

La thèse se fonde sur l'argument selon lequel la compréhension du processus de raisonnement des joueurs dans les jeux nécessite avant tout d'expliquer comment les joueurs forment leurs croyances sur les choix, les perceptions et les croyances des autres joueurs et leur raisonnement dans un contexte stratégique. L'un des objectifs de la thèse est de montrer qu'une théorie psychologique expliquant la formation des croyances des joueurs est nécessaire pour rendre compte de la coordination, et que la théorie de l'esprit (ToM) offre un cadre analytique adéquat. Comme nous l'expliquerons dans le chapitre 1, l'un des principaux problèmes de la théorie des jeux, et en particulier de la théorie des jeux épistémiques, est une ontologie controversée des croyances des joueurs. Rien n'explique comment les joueurs forment leurs croyances concernant les choix, les croyances, les perceptions ou le processus de raisonnement des autres joueurs, puisque ces croyances sont supposées, avant le jeu, être déjà le résultat d'un processus de raisonnement rationnel, à la fin duquel ces croyances individuelles sont mutuellement cohérentes. Ces croyances traduisent déjà l'idée que les choix des joueurs sont rationnels et donc à l'équilibre. Le fondement de la théorie des jeux épistémiques étant la théorie bayésienne de la décision, les croyances et les choix des joueurs ne sont en fin de compte que la représentation d'une décision respectant les axiomes de la rationalité de Bayes. La théorie des jeux épistémique représente les choix et les croyances des joueurs et ne décrit pas une décision ou un processus de formation de croyances (Binmore, 1993 ; Heidl, 2016 ; Hausman, 2012, 2000).

La théorie des jeux standard (classique et épistémique) est restée ancrée à la vision "néoclassique" de l'économie, dans laquelle les préférences individuelles sont données a priori, de sorte que l'économie est libre de tout déterminant psychologique. Notre contribution vise à montrer que la réconciliation de l'économie et de la psychologie cognitive et plus généralement des sciences cognitives est la seule façon d'éviter cette impasse.

La thèse est organisée en cinq chapitres comme suit :

Le premier chapitre décrit les évolutions de la théorie des jeux standard depuis sa création, avec la théorie des jeux classique, jusqu'à sa version moderne avec la théorie des jeux épistémique et analyse l'idée de « concept de solution » et son impact sur la façon dont les joueurs sont représentés, leur rationalité et le processus de résolution du jeu. Ce chapitre vise à mettre en évidence les écueils liés à la conception mathématique de la théorie des jeux tels que le problème d'indétermination et l'incapacité d'expliquer pourquoi et comment les joueurs se coordonnent, s'accordent, sur un même équilibre et une même solution.

Le deuxième chapitre relate la pensée ontologique de Schelling sur la coordination et sa vision d'un jeu comme un processus interactif dans lequel le processus de raisonnement des joueurs explique la solution du jeu. Nous mettons en évidence que Schelling voit les jeux comme des situations de forte interdépendance. Selon lui, la solution du jeu est principalement déterminée par la manière dont les joueurs réagissent les uns aux autres; c'est un processus de découverte.

Le troisième chapitre présente les travaux de Bacharach sur théorie des jeux et principalement deux de ces théories : la Variable Frame Theory (VFT) et le Team Reasoning. Pour comprendre la coordination et expliquer la capacité de coordination des joueurs dans les jeux, il incorpore dans les jeux, à travers la VFT, la perception que les joueurs ont des jeux et de la situation stratégique à laquelle ils sont confrontés. Il ouvre un programme de recherche en théorie des jeux sur le cadrage qui est selon nous une première étape pour réévaluer le type d'intersubjectivité impliquée dans la théorie des jeux non coopérative.

Le quatrième chapitre examine l'apport de la théorie de l'esprit et plus spécifiquement de la théorie de la simulation pour la théorie des jeux. Cela permet d'amener en théorie des jeux une réflexion sur la nature et le type d'intersubjectivité inhérents à un cadre d'interaction stratégique. Nous défendons l'idée selon laquelle la théorie de la simulation fournit une base méthodologique pour expliquer comment les joueurs forment leurs croyances sur les croyances, le raisonnement, les perceptions, les intentions ou les comportements des autres. Elle peut donc fournir un cadre analytique approprié pour intégrer les états mentaux et le processus de raisonnement des joueurs et pour expliquer la coordination.

Nous suggérons enfin au cinquième chapitre de construire une théorie alternative des jeux reposant sur la Variable Frame Theory de Bacharach et sur la théorie de la simulation. Nous définissons ensuite une axiomatique des choix rationnels dans les jeux en présence de joueurs capables de simuler le raisonnement d'autrui.