

**ESSAYS ON SOCIAL CONFORMITY:
BEHAVIOURAL GAME THEORY MODELS AND EXPERIMENT**

by

Alessandro Sontuoso

A thesis submitted in partial satisfaction of the requirements
for the degree of Doctor of Philosophy in Economics

Royal Holloway, University of London

July 2012

Declaration of Authorship

This thesis is my own composition, all sources have been acknowledged and my contribution is clearly identified in the thesis.

Alessandro Sontuoso

London, 23 July 2012

ai miei genitori

Essays on Social Conformity: Behavioural Game Theory Models and Experiment

Alessandro Sontuoso

Abstract

Human conduct is often guided by conformist preferences, with “conformity” being the act of changing one’s behaviour to match the purported beliefs of others. Informal norms regulating human behaviour play a crucial role in directing people’s expectations, thereby favouring uniformity of behaviour. This thesis develops such insights by exploring the conditions for different categories of norms to be in operation. The first essay [Chapter1] considers the motive that drives players when facing a problem of coordinating one another’s actions for their mutual benefit. Chapter 1 suggests that for a “convention” (i.e.: a solution to a coordination game with multiple equilibria) to be in operation, conformity is dependent on the states one is aware of, that is, the specifications of the contingencies that each player perceives in the context of a given game. The second essay [Chapter2] focuses on the motivation that makes people comply with default rules of behaviour when facing a social dilemma (i.e.: a “mixed-motive” game). Chapter 2 suggests that individuals may feel guilt at violating a norm, and this painful emotion generates conformity under precisely stated conditions. The essay models a “norm” as a rule that dictates a set of strategy profiles: it is assumed that players hold a conjecture about the active player’s norm-complying actions; a norm-driven decision maker is then modelled as a player with conformist preferences whose utility function is a linear combination of material and psychological payoffs. The third essay [Chapter3] provides an experimental test for conformist motivations by investigating the extent to which the peers’ behaviour (as presumed by other players) serves the individual as a means to guiding her actions. Specifically, the experiment of Chapter 3 is designed to measure the impact of the beliefs of players in the same role on behaviour; the data show evidence of conformity being present.

Table of Contents

ACKNOWLEDGMENTS	- 9 -
INTRODUCTION.....	- 10 -
I. A THEORY OF CONFORMITY TO CONVENTIONS WITH INCOMPLETELY-AWARE PLAYERS..	- 14 -
I.1. INTRODUCTION	- 15 -
I.2. PRELIMINARIES	- 17 -
<i>I.2.a. Notation on strategic form games</i>	<i>- 17 -</i>
<i>I.2.b. Coordination games, symmetries, and labels.....</i>	<i>- 19 -</i>
<i>I.2.c. Knowledge and unawareness.....</i>	<i>- 23 -</i>
I.3. A GAME AND A COMPARISON OF ALTERNATIVE ANALYSES	- 27 -
I.4. A MODEL OF CONVENTIONS.....	- 32 -
<i>I.4.a. A general framework for perception</i>	<i>- 35 -</i>
<i>I.4.b. From perception to labelling</i>	<i>- 45 -</i>
<i>I.4.c. From labelling to salience comparison</i>	<i>- 52 -</i>
<i>I.4.d. Expected utility maximization</i>	<i>- 57 -</i>
I.5. CONVENTIONS AS EQUILIBRIA	- 68 -
I.6. CONCLUDING REMARKS	- 73 -
II. A THEORY OF BELIEF-DEPENDENT CONFORMITY TO SOCIAL NORMS.....	- 76 -
II.1. INTRODUCTION.....	- 77 -
II.2. BICCHIERI'S ACCOUNT OF NORMS.....	- 84 -
II.3. PRELIMINARIES.....	- 88 -
<i>II.3.a. Notation on extensive form games</i>	<i>- 88 -</i>
<i>II.3.b. Conditional systems of beliefs.....</i>	<i>- 90 -</i>
II.4. A MODEL OF SOCIAL NORMS.....	- 92 -
<i>II.4.a. Norms and perfectly norm-driven decision makers</i>	<i>- 92 -</i>
<i>II.4.b. Belief-dependent conformist preferences</i>	<i>- 98 -</i>
<i>II.4.c. Social norms</i>	<i>- 104 -</i>
II.5. EQUILIBRIUM CONCEPT	- 110 -

II.6. ILLUSTRATIONS	- 115 -
II.6.a. <i>Trust Games</i>	- 117 -
II.7. CLOSING REMARKS.....	- 123 -
II.8. APPENDIX II	- 126 -
II.8.a. <i>Proofs</i>	- 126 -
II.8.b. <i>A review of alternative theories of norm compliance</i>	- 128 -
III. A TEST FOR CONFORMIST MOTIVATIONS IN EXPERIMENTAL GAMES	- 133 -
III.1. INTRODUCTION	- 134 -
III.2. TESTS FOR CONFORMITY AND RELATED BELIEF-DEPENDENT MOTIVATIONS	- 136 -
III.3. EXPERIMENTAL DESIGN	- 141 -
III.3.a. <i>False consensus effects vs. conformity to social norms</i>	- 141 -
III.3.b. <i>Game specification and treatments structure</i>	- 145 -
III.3.c. <i>Hypotheses</i>	- 149 -
III.3.d. <i>Procedure</i>	- 151 -
III.4. RESULTS	- 153 -
III.4.a. <i>Analysis of treatments T0-T1</i>	- 153 -
III.4.b. <i>Analysis of treatment T2</i>	- 159 -
III.5. CONCLUDING REMARKS	- 166 -
III.6. APPENDIX III.....	- 168 -
III.6.a. <i>Additional data</i>	- 168 -
III.6.b. <i>Experimental instructions and screenshots</i>	- 170 -
CONCLUSIONS	- 181 -
REFERENCES	- 183 -

List of Figures

Figure I.1 - Some state spaces and projections in Choose an Object	- 40 -
Figure II.1 - Discrete Dictator Game " <i>DDG</i> "	- 108 -
Figure II.2 - Trust Game " <i>TG</i> "	- 117 -
Figure II.3 - Standardized Trust Game " <i>STG</i> "	- 120 -
Figure III.1 - Trust Game " <i>TG</i> "	- 145 -

List of Tables

Table III-1 - T0 summary statistics	- 154 -
Table III-2 - T1 summary statistics	- 154 -
Table III-3 - Mean values of belief variables	- 155 -
Table III-4 - T0 and T1 Probit regression coefficients.....	- 156 -
Table III-5 - T2 summary statistics	- 159 -
Table III-6 - T2 Probit regression coefficients	- 162 -
Table III-7 - T2 Part I Probit regression coefficients	- 163 -
Table III-8 - Beliefs transmitted in Part I of T2	- 168 -
Table III-9 - Beliefs transmitted in Part II of T2	- 169 -

Acknowledgments

I am grateful for the supervision provided by Dirk Engelmann, and for his brilliant advice and valuable support; in particular, I thank him for having generously offered to continue to provide feedback, especially on the first two essays, when he left Royal Holloway.

I owe particular debts to Michael Naef for supervising me during my work on the third essay, for contributing to the development of its experimental design, and for general support and guidance; also, I am grateful to Bjoern Hartig for managing the experimental lab and for programming the zTree code used in the experiment. I further acknowledge the financial support of the College.

Lastly, I thank the AC/DC for teaching me that «it's a long way to the top (if you wanna rock 'n' roll)».

Introduction

Surveys from various disciplines (including sociology, cognitive psychology and neuroscience) support the view that human conduct is often guided by conformist preferences – which thrive on *behavioural expectations* within a society or group – with “conformity” being the act of changing one’s behaviour to match the purported beliefs of others. *Informal norms* regulating human behaviour play a crucial role in directing people’s expectations, thereby favouring uniformity of behaviour within a given social group. By serving as equilibrium selection devices, norms reduce transaction costs in economic interactions that present multiple equilibria or, in some cases, promote efficient solutions. The present thesis develops these insights, in order to improve our understanding of such norms by explaining the underpinning conditions for different categories of norms to be in operation among players with conformist motivations. Indeed, the literature proposes various mechanisms for uniform social behaviour – which in turn relate to a variety of conformist preferences – including a pure coordination motive, social disapproval and the internalization of absolute norms of conduct.

The first essay [Chapter 1] considers the first mechanism or, precisely, the motive that drives players when facing a problem of coordinating one another’s actions for their mutual benefit. Chapter 1 suggests that for a “convention” (*i.e.*: a certain solution to a coordination game with multiple equilibria) to be in operation, conformity is dependent on the states one is aware of, that is, the specifications of the contingencies that each player perceives in the context of a given game (*e.g.*: contextual cues). The essay proposes a theoretical framework for the player’s own perception of the game so as to show that a convention is in place whenever members of a social

group use, and expect others to use, similar conceptual schemes: this is done by implementing a system of multiple state spaces ordered by expressive power, and a notion of the players' (un)awareness, in such a way as to provide a precise link between the players' perception of the game and the associated strategy labels. In brief, conventions are devised as the result of a four-step procedure: (i) perception; (ii) labelling; (iii) salience comparison; (iv) expected utility maximization.

The second essay [Chapter 2] focuses on the second mechanism or, precisely, the motivation that makes people comply with default rules of behaviour when facing a social dilemma (*i.e.*: a “mixed-motive” game¹). Chapter 2 suggests that individuals may feel guilt at violating a norm, and this painful emotion generates conformity under precisely stated conditions. The essay models a “norm” as a rule that dictates a set of strategy profiles: it is assumed that players, conditional on each history of an extensive form game, hold a conjecture about the active player's norm-complying actions available at that history; a norm-driven decision maker is then modelled as a player

¹ Following Thomas Schelling's ([1960], Ch. 4) classification of games, strategic interactions can be categorized as “pure motive” and “mixed motive” games. The former are situations in which the players' preferences are rank-correlated with respect to outcomes, as in the games of pure coordination (positive correlation) or in the games of pure conflict, also known as zero-sum games (negative correlation). On the other hand, mixed-motive games present a non correlated structure of preferences, due to their mix of coordination opportunities and conflicting motivations: as Schelling puts it, «“[m]ixed-motive” refers not, of course, to an individual's lack of clarity about his own preferences but rather to the ambivalence of his relation to the other player – the mixture of mutual dependence and conflict, of partnership and competition» (p. 89).

with conformist preferences, whose utility function is a linear combination of her material payoff and a component representing the social cost of deviating. A “social norm” is said to exist and to be followed by a population if players have conditionally conformist preferences, hold correct beliefs, and are sensitive enough to the social cost of deviating. (The aforementioned third mechanism for uniform social behaviour, that is, the case of absolute norms of conduct is here considered as the feature of a special family of norm-driven agents, *i.e.*: those with unconditional preferences for conformity to a norm, which therefore constitutes a “moral norm”.)

The third essay [Chapter 3] provides an experimental test for conformist motivations (in mixed-motive games) by investigating the extent to which the peers’ presumed behaviour serves the individual as a means to guiding her own actions. The first hypothesis is that the experimenter should be able to predict a conformist player’s behaviour from the conformist’s guess about the behaviour of other players in the same role. Now, it should be noted that a false consensus effect hypothesis will predict an analogous correlation between beliefs and behaviour (although with an inverse causal relationship); given that, in order to disentangle consensus from conformity, one of the experimental treatments of Chapter 3 introduces an exogenous variation in beliefs by showing subjects some aggregate information about the others’ beliefs. Indeed, if the experimenter can predict a subject’s choice from the subject’s guess (about the behaviour of other participants in the same role) in conjunction with the subsequently transmitted information about others’ guesses, then one has effectively disentangled consensus from conformity and provided evidence in support of a conformity hypothesis. In fact, if false consensus is present, then there will be a causal relationship from behaviour to beliefs, and thus there will not be an effect of providing exogenous information; but if, on the other hand, conformity is present (in which case the causality runs from beliefs to behaviour), one will find that

exogenously varying beliefs has an impact on behaviour. More specifically, the experiment measures the impact of the beliefs of players in the same role, on behaviour, in a Trust Game: in brief, the data show that the transmitted information can influence one's behaviour, with the strength of the impact depending on one's prior beliefs; therefore, the data show evidence of conformity being present.

Before proceeding, I shall stress that the rule-based approach to conformity here pursued does considerably differ from the approach followed in notable alternative accounts of conformity, such as the models of informational cascades (*e.g.*: Banerjee [1992], Bikhchandani *et al.* [1992]) or esteem-based models (*e.g.*: Bernheim [1994]), in a sense that – while those models simply assume an agent to care about the others' information or esteem – there the others' *actions* do not directly enter each agent's utility function. Conversely, the present thesis focuses on conformity in situations where an individual's payoff directly depends on what the others do; in fact, although the present definitions and conditions for conventions or social norms to apply differ from one another (*i.e.*: “conventions” apply to coordination games, while “social norms” apply to mixed-motive games), in both cases the essence of such unwritten rules is defined by the fact that both imply belief-based solutions to problems of strategic interdependence.

***I. A Theory of Conformity to Conventions with
Incompletely-Aware Players***

I.1. Introduction

The “theory of convention” found its first comprehensive formulation within David Hume's theory of justice, elaborated in his *Treatise of Human Nature* (Hume [1740]). Then, the first game-theoretic analysis of conventions was developed by David Lewis in his *Convention: A Philosophical Study* (Lewis [1969]): this defined conventions as behavioural regularities satisfying some special conditions and inducing a “pure coordination equilibrium” (*i.e.*: a regular pattern of behaviour that is a strict Nash equilibrium in a coordination game with two or more strict Nash equilibria).² Note that Lewis acknowledged his debt to Thomas Schelling – from whom he had borrowed the idea of modelling conventions as equilibria of coordination games – according to Schelling [1960] in fact, in solving coordination problems, we are often driven by apparently insignificant factors that make one of the feasible strategies “salient”. Also, it should be recalled that, according to the philosophical literature (Bicchieri [2006]), a convention is simply a descriptive norm which does not imply a commitment to compliance, but is useful since it coordinates people’s expectations by acting as a signal that eases interaction; as Ken Binmore [2007] points out, in the Driving Game nobody cares which convention we use,³ there is no reason why either of the equilibria should be preferred to the other, yet «[i]n practice, we solve many coordination games

² For recent reconstructions of Lewis’ philosophical theory of convention, see: Cubitt and Sugden [2003]; Sillari [2005].

³ The Driving Game is as follows: two players have to choose the side of the road upon which to drive: if they coordinate on either side, both get an equal payoff; if they do not coordinate, neither player receives anything.

by appealing to focal points that are determined by the context in which a game occurs. For example, people drive on the left in Japan and on the right in the United States. Such conventions are usually the result of historical accidents, but not always» (Binmore [2007], p. 268).

The starting point of this essay is that certain features of the game as experienced by the player – which would not explicitly enter the formal description of a standard game – can effectively make some strategy profiles *focal*. Therefore, the core of the problem is to provide a framework for the player's own perception of the strategic situation so as to show that coordination is possible because "normal" players use, and expect others to use, similar conceptual schemes. Thus, here it is suggested that, for a convention to be in operation, conformity is dependent on the states perceived by the agents, that is, the specifications of the contingencies that each player perceives in the context of a given decision problem (*e.g.*: contextual cues). In this respect, I shall build on Heifetz *et al.*'s [2006] model of unawareness so as to account for multiple descriptions of the world: as in their model, here a system of multiple, ordered state spaces and surjective projections (from each state space to every space that is weakly "less expressive") is adopted: the result is a powerful framework allowing for the players to be unaware of some features of the game as captured by alternative state spaces.

Then, building on Sugden's [1995] and Casajus' [2000] approaches, the present theory defines the players' own framing system in such a way as to allow for the possibility that both stochastic and non-stochastic procedures may determine the labelling of strategies. Yet, departing from Sugden's and Casajus' models, here each player's labelling of strategies depends directly on her perception of the game, that is, on the states she is aware of: therefore, this study provides a precise link between a player's information function and her labelled strategies. Also, this study departs from the existing

literature in that it introduces a binary relation (*i.e.*: a complete preordering) on each set of strategy labels, thereby allowing the salience comparison of pairs of labelled strategies. Further, the introduction of two requirements capturing the notions of symmetry and salience has the effect of restricting the set of a player's mixed strategies. It follows that conventions arise as the result of a four-step procedure: (*i*) perception; (*ii*) labelling; (*iii*) salience comparison; (*iv*) expected utility maximization.

Before proceeding, a quick note on methodology: this essay aims at explaining, in a static perspective, why conventions occur. While many commentators suggest alternative analyses of conventions (or, more generally, of coordination problems) based on evolutionary dynamics or on bounded rationality (*i.e.*: Level-*k* models), such approaches are not explored here as the present theory revolves around *one-shot games* played by (fully) rational utility-maximizers. In this respect, this model may be considered part of a body of literature sometimes referred to as "team-reasoning".

The remainder of the study is organized in this manner: I.2. introduces some general notation and concepts on strategic form games, coordination, and unawareness; I.3. proposes a game, and reviews some alternative analyses; I.4. formally expounds the model; I.5. discusses an equilibrium solution, and I.6. concludes.

I.2. Preliminaries

I.2.a. Notation on strategic form games

A strategic form game is formalized by a structure $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$, where: $N = \{1, \dots, n\}$ is the *set of players*, S_i is the *set of player i 's pure strategies*, u_i is *i 's payoff function*.

Each player i has a finite set S_i of pure strategies, with generic element $s_{i,a}$ (the first subscript indicating a certain player, the second subscript a certain strategy, with $a \in \{1, \dots, m\}$ for a given $m \in \mathbb{Z}^+$), where $s_{i,a} \in S_i = \{s_{i,1}, s_{i,2}, \dots, s_{i,m}\}$; in order not to make the notation overly cumbersome, in what follows I shall dispense with the comma that separates the two subscripts and denote a generic strategy-index simply by s_{ia} . Note that, throughout this essay, each of a player's pure strategies is assigned an index which uniquely identifies that strategy: this means that, for example, in the aforementioned Driving Game each player i has a finite set S_i of pure strategies with generic element s_{ia} , where $s_{ia} \in S_i = \{s_{i1}, s_{i2}\}$ for $\forall i \in N = \{1,2\}$; it should be further stressed that the English words "left" and "right" do not enter the formal structure of the game. Given that, a strategy profile s is a tuple of strategies, with one strategy for each player of the game: let $S = \prod_{i \in N} S_i$ be the *set of strategy profiles*; similarly define $S_{-i} = \prod_{j \neq i} S_j$ for players j other than i .

The *material payoffs* of players' strategies (as well as the players' *preferences*) are described by functions $u_i: S \rightarrow \mathbb{R}$, with $i \in N$; the payoffs to player i are therefore written as $u_i(s) \equiv u_i\left(\left(s_{ja}\right)_{j \in N}\right)$ for $\forall s_{ja} \in S_j$, where $s = \left(s_{ja}\right)_{j \in N}$ denotes a strategy profile.⁴ To keep the exposition simple, I shall often focus on 2-player games, although the analysis applies equally well to any n -player normal form game: in the 2-player case, it is common to put the strategy of Player 1 first so that the payoffs to player i , with $i \in N$ ($N = \{1,2\}$), are written as $u_i(s_{1a}, s_{2\bar{a}})$ for $\forall s_{1a} \in S_1, \forall s_{2\bar{a}} \in S_2$; with respect to the order

⁴ The lower subscript $j \in N$ indicates that s contains one element s_{ja} for every $j \in N$.

of the players' strategies when writing a strategy profile, the notation adopted here removes any ambiguity since every strategy-index belongs to exactly one player (e.g.: $(s_{1a}, s_{2\bar{a}})$ indicates the same strategy profile as $(s_{2\bar{a}}, s_{1a})$), hence the order of the players' actions does not matter since the strategies contained in each player's strategy set have different indices, that is, $S_i \cap S_j = \emptyset$ for $\forall j \neq i$.⁵

A *mixed strategy* for player i gives the probabilities that action $s_{ia} \in S_i$ will be played: let a generic mixed strategy for player i be denoted by σ_i , where $\sigma_i \in \Sigma_i \equiv \Delta(S_i)$, with Σ_i denoting the set of i 's mixed strategies and $\Delta(S_i)$ being the set of probability measures over S_i ; a generic mixed strategy for player i can be represented as a vector of probabilities $\sigma_i = (p(s_{i1}), p(s_{i2}), \dots, p(s_{im}))$. In the 2-player case, with a slight abuse of notation, let $u_i(\sigma_1, \sigma_2) = \sum_{s_{1a} \in S_1} \sum_{s_{2\bar{a}} \in S_2} p(s_{1a}) p(s_{2\bar{a}}) u_i(s_{1a}, s_{2\bar{a}})$ indicate the payoffs to player i for the profile of mixed strategies $\sigma = (\sigma_1, \sigma_2)$.

I.2.b. Coordination games, symmetries, and labels

Following in Lewis' [1969] wake – who defined a “convention” as a regular pattern of behaviour that is a strict Nash equilibrium in a coordination game with multiple strict Nash equilibria – the present theory revolves around one-

⁵ Note that the strategies contained in each player's strategy set have different indices because the first subscript of a strategy-index always identifies a certain player; for example in the aforementioned Driving Game, with $N = \{1,2\}$, the strategy sets are $S_1 = \{s_{11}, s_{12}\}$ and $S_2 = \{s_{21}, s_{22}\}$. Also notice that, given that only one-shot games in strategic form are to be analysed here, in this essay I refer to “strategies” and “actions” interchangeably.

shot games (played without communication) in which the payoff table is completely symmetrical between players and strategies.

Before defining symmetries, it may be worth recalling that a coordination problem is a situation with a number of outcomes on which agents can coordinate their actions for mutual benefit; such situations can be categorized as “pure coordination” or “impure coordination” games if – at each equilibrium outcome – all players receive the same payoff or not, respectively. Some pure coordination games, which present the below payoff structure for $\forall i \in N$, are referred to as *matching games*:

$$u_i(s_{1a}, s_{2\tilde{a}}, \dots, s_{n\check{a}}) = \begin{cases} 1 & \text{if } a = \tilde{a} = \dots = \check{a} \\ 0 & \text{otherwise} \end{cases}, \quad (1.2.1)$$

for all $a, \tilde{a}, \dots, \check{a} \in \{1, \dots, m\}$. Compactly, a matching game is a structure $\Gamma_n^m = \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$, where $i \in N = \{1, \dots, n\}$ and $s_{ia} \in S_i = \{s_{i1}, s_{i2}, \dots, s_{im}\}$ for $\forall i \in N$ (for given $n, m \in \mathbb{Z}^+$, $n = m$ or $n \neq m$), with $(u_i)_{i \in N}$ being defined as in formula (1.2.1):⁶ the aforementioned Driving Game is an instance of Γ_2^2 .

The concept of symmetry is developed by Harsanyi and Selten [1988], Ch. 3: here a slightly simpler definition that best suits the purposes of this study is employed.

⁶ Recall that, throughout this essay, it is assumed that material payoffs describe the consequences of the players' actions as well as their preferences.

Definition I.1. Given a strategic form game G , with $S_i \cap S_j = \emptyset$ for $\forall j \neq i$, a *symmetry* of the game is a pair of bijective functions $(\phi, (\zeta_i)_{i \in N})$, where $\phi: N \rightarrow N$ and $\zeta_i: S_i \rightarrow S_{\phi(i)}$ for $\forall i \in N$, such that $u_i(s) = u_{\phi(i)}\left(\left(\zeta_j(s_{ja})\right)_{j \in N}\right)$ for each player $i \in N$ and each strategy profile $s = (s_{ja})_{j \in N}$.

In a nutshell, a symmetry is a way of *exchanging the names* (i.e.: indices) of players and strategies that leaves the payoffs – hence the solution/s – of the game unchanged: obviously, if one exchanges via $(\phi, (\zeta_i)_{i \in N})$ the players and the strategies of a game and this yields in all cases the same payoffs (as those of each strategy profile prior to the exchange of names), then such a pair of bijective functions has generated the same game. Similarly, one could define a new game $G' = \langle N', (S'_i)_{i \in N'}, (u'_i)_{i \in N'} \rangle$ in such a way as to *rename ex novo* the players and the strategies but leave the payoffs as in the original game G : that is easily done by introducing a pair of bijective functions $(\dot{\phi}, (\dot{\zeta}_i)_{i \in N'})$, with $\dot{\phi}: N \rightarrow N'$ and $\dot{\zeta}_i: S_i \rightarrow S'_{\dot{\phi}(i)}$ for $\forall i \in N'$, such that $u_i(s) = u'_{\dot{\phi}(i)}\left(\left(\dot{\zeta}_j(s_{ja})\right)_{j \in N'}\right)$ for each player $i \in N'$ and each strategy profile $s = (s_{ja})_{j \in N'}$.⁷ Again, the solutions of G and G' are necessarily the same, which implies that the solutions of any game in strategic form have to be

⁷ In Harsanyi and Selten's [1988] terminology, such a pair of bijective functions is referred to as a *renaming* or, equivalently, as an *isomorphism with no positive linear payoff transformations*. Two games G and G' are called *isomorphic* if at least one isomorphism from G to G' exists.

independent of the ordering or naming of players/strategies: this result – known as “invariance with respect to isomorphisms” – is formalized by Harsanyi and Selten [1988], Ch. 3; further, since a strategy could be named (*i.e.*: indexed) differently, in equivalent games, invariance with respect to isomorphisms implies that strategies that are distinguishable only with respect to each strategy-index should be assigned the same probabilities in a solution.

Given that, Casajus [2000] suggests a definition of “symmetric strategies” which draws on Harsanyi and Selten’s notion of symmetry of a game: (adapting it to the notation and definition of symmetry of the present essay) two pure strategies $s_{i\alpha} \in S_i$ and $s_{i'\bar{\alpha}} \in S_{i'}$ of G are said to be *symmetric* if there exists a symmetry $(\phi, (\zeta_i)_{i \in N})$ of G such that $\zeta_i(s_{i\alpha}) = s_{i'\bar{\alpha}}$. For example, in the Driving Game Γ_2^2 , s_{11} and s_{12} are symmetrical with s_{21} and s_{22} , respectively; again, this implies that the only symmetry-invariant equilibrium in mixed strategies must assign them the same probability, *i.e.*: $\left(\left(\frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{1}{2}\right)\right)$.

Although neither Schelling [1960] nor Lewis [1969] defined any such formal concept of symmetry of strategic forms, both pioneered the study of coordination in games whose payoff table is symmetrical between players and strategies. In particular, Schelling drew attention to the importance of contextual cues (or “focal points”) in coordination problems: in effect, players can sometimes solve coordination games by resorting to apparently insignificant factors that make one of the feasible actions salient, thereby breaking any symmetry of strategies; put differently, this means that the context in which games appear – or the way games are framed – may affect the way people play them. Building on Schelling’s intuition, Sugden [1995] enriches the mathematical structure of a game by introducing a rule that assigns to each of a player’s strategies a private label representing the way

the very player describes the game to herself (*e.g.*: in Γ_2^2 one player may have a rule such that $s_{i1} \mapsto \textit{left}$ and $s_{i2} \mapsto \textit{right}$, or *vice versa*); there the players' descriptions of strategies are privately observed permutations of the analyst's naming (*i.e.*: indexing) of strategies. As Sugden points out:

The labels that a player uses will depend in part on psychological and cultural factors: for example, where one player sees "left and right", another might see "east and west". There is a sense in which the labels that players use reflect what is salient for them: we might say that the left-right distinction is salient for some players and the east-west distinction for others. (p. 537)

While Sugden's [1995] theory accounts for coordination in games where a *stochastic* procedure determines the (one-dimensional) label each player privately assigns to each of her strategies, Casajus [2000] provides a complementary framework for the analysis of coordination games where a non-stochastic structure called "frame" describes the players' own representation (*i.e.*: labelling) of strategies by associating them with a set of attributes (*e.g.*: colour or shape of an object to choose). More precisely, Casajus builds on Bacharach [1993] and Janssen [2001] by defining a *multi-dimensional* system for the labelling of strategies,⁸ in addition to a requirement on solutions based on Harsanyi and Selten's invariance with respect to isomorphisms.

I.2.c. Knowledge and unawareness

The present theory compromises on Sugden's and Casajus' approaches, as it defines the players' own framing system in such a way as to allow for the

⁸ This body of literature was initiated by Gauthier [1975].

possibility that both stochastic and non-stochastic procedures may determine the labelling of strategies. Moreover, this study postulates that different labellings arise among players by virtue of the different sets of states each player may be aware of (where the multiple state spaces picture the specifications of the contingencies each player perceives in the context of a given decision problem). More precisely, here, *each player's labelling of strategies depends directly on her perception of the game, that is, on the states she is aware of*: therefore, this study provides a precise link between the players' perception of the game (given by their information functions) and their labelled strategies.

Before introducing the concept of (un)awareness, I shall briefly present the standard model of knowledge and the problems associated with it.⁹ For a given strategic form game $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$, a model of knowledge is a structure $\langle (\Omega, q), (\mathcal{S}_i)_{i \in N} \rangle$, where: Ω is the *set of states*, q is a probability measure over Ω , \mathcal{S}_i is the *information partition* of player i (with \mathcal{S}_i being a partition of Ω). Given the state space Ω – where each $\omega \in \Omega$ is a description of the contingencies that the agent considers to be relevant in the context of the game at hand – each player i has an *information (set-valued) function* I_i that associates with every state $\omega \in \Omega$ a non-empty subset $I_i(\omega) \subseteq \Omega$ (with $I_i(\omega)$ being interpreted as the set of states the agent considers possible when the true state is ω). It is assumed that $\langle (\Omega, q), (\mathcal{S}_i)_{i \in N} \rangle$ satisfies the

⁹ The standard model of knowledge corresponds to the S-5 system of epistemic logic, and is due to Hintikka [1962]. The concept of *common knowledge* was suggested by Lewis [1969], whereas Aumann [1976] gave the mathematically-precise (set-theoretic) definition which is habitually used in the economics literature.

following properties: (i) $\omega \in I_i(\omega)$ for $\forall \omega \in \Omega$; (ii) $\omega' \in I_i(\omega) \implies I_i(\omega') = I_i(\omega)$. To sum up, when representing a socio-economic application in game-theoretic terms, given a (unique) state space Ω , uncertainty is generically captured by assuming that the agent i does not know which is the true state, knowing instead only which cell $I_i(\omega)$ of a partition \mathcal{S}_i of Ω contains the true state ω .

Now, *unawareness* characterizes an epistemic state in which “one does not know an event,¹⁰ and does not know that she does not know it, and so on *ad infinitum*”.¹¹ While a notion of unawareness as described here encompasses the perfectly realistic situation in which an agent is simply ignorant as to the existence of some contingency, it turns out that – under the assumptions of the standard partitional model of knowledge – it is not possible to account for such a situation. In fact, the information function I_i associates with every state $\omega \in \Omega$ a non-empty subset $I_i(\omega)$ of Ω , and this implies that the agent cannot be unaware of anything: having an information partition \mathcal{S}_i entails that, if she does not know an event, then she knows she does not know it. To further clarify this point, I shall recall that player i is said to know event E (with $E \subseteq \Omega$), at state ω , if $I_i(\omega) \subseteq E$; writing $\mathcal{K}_i(E)$ as a shorthand for the set of states in which i knows E , one can define a *knowledge function* \mathcal{K}_i (mapping the power set of Ω into itself) by $\mathcal{K}_i(E) =$

¹⁰ As is customary, an event E is defined as a subset of the state space.

¹¹ In a controversial statement to the press Donald Rumsfeld (as the United States Secretary of Defense) claimed: «[T]here are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know».

$\{\omega \in \Omega: I_i(\omega) \subseteq E\}$. It is well known that the knowledge function \mathfrak{K} which is derived from an information function I satisfies the following properties:

$$(k.i: \text{necessitation}) \quad \mathfrak{K}(\Omega) = \Omega$$

$$(k.ii: \text{monotonicity}) \quad E \subseteq F \Rightarrow \mathfrak{K}(E) \subseteq \mathfrak{K}(F)$$

$$(k.iii: \text{conjunction}) \quad \mathfrak{K}(E) \cap \mathfrak{K}(F) = \mathfrak{K}(E \cap F)$$

$$(k.iv: \text{axiom of knowledge}) \quad \mathfrak{K}(E) \subseteq E$$

$$(k.v: \text{axiom of transparency}) \quad \mathfrak{K}(E) \subseteq \mathfrak{K}(\mathfrak{K}(E))$$

$$(k.vi: \text{axiom of wisdom}) \quad \sim\mathfrak{K}(E) \subseteq \mathfrak{K} \sim \mathfrak{K}(E).^{12}$$

A few comments are in order. Properties *(k.i-iii)* are bookkeeping assumptions satisfied by knowledge functions derived from any information function; besides, if the information function is partitional, then *(k.iv-vi)* are satisfied as well. Of particular interest is *(k.vi)*, which implies that whenever an agent does not know an event, she knows she does not know it (which plainly eliminates the possibility of unawareness): for this reason, Geanakoplos [1989] circumvents the problem by using non-partitional information structures; yet, Dekel *et al.* [1998] propose three intuitive properties for unawareness and show that they are not compatible at all with the standard state space specification (and in particular with the above *(k.i-ii)*). As a consequence of Dekel *et al.*'s [1998] impossibility results, a few attempts at modelling multi-person unawareness have been put forward either by making use of the modal syntax within the semantic structures or by

¹² \sim denotes negation.

means of set-theoretic specifications involving a lattice structure:¹³ the latter approach models a system of multiple state spaces explicitly ordered by “expressive power”.

In what follows I shall draw on Heifetz *et al.*'s [2006] specification of such epistemic states, in order to allow for the players to be unaware of some representations of the situation (as captured by different state spaces).¹⁴ Thus, the goal of the essay is to provide a framework for the player's own comprehension of the game so as to show that coordination is possible when certain players use – and expect others to use – similar conceptual schemes: so, for a convention to be in operation, conformity is dependent on the states (*e.g.*: contextual cues) perceived by similar players (*e.g.*: agents sharing a collection of attitudes, values, goals, and practices characterizing a certain group, organization or institution).

I.3. A game and a comparison of alternative analyses

I shall introduce the following game (henceforth “Choose an Object”). Players are presented with a tray with six cubic objects, which are then placed into a

¹³ For the former class of models see Heifetz *et al.* [2008], for the latter see Heifetz *et al.* [2006], among the others.

¹⁴ A related idea – known as “indirect realism” – has been popular in the history of philosophy, being developed by many authors including Bertrand Russell, Baruch Spinoza, René Descartes, and John Locke: indirect realism is a position broadly comparable to a certain view of perception in natural science, according to which individuals do not experience the external world as it really is, but know only interpretations of the way the world is (Hawking and Mlodinow [2010], Ch. 3).

bag all at once.¹⁵ (Such objects are cubic blocks, which the experimenter identifies by the numbers 1 to 6; such numbers are invisible to the players.) The colours of the cubes are as follows: objects no. 1-2 are grey, objects no. 3-4 are red, objects no. 5-6 are black; the objects are otherwise identical in shape, size, material, *etc.*. Given that, the experimenter takes *three blocks* out of the bag, one by one at random in front of all players, and places them on a table in an orderly fashion. Players are then privately asked to *choose one of the objects* (subjects cannot communicate with one another); the players' utility is defined as in formula (1.2.1) above, that is, assuming $|N| = 2$, each player's payoff is 1 if both choose the same object, 0 otherwise.

Using the notation introduced in section I.2.a. above, the analyst's description of the game is as follows: each object represents a distinct strategy, that is, $s_{ia} \in \{s_{i1}, s_{i2}, \dots, s_{i6}\}$ for $\forall i \in N$; notice that no contextual cues enter the analyst's description of the game, in fact – to the analyst – Choose an Object simply consists of $\binom{6}{3} = \frac{6!}{3!3!} = 20$ strategic games, where Nature randomly (and publicly) determines the one game to be played; so, each of the 20 games differs from the others only in the names of the available strategies, with the strategies of each of the 20 strategic games being denoted by $S_i' = \{s_{i1}, s_{i2}, s_{i3}\}$, $S_i'' = \{s_{i1}, s_{i2}, s_{i4}\}$, $S_i''' = \{s_{i1}, s_{i2}, s_{i5}\}$, $S_i'''' = \{s_{i1}, s_{i2}, s_{i6}\}$, *etc.* for $\forall i \in N$. For instance, s_{i1} denotes the strategy of choosing object no. 1, s_{i2} denotes the strategy of choosing object no. 2, *etc.* for $\forall i \in N$; again, it should be stressed that $\{s_{i1}, s_{i2}, s_{i3}\}$ represents only the

¹⁵ The initial positions on the tray are not clearly identifiable.

analyst's naming of the strategies (all such strategy-indices are invisible to the players, as are the numbers on the objects which the analyst uses to identify the blocks).

Now, assume Nature has selected the strategic game of which the set of strategies is $S'_i = \{s_{i1}, s_{i2}, s_{i3}\}$ for $\forall i \in N$: further to the discussion in section I.2.b. above, notice that s_{i1}, s_{i2}, s_{i3} are symmetric strategies; it follows that the only symmetry-invariant equilibrium in mixed strategies must assign them the same probability. Therefore, the only symmetry-invariant equilibrium is the profile of mixed strategies $\sigma = (\sigma_1, \sigma_2)$ where, using the above notation, a mixed strategy is given by the vector of probabilities $\sigma_i = (p(s_{i1}), p(s_{i2}), p(s_{i3})) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for $\forall i \in N$. Similarly, if Nature selects the strategic game with set of strategies $S''_i = \{s_{i1}, s_{i2}, s_{i4}\}$ for $\forall i \in N$, the only symmetry-invariant strategy profile is $\sigma = (\sigma_1, \sigma_2)$, where a mixed strategy is given by the vector of probabilities $\sigma_i = (p(s_{i1}), p(s_{i2}), p(s_{i4})) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for $\forall i \in N$.

From the above analysis it is evident that all such 20 strategic games are isomorphic (see footnote 7), that is, to the analyst their mathematical structures represent exactly the same decision problem, and as such should not be treated differently (Harsanyi and Selten [1988], Ch. 3). Instead, Variable Frame Theory (Bacharach [1993], Bacharach and Bernasconi [1997]) would analyze Choose an Object as follows: once Nature has selected a strategic game, the cubic blocks will present (colour-perceiving) players with either two or three different colours, hence two/three different "feasible acts" (besides a completely randomized act). Therefore, in the case in which the blocks are of only two colours, Variable Frame Theory (along with Janssen [2001]) suggests that players should choose the one block of a different colour: for instance, assume Nature has selected the strategic game with set of strategies $S'_i = \{s_{i1}, s_{i2}, s_{i3}\}$ for $\forall i \in N$; recalling that objects no. 1-

2 are grey and object no. 3 is red, Bacharach would predict that players choose the red object (*i.e.*: s_{i3} for $\forall i \in N$) with probability one.

Furthermore, consider the case in which Nature has selected the strategic game with set of strategies $\{s_{i1}, s_{i3}, s_{i5}\}$ for $\forall i \in N$; here the cubic blocks present players with three different colours. Thus, in this case Bacharach's theory would predict three equilibria in pure strategies, that is, one in which all players choose the grey object with probability one (*i.e.*: $s = (s_{i1})_{i \in N}$), one in which all players choose the red object (*i.e.*: $s = (s_{i3})_{i \in N}$), and one in which all players choose the black object (*i.e.*: $s = (s_{i5})_{i \in N}$). Unlike Bacharach, in this case Janssen [2001] argues that to the players the feasible acts (*i.e.*: the coloured objects) are symmetric, hence – players have no particular reason to choose their part of any one of the three solutions, and so – a theory of rational play should require that players implement the strategy profile $\sigma = (\sigma_1, \sigma_2)$, with each strategy being the vector of probabilities $\sigma_i = (p(s_{i1}), p(s_{i3}), p(s_{i5})) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

As mentioned above, Casajus' [2000] analysis follows in Bacharach's and Janssen's wake by defining a (multi-dimensional) system for the labelling of strategies in such a way as to formalize both (Bacharach's and Janssen's) arguments and take care of Harsanyi and Selten's invariance with respect to isomorphisms.¹⁶ Yet, specifically in the last example (*i.e.*: strategic game with set of strategies $\{s_{i1}, s_{i3}, s_{i5}\}$ for $\forall i \in N$), because the objects differ in colour but are otherwise identical in shape, size, material, *etc.*, Casajus' model would limit itself to defining one "attribute" (*i.e.*: colour), thereby labelling the

¹⁶ A recent contribution in the same line of research is Alós-Ferrer and Kuzmics [2012].

strategies as colour perceiving players see them (*i.e.*: $s_{i1} \mapsto \textit{grey}$, $s_{i3} \mapsto \textit{red}$, and $s_{i5} \mapsto \textit{black}$ for $\forall i \in N$).

On the other hand, departing from the aforementioned models, the theory to be introduced in the next section will argue that in the last example (*i.e.*: strategic game with set of strategies $\{s_{i1}, s_{i3}, s_{i5}\}$ for $\forall i \in N$) – given the same suggested labelling as above (*i.e.*: $s_{i1} \mapsto \textit{grey}$, $s_{i3} \mapsto \textit{red}$, $s_{i5} \mapsto \textit{black}$ for $\forall i \in N$) – colour perceivers can coordinate on a unique pure strategy (*i.e.*: *red*) if they feel that primary colours are most prominent (based on a binary relation allowing the salience comparison of pairs of alternatives), without needing to define any additional attribute.¹⁷ Also, the aforementioned models would not help us in the case in which players simply (ignore colours but) distinguish between the objects by the order in which the experimenter draws the blocks from the bag: in fact, the present theory will argue that the block-order perceivers can end up with a payoff of 1 if they feel that the first object to be randomly drawn is more prominent than the others. (Again, notice that Casajus' framework is based on a non-stochastic labelling structure, like Bacharach's and Janssen's.)

¹⁷ In effect, it should be stressed that here Casajus [2000] would need to define one additional attribute (*i.e.*: primary-/non primary- colour) so as to label strategies as primary-colour distinguishing players see them (*i.e.*: $s_{i1} \mapsto \textit{non primary}$, $s_{i3} \mapsto \textit{primary}$, and $s_{i5} \mapsto \textit{non primary}$ for $\forall i \in N$), in order to reach the same conclusion.

I.4. A model of conventions

The present theory builds on Sugden's and Casajus' frameworks so as to allow for the possibility that both stochastic and non-stochastic procedures may determine the labelling of strategies: for instance, assume it is common knowledge (among the block-order perceivers) that each player identifies the strategies by the order in which the objects are drawn from the bag and placed on the table; then, based on the order of the blocks as drawn by the experimenter, one could define a renaming of the strategic game at hand, thereby labelling the strategies as the players perceive them, namely $\{first, second, third\}$.¹⁸ Now, it should be noticed that such labelled strategies would remain symmetrical with one another, and therefore *first, second, third* should each still be assigned the same probability in a mixed-strategy solution: hence, in order to capture the different degrees of

¹⁸ Sugden [1995], building on Crawford and Haller [1990], proposes a set of properties for labelling procedures, namely: (A1) scrambling of labels for each player; (A2) independent labelling; (A3) common language; (A4) symmetry of labelling between players. Sugden's analysis then focuses on what he dubs "common-pool labelling procedures", that is, labelling procedures satisfying both (A2) and (A4) or, in plain words, procedures requiring that different players' labellings are determined by independent random draws from the same distribution. On the other hand, the present theory departs from Sugden's in that the labelling procedures considered here satisfy both (A1) and (A4), but not (A2): in a nutshell, the labelling procedures I shall focus on have a stochastic element since, for example, the order in which the blocks are drawn from the bag and placed on the table is determined by an exogenous random process, yet all such random draws are publicly observed by players; therefore, here (unlike Sugden [1995], and Crawford and Haller [1990]) players' descriptions of strategies are publicly observed permutations of the analyst's naming (*i.e.*: indexing) of strategies.

salience that players may attach to their labelled strategies, the present theory introduces a binary relation on the set of players' labelled strategies, allowing the salience comparison of pairs of alternatives (*i.e.*: a complete preordering, henceforth referred to as a "salience relation").

A few comments are in order. First, it is assumed that such a binary relation is based on an exogenous criterion that is mutually recognizable, like the order in which the objects are drawn by a third party, the consequent objects' spatial proximity to the players, *etc.*. It is clear that such a binary relation will induce a solution which is arbitrary to a certain degree. As Schelling [1960] puts it:

The solutions are, of course, arbitrary to this extent: any solution is "correct" if enough people think so. [...] Most situations – perhaps every situation for people who are practiced at this kind of game – provide some clue for coordinating behaviour, some focal point for each person's expectation of what the other expects him to expect to be expected to do. Finding the key, or rather finding a key – any key that is mutually recognized as the key becomes *the* key – may depend on imagination more than on logic; it may depend on analogy, precedent, accidental arrangement, symmetry, aesthetic or geometrical configuration, casuistic reasoning, and who the parties are and what they know about each other. (pp. 55-57, italics in original)

For instance, given a binary relation based on the order in which the objects are drawn by the experimenter, here a reason why players may end up attaching a greater degree of salience to the object labelled as *first* is, again, provided by Schelling: «If one [...] asks what number, among all positive numbers, is most clearly unique, or *what rule of selection would lead to unambiguous results*, one may be struck with the fact that the universe of all positive numbers has a "first" or "smallest" number» (Schelling [1960], p. 94, italics in original). So, in Choose an Object, such a rule of selection would result in players – who perceive the strategies as *first, second, third*, respectively – choosing the first block to be drawn by the experimenter (*i.e.*:

the strategy labelled as *first*), whatever is the strategy-index associated with it.

On a different note, it should be stressed that Sugden [1995] (like Crawford and Haller [1990]) accounts for uncertainty regarding the other players' comprehension of the game in that, as mentioned above, the players' descriptions of strategies are privately observed permutations of the analyst's indexing of strategies (see footnote 18). Instead, Casajus [2000] (like Bacharach [1993] and Janssen [2001]) models uncertainty by means of a sort of game of incomplete information in which the frame (*i.e.*: labelling) of a player is associated with a player's "type". Now, the present theory models uncertainty in an original way, which only in part relates to Bacharach, Janssen, and Casajus, in that a certain kind of player is associated with a certain labelling of strategies (and, in the present theory, in turn with a certain binary relation on the set of the player's labelled strategies); however, this study crucially departs from the existing literature because it implements a notion of the players' (un)awareness. In effect, by introducing a notion of unawareness: (*i*) the present theory accounts for both stochastic and non-stochastic labelling procedures, thereby providing a precise link between the players' perception of the game and their labelled strategies; (*ii*) it permits to explain coordination, in certain cases, even between differently-aware players (*i.e.*: between players that have partially-different sets of labelled strategies).

Therefore, the present theory implements Heifetz *et al.*'s [2006] system of multiple state spaces so as to account for such different players' perceptions: for instance, in Choose an Object, players who realize only the block orderings will be aware of state space $\Omega^B = \{\omega^i, \omega^{ii}, \omega^{iii}, \dots, \omega^{cxx}\}$, where each state refers to one of 120 possible orderings of the objects (based on the order of the blocks as drawn by the experimenter); similarly, players who realize only the colour differences will be aware of state space

$\Omega^C = \{\omega^{\bar{i}}, \omega^{\bar{ii}}, \omega^{\bar{iii}}, \dots, \omega^{\bar{xx}}\}$, where each state refers to one of 20 possible combinations of the coloured objects (based on the available coloured objects drawn by the experimenter, irrespective of the order). Then, as mentioned before, each player's labelling of strategies will depend directly on her perception of the game, that is, on the states she is aware of; given that, the present theory restricts the set of each player's (mixed) strategies by imposing two constraints that are implied by the notions of symmetry-invariance and salience relation, respectively.

The exposition of the model is organized in a manner such that each of the following sub-sections will describe one of the steps involved in the operation of a convention: (i) perception; (ii) labelling; (iii) salience comparison; (iv) expected utility maximization.

I.4.a. A general framework for perception

Let $\Omega^* = \bigcup_{k \in K} \Omega^k$ be the *full set of states*, with $(\Omega^k)_{k \in K}$ being disjoint subsets of Ω^* , where K is a space-index set and Ω^k denotes a generic set of states;¹⁹ for each $k \in K$ let (Ω^k, q^k) be a finite probability space, with $q^k \in \Delta(\Omega^k)$, where q^k is a probability measure over Ω^k and $\Delta(\Omega^k)$ is the set of probability measures over Ω^k . The interpretation is that each Ω^k is a collection of mutually exclusive specifications of the contingencies that an agent *perceives* in the context of a given decision problem (*i.e.*: each $\omega \in \Omega^k$ can be regarded as a full description of contingencies that relate to the game, in the player's own perspective). Notice that, for a given state space Ω^k , only one of the

¹⁹ An explicit ordered structure on $\{\Omega^k\}_{k \in K}$ will be introduced below.

states – referred to as the “true state” – *obtains*, thereby depicting how the game is actually expressed through the (k -perceiving) player’s own vocabulary (whereas the other states relate to possible, alternative descriptions of the game).

Example: Choose an Object (cont’d). I can now discuss the game introduced in section I.3. in light of the above definitions. Recall: players who realize only the block orderings will be aware of state space $\Omega^B = \{\omega^i, \omega^{ii}, \omega^{iii}, \dots, \omega^{cxx}\}$, where each state refers to one of 120 possible orderings of the objects; players who realize only the colour differences will be aware of state space $\Omega^C = \{\omega^{\bar{i}}, \omega^{\bar{ii}}, \omega^{\bar{iii}}, \dots, \omega^{\bar{c}\bar{x}\bar{x}}\}$, where each state refers to one of 20 possible combinations of the coloured objects. Hence, $|\Omega^B| = \binom{6}{3} 3! = \frac{6!}{3!(6-3)!} 3! = 120$, and $q^B(\omega) = 1/120$ for each $\omega \in \Omega^B$ (*i.e.*: the probability $q^B(\omega)$ of each of these states occurring, where occurring means “being brought about by the experimenter’s random draws”, is $1/120$);²⁰ similarly, $|\Omega^C| = \binom{6}{3} = 20$, and $q^C(\omega) = 1/20$ for each $\omega \in \Omega^C$. Furthermore, in Choose an Object, players who realize both the block orderings and the colour differences will be aware of the “more expressive” state space $\Omega^{BC} = \{\omega^{\bar{i}}, \omega^{\bar{ii}}, \omega^{\bar{iii}}, \dots, \omega^{\bar{c}\bar{x}\bar{x}}\}$, where each state includes the description of one of 120 possible orderings along with a description of the respective colours of the objects: to sum up, $|\Omega^{BC}| = 120$, and $q^{BC}(\omega) = 1/120$ for each $\omega \in \Omega^{BC}$.

²⁰ For a given set Ω , $|\Omega|$ denotes its cardinality.

Given the role played by the state spaces in depicting the individual's perception of the game, it is now convenient to discuss a notion of the players' (un)awareness. Indeed, as exemplified above, a player may view only some of the features of a (multi-person) decision problem (*i.e.*: some of the state spaces in $\{\Omega^k\}_{k \in K}$); or else, a player may be aware of all the potential contextual cues (*i.e.*: the full set of states $\Omega^* = \bigcup_{k \in K} \Omega^k$), and may yet assume that her co-players' perceptions are limited. Again, a player of Choose an Object (say, Player i), who is aware of both block orderings and colour differences (*i.e.*: aware of the state space Ω^{BC}), may well assume that her counterpart is only aware of Ω^B if she believes that her counterpart is colour-blind; she in turn (say, Player j) – if indeed colour-blind – will not be able to conceive of Ω^C as being relevant to the game perceived by her co-player since she (Player j) will merely ignore the existence of Ω^C . Therefore, I shall draw on Heifetz *et al.*'s [2006] model of unawareness so as to account for such different players' perceptions; as in their model, I use a system of surjective projections from each state space to every space that is weakly less expressive – this sub-section will initially detail such a system of projections and follow up with further discussion of the above illustration – the application will lead the way to a novel theory of conventions.

Consider a *complete lattice of disjoint spaces* $\mathcal{S} = \{\Omega^k\}_{k \in K}$ and let $\Omega^* = \bigcup_{k \in K} \Omega^k$ be the union of such spaces (*i.e.*: the full set of states); recall that a state ω is an element of some space Ω^k . Let \preceq denote an *expressivity relation*, that is, a partial preordering on \mathcal{S} such that, for any $\Omega, \Omega' \in \mathcal{S}$, $\Omega \preceq \Omega'$ means that “ Ω' is weakly more expressive than Ω ”: the interpretation is that

states in Ω' give a more detailed description of contingencies.²¹ Further, let $r = \{r_{\Omega, \Omega'}^{\Omega'}\}_{\Omega, \Omega' \in \mathcal{S}: \Omega \preceq \Omega'}$ denote a *set of surjective projections* from each state space to every space that is weakly less expressive (*i.e.*: $r_{\Omega, \Omega'}^{\Omega'}: \Omega' \rightarrow \Omega$ is a surjective projection from one space, Ω' , to a weakly less expressive one, Ω ; if $\omega \in \Omega'$, then $r_{\Omega, \Omega'}^{\Omega'}(\omega)$ is the projection of ω into a weakly less expressive space); such mappings are required to commute, that is, if $\Omega \preceq \Omega' \preceq \Omega''$ then $r_{\Omega, \Omega''}^{\Omega''} = r_{\Omega, \Omega'}^{\Omega'} \circ r_{\Omega', \Omega''}^{\Omega''}$. If $\omega \in \Omega'$, with a slight abuse of notation, denote by $\omega_{\Omega} := r_{\Omega, \Omega'}^{\Omega'}(\omega)$ the projection of ω into Ω and, similarly, by $\omega_{\Omega'} := r_{\Omega', \Omega''}^{\Omega''}(\omega') = \omega$ the projection of ω' into Ω' , with $\omega' \in \Omega''$. Given a mapping $r_{\Omega, \Omega'}^{\Omega'}: \Omega' \rightarrow \Omega$ and a generic (sub)set of states O , if $O \subseteq \Omega'$ denote by $O_{\Omega} = \{\omega_{\Omega} \in \Omega: \exists \omega \in O \text{ s.t. } \omega_{\Omega} = r_{\Omega, \Omega'}^{\Omega'}(\omega)\}$ the *set of all the images of the elements of O* (*i.e.*: the range of $r_{\Omega, \Omega'}^{\Omega'}|_O$, namely the range of the restriction of $r_{\Omega, \Omega'}^{\Omega'}$ to O). Let $g(\Omega) = \{\Omega': \Omega' \succeq \Omega\}$ be the *set of state spaces that are at least as expressive as Ω* ; for a generic (sub)set of states O – if now $O \subseteq \Omega$ – denote by $O^{\rightarrow} = \bigcup_{\Omega' \in g(\Omega)} (r_{\Omega, \Omega'}^{\Omega'})^{-1}|_O(O)$ the *union of the (inverse) projections from O to spaces weakly more expressive than Ω* (*i.e.*: all the pre-images of a subset O of the range of $r_{\Omega, \Omega'}^{\Omega'}$).²²

²¹ Regarding the interpretation, the fact that “states in Ω' give a more detailed description of contingencies than states in Ω do” does not mean that elements of the latter are also elements of the former, because \mathcal{S} is a lattice of *disjoint* spaces. Also note that it is required that $|\Omega| \leq |\Omega'|$.

²² The above specification of a lattice of state spaces mostly corresponds to that of Heifetz *et al.* [2006], although I give different interpretation and notation; moreover, note that Heifetz *et al.* [2006] do not define probability measures on each Ω^k . (In this respect, notice that here

Given the above apparatus, I can move on to characterize (multi-space) events: as usual, an event is defined as a subset of a state space, although here $E \subseteq \Omega^*$ is an *event* if it is of the form O^\rightarrow for some $O \subseteq \Omega$, with $\Omega \in \mathcal{S}$; notice that this implies that an event contains states lying in multiple spaces (besides, not every subset of Ω^* is an event).²³ Further, if O^\rightarrow is an event (with $O \subseteq \Omega^k$), its negation is defined as $\sim O^\rightarrow := (\Omega^k \setminus O)^\rightarrow$; if $O = \Omega^k$, then $\sim O^\rightarrow \equiv \emptyset^k$ where, for each $\Omega^k \in \mathcal{S}$, \emptyset^k is the vacuous event. Also, the *conjunction* $\bigwedge_{\lambda \in \Lambda}$ of a set of events $\{O_\lambda^\rightarrow\}_{\lambda \in \Lambda}$ is simply the intersection $\bigcap_{\lambda \in \Lambda} O_\lambda^\rightarrow$, whereas *disjunction* $\bigvee_{\lambda \in \Lambda}$ is defined from conjunction by the de Morgan's laws,²⁴ i.e.: $\bigvee_{\lambda \in \Lambda} O_\lambda^\rightarrow = \sim(\bigwedge_{\lambda \in \Lambda} \sim O_\lambda^\rightarrow)$.

Example: Choose an Object (cont'd). From the above discussion it is clear that a complete lattice of disjoint spaces $\mathcal{S} = \{\Omega^k\}_{k \in K}$, with $\Omega^* = \bigcup_{k \in K} \Omega^k$, is given by $\mathcal{S} = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$. Using the above notation, the set of state spaces that are at least as expressive as the – uninformative – empty set $\Omega^\emptyset := \{\emptyset\}$ is denoted by $g(\Omega^\emptyset) = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$. For space constraints the following figure illustrates all the aforementioned state spaces, but only part

the probability measure defined on some space Ω is obviously related to that defined on Ω' , if $\Omega \preceq \Omega'$, but for the purposes of the current study it is not necessary to explore this further.)

²³ In Heifetz *et al.*'s [2006] terminology, if E is an event in the above sense, then O is called the “basis” of E and $\Omega = \Omega(E)$ is the “base-space”. On a different note, it should be stressed that here $f|_O$ indicates the *restriction* of a certain mapping f , e.g.: if $f: X \rightarrow Y$ is a mapping and $O \subseteq X$, I denote the restriction of f to O by $f|_O$, that is, the function from O to Y such that $f|_O(o) = f(o)$ for $\forall o \in O$ (in Heifetz *et al.*'s [2006] terminology, “restriction” refers instead to the projection of a state into a less expressive space).

²⁴ $\sim(\bigcup_{\lambda \in \Lambda} O_\lambda) = \bigcap_{\lambda \in \Lambda} (\sim O_\lambda)$; $\sim(\bigcap_{\lambda \in \Lambda} O_\lambda) = \bigcup_{\lambda \in \Lambda} (\sim O_\lambda)$.

of the states: specifically, it shows only the states associated with the case where Nature has selected the strategic game with set of strategies $S_i = \{s_{i1}, s_{i2}, s_{i3}\}$ for $\forall i \in N$; recalling that objects no. 1-2 are grey and object no. 3 is red, such states are represented in the form of a sequence of blocks (in Ω^B) or a string symbolizing the coloured objects available (in Ω^C) or a sequence of blocks and colours (in Ω^{BC}).²⁵

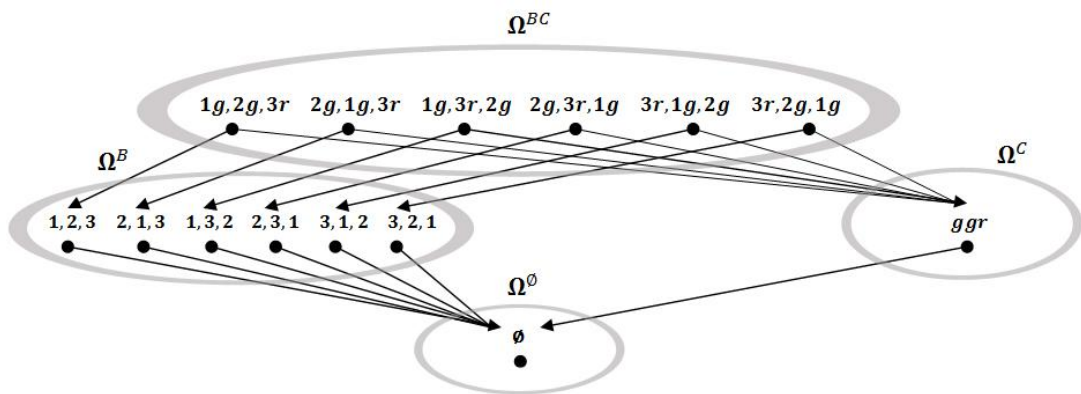


Figure I.1 - Some state spaces and projections in Choose an Object

²⁵ For instance, assume that the state $\omega = (1,2,3)$, with $\omega \in \Omega^B$, obtains: such a state provides information about the order in which the experimenter drew the blocks from the bag, where each integer corresponds to the number of each object. Similarly, in the case of the state $\omega = ggr$, with $\omega \in \Omega^C$, such a state provides information about the colours of the blocks, where g stands for “grey” and r for “red”.

Notice that projections are indicated by arrows:²⁶ it is clear that Ω^B and Ω^C both describe contingencies in a less expressive way than are described in Ω^{BC} ; Ω^\emptyset in turn describes contingencies in such a way as not to provide any information about the order in which the experimenter drew the blocks from the bag or about the colours of the blocks. Recalling that $|\Omega^C| = \binom{6}{3} = 20$, let $\Omega^C = \{\omega^{\bar{i}}, \omega^{\bar{ii}}, \omega^{\bar{iii}}, \dots, \omega^{\bar{xxx}}\} \equiv$

$\{ grr, g'rr, rgg, r'gg, gbb, g'bb, bgg, b'gg, rbb, r'bb, \}$,²⁷ also, recall that state space Ω^B is defined as $\Omega^B = \{\omega^i, \omega^{ii}, \omega^{iii}, \dots, \omega^{xxx}\}$ with each state referring to one of 120 possible orderings of the objects, whereas state space Ω^{BC} is defined as $\Omega^{BC} = \{\omega^{\bar{i}}, \omega^{\bar{ii}}, \omega^{\bar{iii}}, \dots, \omega^{\bar{xxx}}\}$ with each state including the description of one of 120 possible orderings along with a description of the respective colours of the objects. Given that, consider the event E that “Nature selects one or more grey blocks”: let O denote the set of states $\{ grr, g'rr, rgg, r'gg, gbb, g'bb, bgg, b'gg, \}$, hence $|O| = 16$, with $O \subset \Omega^C$; using the inverse projections from O to spaces weakly more expressive than Ω^C (i.e.: the state spaces $g(\Omega^C) = \{\Omega^C, \Omega^{BC}\}$), one can finally define the event

²⁶ Identity maps and compositions of projections are not shown in the figure.

²⁷ Notice that, for example, grr and $g'rr$ describe, respectively: the case in which the experimenter draws two red blocks and one of the grey blocks; the case in which the experimenter draws two red blocks and a grey block other than that of grr . Note that, in practice, not even colour perceiving players would tell grr from $g'rr$; yet, to the analyst the difference matters as each state is associated with a different grey block, hence a different strategy-index. As regards players – if they are indeed colour perceivers – they simply *realize* that, say, grr obtains when the experimenter randomly draws the objects associated with it.

E as the set O^\rightarrow , with $|O^\rightarrow| = 16 + 16 \cdot 6 = 112$. Furthermore, the event that “Nature does *not* select one or more grey blocks” (*i.e.*: the negation of E) is defined by the set $\sim O^\rightarrow$, with $|\sim O^\rightarrow| = 4 + 4 \cdot 6 = 28$. It follows that, unlike in the standard partitional models of knowledge, here it is possible that some states belong neither to an event nor to its negation: in fact, $O^\rightarrow \cup \sim O^\rightarrow \subset \Omega^*$ (*i.e.*: $O^\rightarrow \cup \sim O^\rightarrow \neq \Omega^*$, as $|O^\rightarrow| + |\sim O^\rightarrow| \neq |\Omega^*|$), which implies that there are states one could be unaware of (*e.g.*: those that do not belong to $O^\rightarrow \cup \sim O^\rightarrow$).

This sub-section concludes by detailing a set of properties for information functions I_i (accounting for multiple state spaces), which have been proposed by Heifetz *et al.* [2006]. Recall that, for each player $i \in N$, \mathcal{S}_i was referred to as the information partition of player i , and I_i as the information (set-valued) function associating with every state $\omega \in \Omega$ a non-empty subset $I_i(\omega) \subseteq \Omega$ (with $I_i(\omega)$ being interpreted as the set of states the agent considers possible when the true state is ω).²⁸ At this point it is convenient to extend such an information structure so as to account for multiple state spaces: henceforth, unless otherwise stated, the letter I will stand for a *generalized information function*, that is, a set-valued function (with \mathcal{S} here denoting merely a collection of mutually disjoint subsets of Ω^*) that associates with every state $\omega \in \Omega^*$ (with $\Omega^* = \bigcup_{k \in K} \Omega^k$) a non-empty subset $I(\omega) \subseteq \Omega^*$; formally, for each player $i \in N$, a generalized information

²⁸ See section I.2.c. above.

function is defined as $I_i: \Omega^* \rightarrow 2^{\Omega^*} \setminus \emptyset$.²⁹ It is assumed that such a function satisfies the following properties:

(*gi.i*: confinedness) $\omega \in \Omega'' \Rightarrow \exists \Omega' \preceq \Omega''$ s. t. $I_i(\omega) \subseteq \Omega'$

(*gi.ii*: reflexivity) $\forall \omega \in \Omega^*, \omega \in (I_i(\omega))^\rightarrow$

(*gi.iii*: stationarity) $\omega' \in I_i(\omega) \Rightarrow I_i(\omega') = I_i(\omega)$

(*gi.iv*: projections preserve awareness) ($\omega \in \Omega''$ and $\omega \in I_i(\omega)$ and $\Omega' \preceq \Omega''$)
 $\Rightarrow \omega_{\Omega'} \in I_i(\omega_{\Omega'})$

(*gi.v*: projections preserve ignorance) ($\omega \in \Omega''$ and $\Omega' \preceq \Omega''$)
 $\Rightarrow (I_i(\omega))^\rightarrow \subseteq (I_i(\omega_{\Omega'}))^\rightarrow$

(*gi.vi*: projections preserve knowledge) ($\Omega' \preceq \Omega'' \preceq \Omega'''$ and $\omega \in \Omega'''$ and $I_i(\omega) \subseteq \Omega''$) $\Rightarrow (I_i(\omega))_{\Omega'} = I_i(\omega_{\Omega'})$.

A few comments are in order. Properties (*gi.ii-iii*) reproduce the standard properties of any traditional (partitional) information function: *reflexivity* says that a player never excludes the true state from the set of states she regards as possible; *stationarity* says that a player uses the consistency or inconsistency of states with her information to make inferences about the state. Properties (*gi.i*) and (*gi.iv-vi*) have been proposed by Heifetz *et al.* [2006]: *confinedness* says that the states a player considers as possible at a certain state $\omega \in \Omega''$ are all described with a vocabulary no more expressive than Ω'' ; properties (*gi.iv-vi*) compare a player's information set at a certain

²⁹ For a given set Ω , 2^Ω denotes its power set.

state ω with the information set at the projection of ω into a weakly less expressive space. The rationale is to guarantee that the states a player considers as possible at some $\omega \in \Omega''$ can be described by that very player in a weakly less detailed – yet consistent – way for some $\Omega' \preceq \Omega''$. Given that, one may introduce an unawareness operator which, now, can indeed capture the case in which a player does not know an event and does not know that she does not know it.

The following definition of an “indirect-realism knowledge structure” compactly characterizes the above general framework for perception.

Definition I.2. Given a strategic form game G , an *indirect-realism knowledge structure* (henceforth simply a “knowledge structure”) of G is given by $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_{\Omega}^{\Omega'}\}, (\mathcal{G}_i)_{i \in N} \rangle$, with each component being defined as above.³⁰

Before proceeding, I shall note that a knowledge structure of this form (along with the associated frame, to be introduced in the next sub-section) is comparable to the approach to scientific inquiry referred to as “model-dependent realism”. As physicist Stephen Hawking puts it: «According to model-dependent realism, it is pointless to ask whether a model is real, only whether it agrees with observation. If there are two models that both agree with observation [...], then one cannot say that one is more real than another.

³⁰ \mathcal{G}_i is here informally defined as a collection of mutually disjoint subsets of Ω^* ; see formula (1.4.1) in the next sub-section for a more precise definition of \mathcal{G}_i .

One can use whichever model is more convenient in the situation under consideration. [...] Model-dependent realism applies not only to scientific models but also to the conscious and sub-conscious mental models we all create in order to interpret and understand the everyday world. There is no way to remove the observer – us – from our perception of the world, which is created through our sensory processing and through the way we think and reason. Our perception – and hence the observations upon which our theories are based – is not direct, but rather is shaped by a kind of lens, the interpretive structure of our human brains» (Hawking and Mlodinow [2010], pp. 61-62).

I.4.b. From perception to labelling

I can now turn to define the other ingredients of this theory of conformity in coordination games by introducing a perception-based model of labelling, in such a way as to allow for the possibility that both stochastic and non-stochastic procedures may determine the labelling of strategies. Given a set of properties T , with generic element t , in what follows a “label” is defined as a rule that assigns to every element s_{ia} of the strategy set S_i a unique element π of a set Π_t of instances of property t .

Definition I.3. Given a strategic form game G , a *label* is a function λ_ω^t that assigns to each strategy-index $s_{ia} \in S_i$ one element from a set Π_t of instances of property t ; that is, a label $\lambda_\omega^t: S_i \rightarrow \Pi_t$ is a rule that expresses strategies as an instance of a given property.

In a nutshell, given a class of sets $\wp = \{\Pi_t: \pi \text{ is an instance of property } t\}_{t \in T}$, a label function represents strategies as an instance π of some property t (with $t \in T$). For example, if t is the property “colour”, then Π_t is the set of

instances of colours, *e.g.*: $\Pi_t = \{grey, red, black\}$. Hence, for some $s_{ia} \in S_i$, $\lambda_\omega^t(s_{ia})$ denotes a *perception-based labelled strategy* (henceforth simply a “labelled strategy”) and $\lambda_\omega^t(S_i)$ denotes the *set of labelled strategies*. The interpretation is that a label conveys the descriptions by which players recognize strategies.

Now, it is assumed that a τ -tuple of labels, with $\tau = |T|$, is associated with each state $\omega \in \Omega^*$: let Λ denote the *set of labels*, with generic element λ_ω^t ; (general) “availability” is a rule that assigns to each state a certain number (τ) of labels.

Definition 1.4. Given a strategic form game G and the full set of states $\Omega^* = \bigcup_{k \in K} \Omega^k$, *availability* is an injective set-valued function φ that assigns to each state $\omega \in \Omega^k$ (for all $\Omega^k \in \mathcal{S}$, with $\mathcal{S} = \{\Omega^k\}_{k \in K}$) a τ -tuple of labels $\lambda_\omega^t \in \Lambda$ (with $\tau = |T|$); that is, availability $\varphi: \Omega^* \rightarrow \Lambda$ is a rule associating with each state ω a certain number of labels λ_ω^t by which to express strategies with the vocabulary of Ω^k .

To sum up, given the set Λ of labels, for each state there exists a distinct τ -tuple of labels $\lambda_\omega^t \in \Lambda$, where τ equals the cardinality of the set T of properties. In what follows, for each state $\omega \in \Omega^*$ I shall use bold letters to denote such a τ -tuple of labels, *i.e.*: $\lambda_\omega := (\lambda_\omega^{t'}, \lambda_\omega^{t''}, \dots, \lambda_\omega^{t^\tau})$. Note that it is assumed that, if a certain state ω belongs to some space Ω^k which is not rich enough to express a particular property t' , then $\lambda_\omega^{t'}$ is not defined at S_i .

Example: Choose an Object (cont'd). Recall that a complete lattice of disjoint spaces $\mathcal{S} = \{\Omega^k\}_{k \in K}$, with $\Omega^* = \bigcup_{k \in K} \Omega^k$, is given by $\mathcal{S} = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$. Now, the set of properties can be defined as $T := \{\emptyset, B, C\}$, where $t = B$ is the

“order” property while $t = C$ is the “colour” property; $t = \emptyset$ is the “null” property, which does not explain anything. The set of labels can be defined as $\Lambda := \{\lambda_\omega\}_{\omega \in \Omega^*}$ where, for each $\omega \in \Omega^*$, $\lambda_\omega := (\lambda_\omega^\emptyset, \lambda_\omega^B, \lambda_\omega^C)$. Again, it should be stressed that if ω belongs to some Ω^k not rich enough to characterize a particular property t' , then $\lambda_\omega^{t'}$ is not defined at S_i , *e.g.*: if some state ω belongs to Ω^B , then λ_ω^C is not defined at S_i (similarly, if $\omega \in \Omega^C$, λ_ω^B is not defined at S_i). Given that, consider the case where the experimenter randomly draws the objects from the bag in a way as captured by state $\omega = (4r, 5b, 6b)$, with $\omega \in \Omega^{BC}$: clearly, $\omega = (4r, 5b, 6b)$ is the state associated with the case where Nature has selected the strategic game with set of strategies $S_i = \{s_{i4}, s_{i5}, s_{i6}\}$ for $\forall i \in N$. Hence, using the above notation, the value of the availability function at $\omega = (4r, 5b, 6b)$ is $\varphi(\omega) = (\lambda_\omega^\emptyset, \lambda_\omega^B, \lambda_\omega^C)$ and the label functions $\lambda_\omega^B: S_i \rightarrow \Pi_B$, $\lambda_\omega^C: S_i \rightarrow \Pi_C$, and $\lambda_\omega^\emptyset: S_i \rightarrow \Pi_\emptyset$ are given as follows.³¹

³¹ Recall that the analyst identifies the strategies with the numbers of the objects (object no. 4 is red and objects no. 5-6 are black). In this connection, it should be stressed once again that – although such numbers are invisible to the players – the experimenter’s random draws are publicly observed by the players: therefore, here players’ descriptions of strategies are publicly observed permutations of the analyst’s indexing of strategies (see footnote 18).

S_i	$\lambda_\omega^B(s_{ia})$	$\lambda_\omega^C(s_{ia})$	$\lambda_\omega^\emptyset(s_{ia})$
s_{i4}	<i>first</i>	<i>red</i>	<i>object</i>
s_{i5}	<i>second</i>	<i>black</i>	<i>object</i>
s_{i6}	<i>third</i>	<i>black</i>	<i>object</i>

Furthermore, let $r = \{r_\Omega^{\Omega^{BC}}\}_{\Omega, \Omega^{BC} \in \mathcal{S}: \Omega \preceq \Omega^{BC}}$ denote a set of surjective projections from Ω^{BC} to every space Ω that is weakly less expressive: for example, if $\Omega = \Omega^B$, then $\omega_\Omega \equiv r_{\Omega^B}^{\Omega^{BC}}(\omega)$ represents the projection of ω into the weakly less expressive space Ω^B , that is, $\omega_\Omega = (4,5,6)$; from the above discussion it follows that, in this case, $\lambda_{\omega_\Omega}^C(S_i)$ is not defined.

Given that each player's labelling of strategies should depend on her possibly limited perception of the game, it is convenient to define each "individual availability" as the restriction of (general) availability φ to a particular subset of Ω^* , namely to the "set of player i 's perceivable states" (henceforth denoted by Ω_i^* for each $i \in N$). In order to do so, I shall first recall that (given a complete lattice of disjoint spaces $\mathcal{S} = \{\Omega^k\}_{k \in K}$) \preceq is a partial preordering on \mathcal{S} such that, for any $\Omega, \Omega' \in \mathcal{S}$, $\Omega \preceq \Omega'$ means that Ω' is weakly more expressive than Ω . Then, for each player $i \in N$ let $\widehat{K}_i \subseteq K$ denote player i 's space-index set and, given that, let Ω_i^{max} denote the *maximum* (or

greatest) state space player i is aware of, that is, $\Omega_i^{max} = \max_{k \in \hat{K}_i} \{\Omega^k\}_{k \in K}$: the interpretation is that Ω_i^{max} depicts “the most expressive state space perceivable to player i ”.³² Hence, for each $i \in N$ let \mathcal{S}_i be defined as:

$$\mathcal{S}_i = \{\Omega^k\}_{\Omega^k \in \mathcal{S}: \Omega^k \leq \Omega_i^{max}} . \tag{1.4.1}$$

Given that, one can derive the *set of player i 's perceivable states*, for each $i \in N$, as $\Omega_i^* := \bigcup_{\Omega^k \in \mathcal{S}: \Omega^k \leq \Omega_i^{max}} \Omega^k \equiv \bigcup_{k \in \hat{K}_i} \Omega^k$. I can now proceed to define each player's *individual availability* – henceforth denoted by φ_i for $\forall i \in N$ – as the restriction of (general) availability φ to the set Ω_i^* of i 's perceivable states:

$$\varphi_i = \varphi|_{\Omega_i^*} . \tag{1.4.2}$$

³² Note that – given a partial preordering on \mathcal{S} , and an information function I_i – the *maximum (or greatest) state space player i is aware of* can be derived for each $i \in N$ by taking the maximum across the sets of the images of the states in Ω^k (for $\forall \Omega^k \in \mathcal{S}$), i.e.: $\Omega_i^{max} = \max_{\Omega^k \in \mathcal{S}} \{\bigcup_{\omega \in \Omega^k} I_i(\omega)\}_{\Omega^k \in \mathcal{S}}$. In plain words, Ω_i^{max} is obtained by taking the maximum across all the state spaces player i considers possible (with $\bigcup_{\omega \in \Omega^k} I_i(\omega) = I_i(\Omega^k)$), therefore Ω_i^{max} simply represents to player i the most expressive possible state space.

To sum up, given a strategic form game G and the associated knowledge structure $\langle \Omega^*, (\Omega^k, q^k)_{k \in K}, (\mathcal{S}_i)_{i \in N} \rangle$, i 's individual availability is an injective set-valued function φ_i that assigns to each state $\omega \in \Omega^k$ – for all $\Omega^k \in \mathcal{S}_i$ – a τ -tuple of labels $\lambda_\omega^t \in \Lambda$. Again, the intuition is that each player's labelling of strategies depends on her own (*i.e.*: possibly limited) perception of the game.

The following definition compactly characterizes a “frame” as a structure that, in conjunction with a knowledge structure, fully determines the player's own comprehension and description of a certain strategic form game.

Definition 1.5. Given a strategic form game G and the associated knowledge structure $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_\Omega^\Omega\}, (\mathcal{S}_i)_{i \in N} \rangle$, a *frame* is a description of G and is given by $\langle \Omega^*, \varphi, \Lambda, T, (\Pi_t)_{t \in T} \rangle$, with each component being defined as above.

It should be noticed that the current formalization of the players' comprehension and description of a game, although somehow related to that of Casajus [2000], crucially differs from it by accounting for both stochastic and non-stochastic procedures in the specification of the labelling of strategies (see footnote 18). Moreover, while Casajus directly allows for different label functions across players, here the frame does *not vary directly* with players (in fact, it varies with properties $t \in T$), however different labellings may be employed among players by virtue of the different players' information functions: once again, here each player's labelling of strategies depends on her perception of the game.

Further, going back to the game of Choose an Object – and considering again the case where state $\omega = (4r, 5b, 6b)$ occurs – it is clear that λ_ω^B generates an isomorphic game (see footnote 7), that is, to the analyst

the mathematical structures of the original game with strategies $S_i = \{s_{i4}, s_{i5}, s_{i6}\}$ and of the game with labelled strategies $\lambda_\omega^B(S_i) = \{first, second, third\}$ represent exactly the same decision problem, and as such should not be treated differently (Harsanyi and Selten [1988], Ch. 3).³³ Thus, since $\lambda_\omega^B(s_{i4}), \lambda_\omega^B(s_{i5}), \lambda_\omega^B(s_{i6})$ are symmetric strategies, players ought to assign them the same probability in a symmetry-invariant equilibrium: what follows formalizes this argument by making use of the notion of a frame.

Remark I.1. Given a strategic form game G , and the associated knowledge structure $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_\Omega^{\Omega'}\}, (\mathcal{S}_i)_{i \in N} \rangle$ and frame $\langle \Omega^*, \varphi, \Lambda, T, (\Pi_t)_{t \in T} \rangle$:

$$\begin{aligned} & \text{(for } \forall s_{ia}, s_{i\tilde{a}} \in S_i: a \neq \tilde{a}; \lambda_\omega^t(s_{ia}) = \lambda_\omega^t(s_{i\tilde{a}}) \text{ s.t. } \lambda_\omega^t \in \varphi_i(\Omega_i^*)) \\ & \implies p(s_{ia}) = p(s_{i\tilde{a}}). \end{aligned}$$

A few observations are in order. First, notice that remark I.1 simply says that if two strategies with different indices (*i.e.*: $s_{ia}, s_{i\tilde{a}} \in S_i: a \neq \tilde{a}$) are labelled in the same way for a given property $t \in T$ (*i.e.*: $\lambda_\omega^t(s_{ia}) = \lambda_\omega^t(s_{i\tilde{a}})$), then the probability with which they are chosen by player i must be the same (*i.e.*: $p(s_{ia}) = p(s_{i\tilde{a}})$) for some λ_ω^t , *provided that i can actually conceive of those labelled strategies* (*i.e.*: $\lambda_\omega^t \in \varphi_i(\Omega_i^*)$, with $\varphi_i(\Omega_i^*)$ denoting the range of individual availability φ_i). Also, note that remark I.1 (which, following Casajus [2000], is derived from Harsanyi and Selten's requirement of invariance with

³³ Conversely, it is also clear that λ_ω^C does not generate an isomorphic game.

respect to isomorphisms) uses some version of Bernoulli’s “principle of insufficient reason” (*i.e.*: if the strategies are indistinguishable except for their names, then each should be assigned the same prior belief); similarly the frameworks of Bacharach [1993] and Janssen [2001] too, more or less implicitly, use some version of Bernoulli’s principle of insufficient reason.

I.4.c. From labelling to salience comparison

From the above discussion it is now clear that the solution concepts of all the aforementioned models – as a consequence of their use of the principle of insufficient reason – depend on the number of strategies which are labelled in the same way: in other words, coordination heavily depends on the rarity of some labelled strategies relative to all other labelled strategies. Indeed, what those models lack is an appreciation of the degree of salience of each labelled strategy (for a given property).

Example: Choose an Object (cont’d). Consider the case where the experimenter randomly draws the objects from the bag in a way as captured by state $\omega = (4r, 5b, 6b)$, with $\omega \in \Omega^{BC}$: again, $\omega = (4r, 5b, 6b)$ is the state associated with the case where Nature has selected the strategic game with set of strategies $S_i = \{s_{i4}, s_{i5}, s_{i6}\}$ for $\forall i \in N$. Now, I shall temporarily make the assumption that no player is aware of the block orderings, hence no player conceives of any labelling more expressive than the colour identifier: therefore, given $\mathcal{S}_i = \{\Omega^\emptyset, \Omega^C\}$ for $\forall i \in N$ – and denoting $\Omega \equiv \Omega^C$ – let $\omega_\Omega \equiv r_{\Omega^C}^{\Omega^{BC}}(\omega)$ be the projection of ω into the weakly less expressive space

Ω^C ; it follows that $\lambda_{\omega_\Omega}^C \in \varphi_i(\omega_\Omega)$, with $\lambda_{\omega_\Omega}^C: S_i \rightarrow \Pi_C$ being given as before for $\forall i \in N$.³⁴ Recall that remark I.1 states that if two different strategies are labelled in the same way (for a given property $t \in T$), then the probability with which they are chosen by player i must be the same for some $\lambda_\omega^t \in \varphi_i(\Omega_i^*)$. So, in this case remark I.1 implies that $p(\text{black}) \equiv p(\lambda_{\omega_\Omega}^C(s_{i5})) = p(\lambda_{\omega_\Omega}^C(s_{i6}))$ for $\lambda_{\omega_\Omega}^C \in \varphi_i(\omega_\Omega)$,³⁵ however it does not say anything about $p(\text{black})$ being greater or less than $p(\text{red})$. In effect, many may probably feel that red is more salient than black.

In order to capture the different degrees of salience that players may attach to their labelled strategies, this sub-section introduces a binary relation on the set of instances of property t (or, equivalently, on the set of labelled strategies via λ_ω^t).

Definition I.6. Given a set Π_t of instances of property t , with $t \in T$, *salience* is a binary relation \succsim^t defined on Π_t which allows the comparison of pairs of instances π, π' of t (hence, the comparison of pairs of alternative labelled strategies via λ_ω^t , when $\lambda_\omega^t(s_{i\alpha}) = \pi$, $\lambda_\omega^t(s_{i\bar{\alpha}}) = \pi'$ for some $\lambda_\omega^t \in \Lambda$); that is,

³⁴ The above assumption implies that, in this example, λ_ω^B is not defined for any $\omega \in \Omega_i^*$.

³⁵ Similarly, if one lets $\omega_{\Omega^\emptyset} \equiv r_{\Omega^\emptyset}^{\Omega^{BC}}(\omega)$ be the projection of ω into the least expressive space Ω^\emptyset , obviously given that $\lambda_{\omega_{\Omega^\emptyset}}^\emptyset \in \varphi_i(\Omega_i^*)$ for $\forall i \in N$, in this example remark I.1 also implies that $p(\text{object}) \equiv p(\lambda_{\omega_{\Omega^\emptyset}}^\emptyset(s_{i4})) = p(\lambda_{\omega_{\Omega^\emptyset}}^\emptyset(s_{i5})) = p(\lambda_{\omega_{\Omega^\emptyset}}^\emptyset(s_{i6}))$ for $\lambda_{\omega_{\Omega^\emptyset}}^\emptyset \in \varphi_i(\omega_{\Omega^\emptyset})$.

salience is a complete preordering on Π_t such that, for any $\pi, \pi' \in \Pi_t$, $\pi \succcurlyeq^t \pi'$ means that “ π is weakly more salient than π' ”.

In brief, given a class of sets $\wp = \{\Pi_t : \pi \text{ is an instance of property } t\}_{t \in T}$, recall that a label function represents strategies as an instance π of some property t (with $t \in T$). For example, denoting by $t = C$ the colour property, Π_C is the set of instances of colours, *e.g.*: $\Pi_C = \{grey, red, black\}$; hence, in this case, salience \succcurlyeq^C allows the comparison of pairs of alternative colour-labelled strategies, *e.g.*: $red \succcurlyeq^C grey \succcurlyeq^C black$. Similarly, in the case where players identify the strategies by the order in which the objects are randomly drawn from the bag, denote by $t = B$ the order property and let Π_B be the set $\{first, second, third\}$; hence, in this case, salience \succcurlyeq^B allows the comparison of pairs of alternative order-labelled strategies, *e.g.*: $first \succcurlyeq^B second \succcurlyeq^B third$.³⁶

Note that, for the sake of simplicity, the present theory assumes that *salience is based on an exogenous criterion*: in effect, one may think of salience as a (biology- or culture-dependent) binary relation based on a mutually recognizable criterion; as a consequence, for a given game and associated knowledge and frame structures, it is assumed that there exists a unique salience relation \succcurlyeq^t for each property $t \in T$.

³⁶ The symbol used to denote *salience*, for some t (*i.e.*: a binary relation \succcurlyeq^t defined on the set Π_t of instances of property t) must not be confused with the symbol used to denote *expressivity* (*i.e.*: a binary relation \preceq defined on a complete lattice \mathcal{S} of disjoint state spaces).

Example: Choose an Object (cont'd). In the case of the colour property, one may argue that primary colours are most salient: this could be captured adopting the “RGB” colour model. RGB is an additive colour model, used in computer graphics, in which red, green and blue light are added together in various ways to reproduce a broad range of colours (the name of the model comes from the initials of the three additive primary colours, *i.e.*: R =red, G =green, and B =blue). In computing, the three component values are usually inputted as integers in the range 0 to 255; here – given the set of integers $X = \{0,1,2, \dots, 255\}$, with generic element x – one can easily define the set Π_C of instances of colours as the set of 3-dimensional vectors $\Pi_C \equiv X^3 = \{\pi = (x_R, x_G, x_B): x_v \in X \text{ for } \forall v = R, G, B\}$. Next, one needs to define a notion of length of a vector in X^3 , that is, a norm on X^3 (*i.e.*: $\|\cdot\|: X^3 \rightarrow \mathbb{R}^+$): for example, consider the *sup norm*, defined as the absolute value of the largest component of a vector, that is, $\|\pi\| := \max_v \{x_v | v = R, G, B\}$. Now, consider the case where Nature has selected the strategic game with set of strategies $\{s_{i1}, s_{i3}, s_{i5}\}$ for $\forall i \in N$, and let the set of colour-labelled strategies be given by $\{grey, red, black\}$, with $s_{i1} \mapsto grey$, $s_{i3} \mapsto red$, and $s_{i5} \mapsto black$ for $\forall i \in N$. Then, from computer graphics we know that a common shade of grey is given by the vector $\pi = (127,127,127)$, red is given by the vector $\pi = (255,0,0)$, and black by the vector $\pi = (0,0,0)$: it is clear that the sup norm implies that $\|red\| > \|grey\| > \|black\|$. Finally, salience relation \succ^C can be defined accordingly, thereby indicating that primary colours are most salient, that is, $red \succ^C grey \succ^C black$.³⁷

³⁷ The assumption that I have arbitrarily made here (that primary colours are more salient

The next assumption follows from the introduction of a salience relation.

Assumption I.1. Given a strategic form game G and the associated knowledge structure $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_\Omega^\Omega\}, (\mathcal{S}_i)_{i \in N} \rangle$, frame $\langle \Omega^*, \varphi, \Lambda, T, (\Pi_t)_{t \in T} \rangle$, and salience relations $(\succsim^t)_{t \in T}$:

$$\left(\begin{array}{l} \text{for } \forall s_{ia}, s_{i\tilde{a}} \in S_i: a \neq \tilde{a}; \lambda_\omega^t(s_{ia}), \lambda_\omega^t(s_{i\tilde{a}}) \in \Pi_t \text{ s.t. } \lambda_\omega^t(s_{ia}) = \pi, \lambda_\omega^t(s_{i\tilde{a}}) = \pi' \\ \text{and } \pi \succsim^t \pi' \text{ and } \lambda_\omega^t \in \varphi_i(\Omega_i^*) \end{array} \right)$$

$$\Rightarrow p(s_{ia}) \geq p(s_{i\tilde{a}}).$$

Assumption I.1 simply says that if two strategies with different indices (*i.e.*: $s_{ia}, s_{i\tilde{a}} \in S_i: a \neq \tilde{a}$) are labelled differently for a given property $t \in T$ (*i.e.*: $\lambda_\omega^t(s_{ia}) = \pi, \lambda_\omega^t(s_{i\tilde{a}}) = \pi'$) and if the instances of property t with which s_{ia} and $s_{i\tilde{a}}$, respectively, are associated are the first more salient than the second (*i.e.*: $\pi \succsim^t \pi'$), then the probability with which they are chosen by player i should be such that the former is weakly more likely than the latter (*i.e.*:

than others), in fact, has some sort of scientific foundation. The modern RGB colour model is derived from the Young-Helmholtz tri-chromatic colour vision theory, which was developed in the 19th century (by polymaths Thomas Young and Hermann von Helmholtz) in order to explain the way the photoreceptor cells of human eyes enable colour vision. According to such a theory, in fact, there exist three types of photoreceptors (now referred to as “cone cells”) in the eye, each of which is sensitive to a particular range of visible light. Evidence that the eye does contain three types of cone has effectively been provided relatively recently by examining the light emerging from the eye after reflection off the retina.

$p(s_{ia}) \geq p(s_{i\bar{a}})$) for some λ_ω^t , provided that i can actually conceive of those labelled strategies (i.e.: $\lambda_\omega^t \in \varphi_i(\Omega_i^*)$).

Example: Choose an Object (cont'd). Going back to the case where $\lambda_{\omega_\Omega}^C \in \varphi_i(\omega_\Omega)$, with label function $\lambda_{\omega_\Omega}^C: S_i \rightarrow \Pi_C$ and salience relation \succsim^C being defined as above, remark I.1 and assumption I.1 imply that $p(\lambda_{\omega_\Omega}^C(s_{i5})) = p(\lambda_{\omega_\Omega}^C(s_{i6})) \equiv p(\text{black}) \leq p(\text{red}) \equiv p(\lambda_{\omega_\Omega}^C(s_{i4}))$ for $\lambda_{\omega_\Omega}^C \in \varphi_i(\omega_\Omega)$. Therefore, player i 's (mixed) strategies – respecting remark I.1 and assumption I.1 for $\lambda_{\omega_\Omega}^C$ – can be represented as the vector of probabilities $\sigma_i = (p(s_{i4}), p(s_{i5}), p(s_{i6})) = (p_i, \frac{1}{2}\tilde{p}_i, \frac{1}{2}\tilde{p}_i)$ with $p_i \geq \tilde{p}_i$ ($p + \tilde{p}_i = 1$).

To sum up, remark I.1 and assumption I.1 have the effect of restricting the set of (mixed) strategies by imposing constraints that capture the notions of symmetry and salience, respectively. Before proceeding to the next subsection, I shall abuse notation denoting by $\rho_i(\lambda^t)$ a generic (mixed) strategy of player i respecting remark I.1 and assumption I.1 for some λ^t . (In the above example, $\rho_i(\lambda_{\omega_\Omega}^C) \equiv (p_i, \frac{1}{2}\tilde{p}_i, \frac{1}{2}\tilde{p}_i)$ with $p_i \geq \tilde{p}_i$ ($p + \tilde{p}_i = 1$).)

I.4.d. Expected utility maximization

The last step involved in the operation of a convention is an expected utility maximization. As mentioned above, the present theory models uncertainty in an original way, which is only in part similar to Bacharach [1993] in that a certain kind of player – or rather, here, a player perceiving certain states – is associated with a certain labelling of strategies (and, in the present theory, in turn with a certain binary relation on the set of the player's labelled strategies). In fact, given the set of a player's actually realized attributes,

Bacharach assumes that the distribution of such sets of realized attributes in the population of players is exogenously given.³⁸ Bacharach uses this distribution to derive the “subjective probability that a player having a repertoire assigns to the other player having some other repertoire”; in doing so, Bacharach [1993] (as well as Janssen [2001] and Casajus [2000]) assumes that each player believes that her co-player’s repertoire is some subset of, or equal to, her own repertoire. Now, while Bacharach [1993] does not provide a solid foundation for those assumptions (in terms of a knowledge structure),³⁹ the present theory justifies its own construction by employing an indirect-realism knowledge structure as defined in section I.4.a. above. Besides, this study departs from the existing literature also because, as it will soon be clear, by implementing a notion of the players’ (un)awareness it effectively permits to explain coordination, in certain cases, even between differently-aware players (*i.e.*: between players that have partially-different sets of labelled strategies).

Thus, I shall introduce a few more definitions. Given a complete lattice of disjoint spaces \mathcal{S} and an expressivity relation \preceq (*i.e.*: a partial preordering on \mathcal{S} as defined in section I.4.a. above), let $\underline{\mathcal{S}}$ denote the *set of minimal*

³⁸ The sets of attributes are referred to as “families” in Bacharach’s [1993] terminology (each family roughly corresponds to the *set of instances of property t* of the present theory); Bacharach refers to a set of such families as a “repertoire”.

³⁹ Bacharach and Stahl [2000] justify those assumptions by implementing a Level- k model of bounded rationality.

elements of $\mathcal{S} \setminus \Omega^\emptyset$.⁴⁰ Similarly, given $\mathcal{S}_i = \{\Omega^k\}_{\Omega^k \in \mathcal{S}: \Omega^k \leq \Omega_i^{\max}}$ for each $i \in N$ (as defined in formula (1.4.1) above), let $\underline{\mathcal{S}}_i$ denote the set of minimal elements of $\mathcal{S}_i \setminus \Omega^\emptyset$. Before proceeding, it should be noted that the reason why I have defined such sets is that each space $\Omega \in \underline{\mathcal{S}}$ may (or may not) describe the states $\omega \in \Omega^*$ according to one property (or basic vocabulary), whereas all the spaces more expressive than those in $\underline{\mathcal{S}}$ just employ various combinations of those basic vocabularies. (For example, in the game of Choose an Object, $\mathcal{S} := \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$, therefore $\underline{\mathcal{S}} = \{\Omega^B, \Omega^C\}$; if player i is not aware of the block orderings, then $\mathcal{S}_i = \{\Omega^\emptyset, \Omega^C\}$ and $\underline{\mathcal{S}}_i = \{\Omega^C\}$, that is, Ω^C is his only basic vocabulary.) Given that, let $v(\Omega)$ denote the objective probability of a player being aware of some space $\Omega \in \underline{\mathcal{S}}$. So, for each $\Omega \in \underline{\mathcal{S}}$, $v(\Omega)$ represents the probability of success in a Bernoulli trial, where the outcomes of the experiment are success (being interpreted as the “outcome that a player is aware of Ω ”) and failure (being interpreted as the “outcome that a player is not aware of Ω ”). It is assumed that the probabilities of a player being aware of different spaces in $\underline{\mathcal{S}}$ are mutually independent (and exogenously given).

Now, the present theory assumes that players are naïve in the following way: if player i is aware of some state spaces (*i.e.*: aware of those contained in \mathcal{S}_i and unaware of those in $\mathcal{S} \setminus \mathcal{S}_i$), then she presumes that her co-player may be aware of *one* space in $\underline{\mathcal{S}}_i \cup \Omega^\emptyset$ *only*; that is, player i

⁴⁰ Given a relation \leq on \mathcal{S} , $\Omega \in \mathcal{S}$ is a *minimal element* if there is no element $\tilde{\Omega} \in \mathcal{S}$ (with $\tilde{\Omega} \neq \Omega$, where \neq means “not equivalent” or rather, here, “not iso-expressive”) such that $\tilde{\Omega} \leq \Omega$.

believes that $\mathcal{S}_j \subseteq \underline{\mathcal{S}}_i \cup \Omega^\emptyset$ and $|\mathcal{S}_j| = 1$. Hence, the assumption of independence across state spaces $\Omega \in \underline{\mathcal{S}}$ makes it possible that *player i's prior belief* $\tau_i(\widehat{\Omega})$ *about j being aware of* $\widehat{\Omega}$ is derived from v as follows:

$$\tau_i(\widehat{\Omega}) := \begin{cases} \frac{v(\widehat{\Omega}) \prod_{\Omega \in \underline{\mathcal{S}}_i \setminus \widehat{\Omega}} (1 - v(\Omega))}{\sum_{\widetilde{\Omega} \in \underline{\mathcal{S}}_i} [v(\widetilde{\Omega}) \prod_{\Omega \in \underline{\mathcal{S}}_i \setminus \widetilde{\Omega}} (1 - v(\Omega))] + \prod_{\Omega \in \underline{\mathcal{S}}_i} (1 - v(\Omega))} & : \widehat{\Omega} \in \underline{\mathcal{S}}_i, \\ \text{not defined} & : \widehat{\Omega} \notin \underline{\mathcal{S}}_i \cup \Omega^\emptyset, \\ \frac{\prod_{\Omega \in \underline{\mathcal{S}}_i} (1 - v(\Omega))}{\sum_{\widetilde{\Omega} \in \underline{\mathcal{S}}_i} [v(\widetilde{\Omega}) \prod_{\Omega \in \underline{\mathcal{S}}_i \setminus \widetilde{\Omega}} (1 - v(\Omega))] + \prod_{\Omega \in \underline{\mathcal{S}}_i} (1 - v(\Omega))} & : \widehat{\Omega} = \Omega^\emptyset, \underline{\mathcal{S}}_i \neq \emptyset, \\ 1 & : \widehat{\Omega} = \Omega^\emptyset, \underline{\mathcal{S}}_i = \emptyset. \end{cases} \quad (1.4.3)$$

A few comments are in order. First, it should be noticed that formula (1.4.3) implies that *player i* presumes the other player not to be aware of state spaces weakly more expressive than those in $\underline{\mathcal{S}}_i \cup \Omega^\emptyset$; besides, the first line of the formula shows that, for all the spaces in $\underline{\mathcal{S}}_i$, $\tau_i(\widehat{\Omega})$ is given by the ratio of the “objective probability of being aware of $\widehat{\Omega}$ only” to the “sum of the objective probabilities of being aware of (only) each of the spaces in $\underline{\mathcal{S}}_i$ and the probability of being aware of Ω^\emptyset only”.⁴¹ (This ensures that the $\tau_i(\widehat{\Omega})$ sum

⁴¹ The “probability of being aware of Ω^\emptyset only” is given by the product of the objective probabilities of being unaware of all the spaces in $\underline{\mathcal{S}}_i$.

to unity for $\forall \widehat{\Omega} \in \underline{\mathcal{S}}_i \cup \Omega^\emptyset$.) Further, the third line of the formula gives the value of $\tau_i(\widehat{\Omega} = \Omega^\emptyset)$ when player i is aware of at least one space weakly more expressive than Ω^\emptyset (*i.e.*: when $\underline{\mathcal{S}}_i$ is non-empty), while the fourth line gives the value of $\tau_i(\widehat{\Omega} = \Omega^\emptyset)$ when player i is unaware of any space weakly more expressive than Ω^\emptyset (*i.e.*: when $\underline{\mathcal{S}}_i$ is empty).

The interpretation is the following. Suppose that G is played many times between players i and j , with players drawn at random from large populations: then, $\tau_i(\cdot)$ may be interpreted as the statistical distribution of the “mixed strategies $\rho_j(\lambda^t)$ respecting remark I.1 and assumption I.1 for the available λ^t ”, in a population of agents playing G in the role of player j (in other words, $\tau_i(\cdot)$ may be thought of as the distribution of the basic vocabularies, hence the distribution of the strategies labelled according to the available $t \in T$, hence the distribution of $\rho_j(\lambda^t)$).⁴² In fact, if these statistical distributions are observed, each $\tau_i(\widehat{\Omega})$ may also represent i 's probabilistic belief about player j respecting remark I.1 and assumption I.1 for the available λ^t , in a play of G . It should be highlighted that if the agent who is drawn to play in the role of player j happened to be aware indeed of state spaces in $\mathcal{S}_j \subseteq \mathcal{S}_i$ – yet of spaces weakly more expressive than those in $\underline{\mathcal{S}}_i \cup \Omega^\emptyset$ (*e.g.*: in other words, if j herself were aware of Ω^{BC}) – then by

⁴² Obviously, for each set $\lambda^t(\mathcal{S}_j)$ there is an infinite number of mixed strategies $\rho_j(\lambda^t)$ respecting remark I.1 and assumption I.1 (for some given t). Yet, for simplicity, one can think of the belief about $\rho_j(\lambda^t)$ as a degenerate point belief, for each $t \in T$: as it will soon be clear, expected utility maximization will imply that it is so (*i.e.*: for each $t \in T$, individuals play one and only one profile of mixed strategies).

disregarding $\tau_i(\Omega^{BC})$ little would be lost: in effect, although j will actually be able to describe the current situation according to each of the basic vocabularies (properties) available in $\underline{\mathcal{S}}_i$, she will still have to play a mixed strategy respecting remark I.1 and assumption I.1 for *one* of the available λ^t *only*; as a consequence, from i 's perspective, j will be expected to play a mixed strategy respecting remark I.1 and assumption I.1 for one of the λ^t which the vocabularies available in $\underline{\mathcal{S}}_i \cup \Omega^\emptyset$ permit to define, as captured by formula (1.4.3) above.⁴³ Therefore, $\tau_i(\Omega^t)$ is interpreted as “player i 's prior belief about player j respecting remark I.1 and assumption I.1 for λ^t , in a play of G ”.

Example: Choose an Object (cont'd). Recall that $\mathcal{S} := \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$, which implies that $\underline{\mathcal{S}} = \{\Omega^B, \Omega^C\}$; further, if player i is aware of both block orderings and colour differences, then $\mathcal{S}_i = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$ and $\underline{\mathcal{S}}_i = \{\Omega^B, \Omega^C\}$. Now, the above assumption about the players' naïveté implies that player i will presume that player j must be aware of either Ω^\emptyset or Ω^B or Ω^C (but not of a union of Ω^B and Ω^C). In the case in which the agent who is drawn to play in the role of player j happened to be aware of state spaces in $\mathcal{S}_j \subseteq \mathcal{S}_i$ – yet of spaces weakly more expressive than those in $\underline{\mathcal{S}}_i \cup \Omega^\emptyset$ (i.e.: Ω^{BC}) – then she will still have to play a mixed strategy respecting remark I.1 and assumption I.1 for *one* of the available λ^t *only*: for instance, if player j happens to be

⁴³ On a different note it should be stressed that, obviously, in the case in which it is possible for some j in the population to be aware of state spaces *not* in \mathcal{S}_i (i.e.: if $\mathcal{S}_j \not\subseteq \mathcal{S}_i$), then the notion of unawareness implies that (for some player i) τ_i will not be defined at those spaces.

aware of both Ω^B and Ω^C (i.e.: $\mathcal{S}_j = \mathcal{S}_i = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$), then from i 's perspective j will be expected to play a mixed strategy respecting remark I.1 and assumption I.1 for one of the λ^t which the vocabularies available in $\underline{\mathcal{S}}_i \cup \Omega^\emptyset$ permit to define, that is, either $\rho_j(\lambda^C)$ or $\rho_j(\lambda^B)$ or $\rho_j(\lambda^\emptyset)$.

Before proceeding it should be noted that (while the reason for the above simplifying assumption is that the introduction of salience $(\succsim^t)_{t \in T}$ allows the comparison of pairs of strategies, each expressed as an instance of an *exclusive* property $t \in T$) such an assumption is likely to make no difference to expected utility maximizing behaviour, besides sparing us some complications.

Given that, I can now turn to define the expected payoff to player i , as follows.

Definition I.7. Given a strategic form game G and the associated knowledge structure $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_\Omega^\Omega\}, (\mathcal{S}_i)_{i \in N} \rangle$, frame $\langle \Omega^*, \varphi, \Lambda, T, (\Pi_t)_{t \in T} \rangle$, salience relations $(\succsim^t)_{t \in T}$, and distribution v , the *expected payoff* to player i is defined as:

$$E_{\sigma_i, \rho_{-i}, v}[u_i | \mathcal{S}_i] = \sum_{\Omega \in \underline{\mathcal{S}}_i} \sum_{t \in T: t \neq \emptyset} \tau_i(\Omega) u_i(\sigma_i, \rho_{-i}(\lambda^t)) + \tau_i(\Omega^\emptyset) u_i(\sigma_i, \rho_{-i}(\lambda^\emptyset)),$$

where σ_i is a generic mixed strategy of player i , and $\rho_{-i}(\lambda^t)$ is a profile of strategies of all players other than i respecting remark I.1 and assumption I.1 for some λ^t (with $\rho_{-i}(\lambda^t) = (\rho_j(\lambda^t))_{j \in N: j \neq i}$ and $\rho_{-i}(\lambda^\emptyset) = (\rho_j(\lambda^\emptyset))_{j \in N: j \neq i}$).

It is clear that the expected payoff to player i depends on her perception (hence labelling) and salience comparison, which implies that players with different knowledge functions face different games. In a nutshell, the second term of the above expected utility function represents the share of the payoff due to i 's belief about her co-player(s) being unaware of any state space more expressive than the uninformative Ω^\emptyset : as a result, (before playing i presumes that) every other player completely randomizes as imposed by remark I.1 on the case where $\Omega = \Omega^\emptyset$ (i.e.: $\rho_{-i}(\lambda^\emptyset)$). Then, the first term represents the share of the payoff due to i 's belief about her co-player(s) being aware of (only) each of the spaces in $\underline{\mathcal{S}}_i$ (which is the reason for the summation across $\Omega \in \underline{\mathcal{S}}_i$); notice that, for each $\Omega \in \underline{\mathcal{S}}_i$, players are required to take a mixed strategy respecting remark I.1 and assumption I.1 for $\lambda^t \neq \lambda^\emptyset$ (i.e.: $\rho_{-i}(\lambda^t)$, which is the reason for the summation across $\forall t \in T: t \neq \emptyset$). In this regard, it should be highlighted that – for each space $\Omega \in \underline{\mathcal{S}}$ – there exists only one property $t \neq \emptyset$ (with $t \in T$) according to which λ^t may express strategies: this implies that only one $\lambda^t \neq \lambda^\emptyset$ is actually defined for each state $\omega \in \Omega$, with $\Omega \in \underline{\mathcal{S}}$. It follows that in the first term of definition I.7, for each space $\Omega \in \underline{\mathcal{S}}_i$, $\rho_{-i}(\lambda^t)$ denotes a profile of strategies of all players $j \neq i$ respecting remark I.1 and assumption I.1 for the only one label function $\lambda^t \neq \lambda^\emptyset$ which is actually defined.

Example: Choose an Object (cont'd). Given $S = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$, recall that the set of properties is defined as $T := \{\emptyset, B, C\}$, where $t = B$ is the “order” property while $t = C$ is the “colour” property; $t = \emptyset$ is the “null” property, which does not explain anything. The set of labels is defined as $\Lambda := \{\lambda_\omega\}_{\omega \in \Omega^*}$ where, for each $\omega \in \Omega^*$, $\lambda_\omega := (\lambda_\omega^\emptyset, \lambda_\omega^B, \lambda_\omega^C)$. Again, it should be stressed that if

ω belongs to some Ω not rich enough to characterize a particular property t' , then $\lambda_{\omega}^{t'}$ is not defined at S_i , e.g.: if some state ω belongs to Ω^B , then λ_{ω}^C is not defined at S_i (similarly, if $\omega \in \Omega^C$, λ_{ω}^B is not defined at S_i).

To make the analysis more interesting, I shall now return to one of the cases considered in section I.3. above, namely the case where the experimenter randomly draws the objects from the bag in a way as captured by the following state: $\omega = (1g, 3r, 5b)$, with $\omega \in \Omega^{BC}$; more specifically, $\omega = (1g, 3r, 5b)$ is the state associated with the case where Nature has selected the strategic game with set of strategies $S_i = \{s_{i1}, s_{i3}, s_{i5}\}$ for $\forall i \in N$. Besides, let $(\succsim^t)_{t \in T}$ be defined as before, i.e.: *red* \succsim^C *grey* \succsim^C *black* and *first* \succsim^B *second* \succsim^B *third*.

Recall that $\rho_j(\lambda^t)$ denotes a generic mixed strategy of player j respecting remark I.1 and assumption I.1 for some λ^t . Now, considering the projection of ω into Ω^C , in this case one gets $\rho_j(\lambda_{\omega_{\Omega^C}}^C) \equiv (p(s_{j1}), p(s_{j3}), p(s_{j5}))$, where $p(s_{j3}) \geq p(s_{j1}) \geq p(s_{j5})$ (with $p(s_{j1}) + p(s_{j3}) + p(s_{j5}) = 1$); on the other hand, considering the projection of ω into Ω^B , one gets $\rho_j(\lambda_{\omega_{\Omega^B}}^B) \equiv (p(s_{j1}), p(s_{j3}), p(s_{j5}))$, where $p(s_{j1}) \geq p(s_{j3}) \geq p(s_{j5})$ (with $p(s_{j1}) + p(s_{j3}) + p(s_{j5}) = 1$); moreover, considering the projection of ω into Ω^{\emptyset} , one gets $\rho_j(\lambda_{\omega_{\Omega^{\emptyset}}}^{\emptyset}) \equiv (p(s_{j1}), p(s_{j3}), p(s_{j5}))$, where $p(s_{j1}) = p(s_{j3}) = p(s_{j5}) = \frac{1}{3}$.

Further, in order to simplify the notation, I shall drop the player-index and simply write p_1, p_3, p_5 for $p(s_{j1}), p(s_{j3}), p(s_{j5})$ ($\forall j \in N$), respectively. Thus, it is clear that – if j is aware of Ω^C only – a mixed strategy of player j respecting remark I.1 and assumption I.1 for $\lambda_{\omega_{\Omega^C}}^C$ can be written as $\rho_j(\lambda_{\omega_{\Omega^C}}^C) = (p_1, p_3, p_5)$, where $p_3 \geq p_1 \geq p_5$; then, assuming the game is played between two players i, j , the expected payoff to both players is

maximized only when players i and j play the same mixed strategy, and hence amounts to $(p_1)^2 + (p_3)^2 + (p_5)^2$, which in turn implies that the mixed strategy that uniquely maximizes player i 's payoff subject to the above constraint is given by the vector of probabilities $(p_1, p_3, p_5) = (0, 1, 0)$ (*i.e.*: more explicitly, in definition 1.7, if $\tau_i(\Omega^C) = 1$ the payoff to player i is maximized when $\sigma_i = \rho_j(\lambda_{\omega_{\Omega^C}}^C) \Rightarrow \sigma_i = (0, 1, 0)$). On the other hand – if j is aware of Ω^B only – a mixed strategy of player j respecting remark 1.1 and assumption 1.1 for $\lambda_{\omega_{\Omega^B}}^B$ can be written as $\rho_j(\lambda_{\omega_{\Omega^B}}^B) = (p_1, p_3, p_5)$, where $p_1 \geq p_3 \geq p_5$; then, the expected payoff to both players is again maximized only when players i and j play the same mixed strategy, and hence amounts to $(p_1)^2 + (p_3)^2 + (p_5)^2$, which here implies that the mixed strategy that uniquely maximizes player i 's payoff subject to the above constraint is given by the vector of probabilities $(p_1, p_3, p_5) = (1, 0, 0)$ (*i.e.*: in definition 1.7, if $\tau_i(\Omega^B) = 1$ the payoff to player i is maximized when $\sigma_i = \rho_j(\lambda_{\omega_{\Omega^B}}^B) \Rightarrow \sigma_i = (1, 0, 0)$). Moreover – if j is aware of Ω^\emptyset only – the mixed strategy of player j respecting remark 1.1 and assumption 1.1 for $\lambda_{\omega_{\Omega^\emptyset}}^\emptyset$ can be written as $\rho_j(\lambda_{\omega_{\Omega^\emptyset}}^\emptyset) = (p_1, p_3, p_5)$, which here implies that the mixed strategy that uniquely maximizes player i 's payoff is $\sigma_i = \rho_j(\lambda_{\omega_{\Omega^\emptyset}}^\emptyset) \Rightarrow \sigma_i = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

Now, assume that $\mathcal{S}_i = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$, hence $\underline{\mathcal{S}}_i = \{\Omega^B, \Omega^C\}$; then, the assumption about the players' naïveté in forming beliefs about the opponents' (un)awareness implies that, before playing, i will presume that player j must be aware of either Ω^\emptyset or Ω^B or Ω^C (but not of a union of Ω^B and Ω^C). It is clear that player i 's expected payoff is given by: $E_{\sigma_i, \rho_{-i}, v}[u_i | \mathcal{S}_i] = \sum_{\Omega \in \underline{\mathcal{S}}_i} \sum_{t \in T: t \neq \emptyset} \tau_i(\Omega) u_i(\sigma_i, \rho_{-i}(\lambda^t)) + \tau_i(\Omega^\emptyset) u_i(\sigma_i, \rho_{-i}(\lambda^\emptyset)) \equiv \tau_i(\Omega^C) u_i(\sigma_i, \rho_{-i}(\lambda_{\omega_{\Omega^C}}^C)) + \tau_i(\Omega^B) u_i(\sigma_i, \rho_{-i}(\lambda_{\omega_{\Omega^B}}^B)) +$

$\tau_i(\Omega^\emptyset)u_i\left(\sigma_i, \rho_{-i}\left(\lambda_{\omega, \Omega^\emptyset}^\emptyset\right)\right)$. Therefore, the mixed strategy of player i that uniquely maximizes $E_{\sigma_i, \rho_{-i}, v}[u_i | \mathcal{S}_i]$ is as follows:

- $\sigma_i = (0,1,0)$ for $\tau_i(\Omega^C) > \tau_i(\Omega^B)$ and $\tau_i(\Omega^C) > \frac{1}{3}\tau_i(\Omega^\emptyset)$;
- $\sigma_i = (1,0,0)$ for $\tau_i(\Omega^B) > \tau_i(\Omega^C)$ and $\tau_i(\Omega^B) > \frac{1}{3}\tau_i(\Omega^\emptyset)$;
- $\sigma_i = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ for $\tau_i(\Omega^\emptyset) > 3\tau_i(\Omega^C)$ and $\tau_i(\Omega^\emptyset) > 3\tau_i(\Omega^B)$;
- $\sigma_i = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$ for $\tau_i(\Omega^C) = \tau_i(\Omega^B) > \frac{2}{3}\tau_i(\Omega^\emptyset)$.

For instance, consider the case of $v(\Omega)$ being defined as $v(\Omega^B) = \frac{1}{4}$, $v(\Omega^C) = \frac{1}{2}$. Formula (1.4.3) above implies:

$$\tau_i(\Omega^C) \equiv \frac{\frac{1}{4} \frac{3}{2}}{\frac{1}{8} + \frac{3}{8} + \frac{3}{8}} = \frac{3}{7};$$

$$\tau_i(\Omega^B) \equiv \frac{\frac{1}{4} \frac{1}{2}}{\frac{1}{8} + \frac{3}{8} + \frac{3}{8}} = \frac{1}{7};$$

$$\tau_i(\Omega^\emptyset) \equiv \frac{\frac{3}{4} \frac{1}{2}}{\frac{1}{8} + \frac{3}{8} + \frac{3}{8}} = \frac{3}{7}.$$

It follows that, in this case, the mixed strategy of player i that uniquely maximizes $E_{\sigma_i, \rho_{-i}, v}[u_i | \mathcal{S}_i]$ is $\sigma_i = (0,1,0)$, that is, i chooses the red object.⁴⁴

⁴⁴ Notice that, as mentioned above, players with different knowledge functions face different games. Here this means that if player i is unaware of the block orderings, then $\mathcal{S}_i = \{\Omega^\emptyset, \Omega^C\}$

The above example shows that player i 's best response (to the expected strategies of player j respecting remark I.1 and assumption I.1) varies with the objective probability $v(\Omega)$ of a player being aware of some space $\Omega \in \underline{\mathcal{S}}$.

I.5. Conventions as equilibria

This section defines a “convention” as an equilibrium of a game with which a frame and a set of salience relations are associated.

Definition I.8. Given a strategic form game G and the associated knowledge structure $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_\Omega^\Omega\}, (\mathcal{S}_i)_{i \in N} \rangle$, frame $\langle \Omega^*, \varphi, \Lambda, T, (\Pi_t)_{t \in T} \rangle$, salience relations $(\succsim^t)_{t \in T}$, and distribution v , a *convention* of G is a profile of mixed strategies $\sigma^* = (\sigma_i^*)_{i \in N}$ such that for $\forall i \in N$:

$$\sigma_i^* \in \arg \max_{\sigma_i \in \Delta(\mathcal{S}_i)} E_{\sigma_i, \rho_{-i}, v} [u_i | I_i(\omega)].$$

In plain words, a convention is in place if every player $i \in N$ maximizes $E_{\sigma_i, \rho_{-i}, v} [u_i | I_i(\omega)]$, given her knowledge function: I give a dual interpretation.

and $\underline{\mathcal{S}}_i = \{\Omega^C\}$. Hence, in this case, player i 's expected payoff is simply given by:

$$E_{\sigma_i, \rho_{-i}, v} [u_i | \mathcal{S}_i] = \sum_{\Omega \in \underline{\mathcal{S}}_i} \sum_{t \in T: t \neq \emptyset} \tau_i(\Omega) u_i(\sigma_i, \rho_{-i}(\lambda^t)) + \tau_i(\Omega^\emptyset) u_i(\sigma_i, \rho_{-i}(\lambda^\emptyset)) \equiv \\ \tau_i(\Omega^C) u_i(\sigma_i, \rho_{-i}(\lambda_{\omega_{\Omega^C}}^C)) + \tau_i(\Omega^\emptyset) u_i(\sigma_i, \rho_{-i}(\lambda_{\omega_{\Omega^\emptyset}}^\emptyset)).$$

Conventions as the result of a game of incomplete information with “blind” players. Again, suppose that G is played many times between i and j , with players drawn at random from large populations: as before, $\tau_i(\cdot)$ is thought of as the statistical distribution of the “mixed strategies $\rho_j(\lambda_j)$ ” respecting remark I.1 and assumption I.1 for the available λ^t , in a population of agents playing G in the role of player j ; besides, each $\tau_i(\hat{\Omega})$ also represents i 's prior belief about player j respecting remark I.1 and assumption I.1 for the available λ^t , in a play of G . Now – assuming that an individual does *not* get to see who the matched co-player is until after having played – a convention is in place if every player maximizes $E_{\sigma_i, \rho_j, v}[u_i | I_i(\omega)]$, given her beliefs $(\tau_i(\Omega))_{\Omega \in \underline{\mathcal{S}}_i \cup \Omega^\emptyset}$: since i 's way of deriving $(\tau_i(\Omega))_{\Omega \in \underline{\mathcal{S}}_i \cup \Omega^\emptyset}$ from an exogenously given $(v(\Omega))_{\Omega \in \underline{\mathcal{S}}}$ depends on $\underline{\mathcal{S}}_i$, it follows that a convention is implemented when every pair of matched players is characterized by the same knowledge function.

Conventions as the result of a game with differently-aware players. Consider an awareness operator $A_i(E) = \mathcal{K}_i(E) \cup \mathcal{K}_i(\sim \mathcal{K}_i(E))$, where the knowledge operator is derived from a generalized information function I_i respecting properties $(g.i.i-vi)$, as in section I.4.a. above. Similarly to the “everybody knows” and the “common knowledge” operators, one can define “everybody is aware” and “common awareness” operators. (Indeed, Heifetz *et al.* [2006] prove that when everybody is aware of an event E , then everybody is also aware that everybody is aware of E : it follows that the events “everybody is aware of E ” and “common awareness of E ” coincide.) Now – assuming that an individual *does* get to see who the matched co-player is before playing – a convention is in order if every player maximizes $E_{\sigma_i, \rho_j, v}[u_i | I_i(\omega)]$, given her beliefs $(\tau_i(\Omega))_{\Omega \in \underline{\mathcal{S}}_i \cup \Omega^\emptyset}$: yet, contrary to the previous case, here player i has access to j 's information, therefore she can use the

awareness operator in a way similar to the knowledge operator so as to define what “ i is aware of j being aware of”. For example, in the game of Choose an Object, let player i be aware of both block orderings and colour differences (*i.e.*: $\mathcal{S}_i = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$ and $\underline{\mathcal{S}}_i = \{\Omega^B, \Omega^C\}$) while player j be aware of colour differences only (*i.e.*: $\mathcal{S}_j = \{\Omega^\emptyset, \Omega^C\}$ and $\underline{\mathcal{S}}_j = \{\Omega^C\}$). Using the awareness operator, it follows that $A_i(\sim A_j(\Omega^B))$: as a consequence, i can be thought of as updating her beliefs so that $\hat{\tau}_i(\Omega^C) = 1$; it is clear that, in this case, a convention corresponds to the profile of mixed strategies $(\rho_i(\lambda^C))_{i \in N}$ respecting remark I.1 and assumption I.1 for λ^C such that $(\sigma_i = (0,1,0))_{i \in N}$. It should be noticed that here, contrary to the previous case, for a convention to be in operation it is not necessary that every pair of players is characterized by the same knowledge function, but just that every player is commonly aware of a state space, that is, $\mathcal{S}_i \cap \mathcal{S}_j \neq \emptyset$.⁴⁵ Once again, it should be stressed that it would be impossible to model such a situation by using standard information structures in which the knowledge function only satisfies the standard properties (*k.i-vi*) described in section I.2.c. above: as a matter of fact, in a standard single-space information structure there will always be common knowledge of the unique state space.

⁴⁵ In the case in which $\mathcal{S}_i \cap \mathcal{S}_j \neq \emptyset$ and there is *not* a *unique* optimal convention (*i.e.*: as in the case where $\mathcal{S}_i = \mathcal{S}_j = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$ for a given pair of players (i, j)), then one may assume that individuals play the mixed strategy respecting remark I.1 and assumption I.1 with strategies labelled and ranked according to the property (or basic vocabulary) t that has maximum *prior* probability $\tau_i(\Omega^t)$.

The following proposition applies to the latter of the above interpretations/settings, thereby relating conventions to Aumann's [1974, 1987] notion of correlated equilibrium.

Proposition I.1. Take a strategic form game G and the associated knowledge structure $\langle K, (\Omega^k, q^k)_{k \in K}, \preceq, \{r_\Omega^\Omega\}, (\mathcal{S}_i)_{i \in N} \rangle$, frame $\langle \Omega^*, \varphi, \Lambda, T, (\Pi_t)_{t \in T} \rangle$, and salience relations $(\succsim^t)_{t \in T}$. Let $\sigma_i^t(\omega) \in \arg \max_{\sigma_i \in \Delta(S_i)} E_{\sigma_{-i}, \rho_{-i}, v} [u_i | I_i(\omega^*)]$ denote an optimal strategy when strategy-indices are expected to be labelled and ranked according to some $t \in T$, given that $\omega \in \Omega^*$ obtains. Let $\sigma^t(\omega^*)$ denote a convention defined by strategy profile $(\sigma_i^t(\omega^*))_{i \in N}$: the set of all such strategy profiles (one for each $(I_i(\omega))_{i \in N}$) is a *correlated equilibrium* of G .

Proof. Consider a game (*e.g.*: Choose an Object, with $S = \{\Omega^\emptyset, \Omega^B, \Omega^C, \Omega^{BC}\}$), where every pair of matched players is characterized by the same knowledge function: for simplicity, assume that every pair of players is aware of either $t = C$ or $t = B$ (*i.e.*: $\mathcal{S}_i = \mathcal{S}_j = \{\Omega^\emptyset, \Omega^C\}$ or $\mathcal{S}_i = \mathcal{S}_j = \{\Omega^\emptyset, \Omega^B\}$ for every pair (i, j)). Remark I.1 and assumption I.1, along with the definition of each player's expected payoff (definition I.7), imply that a convention is a profile of strategies in which a pair (i, j) plays, with probability one, the most salient labelled strategy (as ranked by \succsim^C or \succsim^B) in $(\lambda_\omega^C(S_i), \lambda_\omega^C(S_j))$ or $(\lambda_\omega^B(S_i), \lambda_\omega^B(S_j))$, respectively (depending on the state obtaining, and the players' "awareness type"). Denote a strategy profile given by a pair of such strategies by $\sigma^t(\omega)$, with $t \in \{B, C\}$. Then, it is straightforward to see that the set of all such strategy profiles (one for each $(I_i(\omega))_{i \in N}$, as strategies are such that $\sigma_i^t(\omega) = \sigma_i^t(\omega')$ whenever ω and ω' are in the same cell of the information partition) is a correlated equilibrium of G , which is simply defined

by $\langle (\Omega^t, q^t), (\mathcal{S}_i, \mathcal{S}_j), (\sigma_i^t, \sigma_j^t) \rangle$, where: (Ω^t, q^t) is a finite probability space; $(\mathcal{S}_i, \mathcal{S}_j)$ is a profile of partitions of $\Omega^t \cup \Omega^\emptyset$; (σ_i^t, σ_j^t) is a profile of decision functions, with $\sigma_i^t: \Omega^t \rightarrow \lambda^t(S_i)$ and $\sigma_j^t: \Omega^t \rightarrow \lambda^t(S_j)$. Note that, in order for $\langle (\Omega^t, q^t), (\mathcal{S}_i, \mathcal{S}_j), (\sigma_i^t, \sigma_j^t) \rangle$ to define a correlated equilibrium, the following inequality must hold for all players: $\sum_{\omega \in \Omega^t} q^t(\omega) u_i(\sigma_i^t(\omega), \sigma_j^t(\omega)) \geq \sum_{\omega \in \Omega^t} q^t(\omega) u_i(\tilde{\sigma}_i^t(\omega), \sigma_j^t(\omega))$; in other words, for every state $\omega \in \Omega^t$ (of which the probability is $q^t(\omega)$), the strategy $\sigma_i^t(\omega)$ must be optimal given the other player's strategy and i 's knowledge about ω . It is clear that the inequality holds when $\sigma_i^t(\omega)$ is given as in proposition 1.1, hence $u_i^t := \sum_{\omega \in \Omega^t} q^t(\omega) u_i(\sigma_i^t(\omega), \sigma_j^t(\omega))$ may be referred to as the correlated equilibrium payoff to player i (generated by conventions $(\sigma_i^t(\omega), \sigma_j^t(\omega))_{\omega \in \Omega^t}$); similarly, u_j^t is the correlated equilibrium payoff to player j . ■

It should be noted that, unlike Aumann's traditional notion of correlated equilibrium (where strategy choices are pegged on events defined on a single state space), here ω, ω' are actually the projections of states in Ω^{BC} into a weakly less expressive space Ω^t , with $t \in \{B, C\}$. Also, it should be stressed that Aumann's notion of correlated equilibrium assumes that players enter an agreement in the form of a pre-determined collection of decision functions; instead – given a set of strategies labelled and ranked according to some property (or basic vocabulary) t – for a convention to be in place players are required to respect remark 1.1 and assumption 1.1, with the strategies to be played being determined as the result of expected utility maximization. Interestingly, notice that the idea of conventions (in the sense of Lewis [1969]) defined as Aumann's correlated equilibria has been explored informally by philosopher Peter Vanderschraaf [1998]. Yet, it should be noted

that, since Vanderschraaf's application of the correlated equilibrium would require an explicit agreement between players, that may not capture a great deal of social phenomena which involve no explicit act of agreement. (In this connection, John Locke [1689] highlighted the notion of a "tacit agreement", with a tacit agreement occurring when there has been no explicit agreement but matters are otherwise *as if* an explicit agreement occurred.)

I.6. Concluding remarks

This study has presented an original theory of conformity in coordination games. The core of the problem has been to provide a framework for the player's own perception of the strategic situation so as to show that coordination may occur when "normal" players use, and expect others to use, similar conceptual schemes. Again, here it is suggested that, for a convention to be in operation, conformity is dependent on the states perceived by the agents: therefore, the model has implemented a notion of unawareness so as to account for multiple descriptions of the world. Thus, the present theory has defined the player's own framing system in such a way as to allow for the possibility that both stochastic and non-stochastic procedures may determine the labelling of strategies; besides, it has provided a precise link between a player's information function and her labelled strategies. Further, introducing a salience relation on each set of strategy labels, along with two requirements (relating to symmetry and salience) has resulted in the set of a player's mixed strategies being restricted. To sum up, this essay has suggested that conventions may arise as the result of a four-step procedure: (i) perception; (ii) labelling; (iii) salience comparison; (iv) expected utility maximization. Such a theory is consistent with the intuitions of Lewis [1969] and Schelling [1960] in that conventions are defined as solutions to

coordination games where there are multiple equilibria; also, certain characteristics of a strategy set as perceived by the player – which would not explicitly enter the formal description of a standard game – here can make some strategies more salient than others.

It should be noted that the game introduced during the exposition of the theory is indeed meaningful, as it shows that conventions are determined by the context in which the game itself appears and, in part, are the result of random eventualities. Moreover, it should be stressed that, in the spirit of Schelling [1960], such a theory can work even in the case of impure coordination games (or tacit bargaining problems). The intuition is confirmed by an experimental study (Mehta *et al.* [1992]) in which two subjects had to agree on how to divide £10 between them, with each bargainer receiving zero if no agreement was reached. Before the bargaining took place, subjects were dealt four playing cards each, from a deck consisting of 8 cards, including 4 aces and 4 twos. Subjects were told that a set of 4 aces was worth £10 (whereas any other combination of cards was worth nothing), and in order to be paid they had to pool their aces and agree on how to divide the £10; notice that only situations in which neither player held all 4 aces were considered in the analysis. Although Mehta *et al.* observed that equal divisions were the modal proposal by holders of 1, 2, or 3 aces, they also noted a tendency to give more to the bargainer with more aces: in effect, they recorded a second modal demand of only £2.50 by holders of 1 ace. Their suggestion is that the bargainers use the cards dealt to them as cues to help solve the coordination problem which lies at the heart of the bargaining game. Interestingly, reading their essay through the lens of the present theory, Mehta *et al.* seem to implicitly suggest that in their experiment players could be aware of (at least) two state spaces (*i.e.*: one containing three states, with each state corresponding to the number of aces held by a player;

the other containing only one state) associated with a salience relation, as follows:

We assumed that participants may perceive aces to be worth £2.50 each and our prior expectation was that subjects would use this perception in conjunction with one of two principles. The principle of “closeness” would lead subjects to the rule “distribute according to the value of the cards”. The principle of “equality” would lead to the rule “distribute by equal shares”. Thus, we would expect the predominant demands to be as follows: (p. 215)

		“Closeness”	“Equality”
Subjects holding	1 ace	£2.5	£5
	2 aces	£5	£5
	3 aces	£7.5	£5

Indeed, their hypothesis testing confirms that the average demand of subjects holding 1 ace is significantly lower than that of subjects holding 2 aces, which in turn is significantly lower than that of subjects holding 3 aces.

To conclude, such a game-theoretic study of conventions may contribute to shed some light on the mental *tâtonnements* enabling the convergence of players’ beliefs and the emergence of a stable pattern of behaviour in problems with many possible equilibria, thereby tackling the widespread equilibrium selection problem for which no satisfactory solution is yet known. Possible applications may involve a number of economic situations in which all parties can realize mutual gains, but only by making mutually consistent decisions: for example, think of the choice of technological standards (a product which becomes generally accepted and dominant is often considered a *de facto* standard, even without an established norm or requirement about technical systems, *e.g.*: the QWERTY keyboard) or the emergence of aesthetic rules and trends (as a matter of fact, every year or so a different colour is made salient by the fashion industry).

II.A Theory of Belief-Dependent Conformity to Social Norms

II.1. Introduction

Socio-economic behaviour is generally modelled on rational choice theory's prescriptions: economic theory assumes that an agent has preferences satisfying some rationality requirements, yet most traditional economic applications simply view those requirements as implying that the self-interest of the agent is narrowly *self-centred* and unaffected by the others' outcome. On the other hand, the widely documented regularities of behaviour inconsistent with the standard predictions of models with rational self-centred individuals have motivated alternative accounts. Everyday life examples of such "incidents" might be brought about by norms that informally prescribe how people ought to behave in the community or workplace, and which are enforced out of fear of social sanctions: Arrow's [1972] investigation suggests that entrepreneurs, who could turn a profit on hiring labour cheaply from a racially discriminated group, were restrained from doing so owing to the establishment of social customs involving discriminatory tastes; or rather, as Akerlof [1980] claims, if the custom prohibits an employer from hiring labour at a reduced wage, employees will not cooperate in training new workers (who undercut existing wages), because by doing so they would suffer a loss of reputation for participating in disobeying the norm. Other situations that may be explained by the enforcement of informal norms regulating social behaviour include the voluntary supply of public goods (Sugden [1984]) and reciprocity-based transactions such as gift-giving, *etc.* (Sacco *et al.* [2006]).

The above instances seem to be validated by a wealth of experimental evidence in *social dilemma* (*i.e.*: mixed-motive) games, which is thoroughly collected by Camerer [2003], Ch. 2; Fehr and Schmidt [2006]; Ledyard [1995]. Such paradigms as the Prisoner's Dilemma, the Ultimatum Game or the Trust Game are particularly meaningful since they clearly embody a tension – between individual incentives and other motivations – which can be

well observed in many everyday social interactions: indeed, the aforementioned experimental games all provide support against the traditional self-centred view of economic agents and its related *descriptive* predictions. In this connection, the present investigation focuses on the very motivation that makes people comply with default rules of behaviour when facing a social dilemma; this essay suggests that individuals may feel guilt at violating a norm (and this painful emotion generates conformity under precisely stated conditions), with a *norm* being modelled as a rule that dictates the strategy profile/s most “appropriate” for each decision node of a given mixed-motive game in extensive form.

According to a line of thought which has been popular especially among philosophers, mixed-motive games make it possible to highlight that, in some cases, traditional economic applications do not seem *normatively* adequate either (Gold and Sugden [2007; 2008]). A famous puzzle is suggested by the Prisoner’s Dilemma: there a positive theory (implicitly assuming that the agent only cares about her own material payoff) would forecast that both subjects *will* defect, whereas experimental evidence (collected in Sally [1995]) shows that almost half of the subjects in the laboratory cooperate;⁴⁶ besides, a normative theory – neutral on matters of social welfare – would suggest that both players *ought to* defect, however both would benefit from mutual cooperation. In a nutshell, mixed-motive games raise some problems since the “conventional” fashion, in which

⁴⁶ For a review of Prisoner’s Dilemma experiments, see also: Colman [1995], Ch. 9; Cooper *et al.* [1996]; Davis and Holt [1993], Ch. 9; Goeree and Holt [2001]; Ledyard [1995]. Early experiments are more extensively discussed in: Lave [1962]; Rapoport and Chamah [1965], Pt. 1.

economic applications are modelled, may appear to be unconvincing with regard to both its positive and normative facet. In this respect, it should be stressed that game theory (in its classic, normative form) is certainly effective in prescribing strategies that maximize the individual player's payoff, even though in most mixed-motive games it fails to induce collectively desirable outcomes; on the other hand, most economists reply to this argument by pointing out that there is no such normative problem or, equivalently, that the above analysis would be correct only if the Prisoner's Dilemma were a game with just one player (Binmore [2007], Ch. 19). Now, whichever view is correct, the Prisoner's Dilemma leaves us with a problem of externalities, and a consequent policy-modelling problem: indeed, in most economic applications, the precise tools of analytical game theory are used to *describe* and *prescribe* the best policy options for actual economic agents, with the unfortunate result of promoting socially inefficient allocations.⁴⁷

⁴⁷ The one-shot Prisoner's Dilemma is an example of a game with a unique Pareto-inefficient Nash equilibrium; likewise, the only subgame perfect equilibrium of the finitely-repeated Prisoner's Dilemma, as every Nash equilibrium, consists of strategies that prescribe defection in every period (again, it is Pareto-inefficient). Similarly other puzzles of game theory, generally referred to as social dilemmas (including the Investment Game, the Centipede Game, *etc.*), all present a unique Pareto-inefficient subgame perfect equilibrium – with the exception of the Ultimatum Game. In effect, the Ultimatum Game is an example of a mixed-motive game where the only subgame perfect equilibrium (*i.e.*: the strategy pair in which the "Proposer" offers the smallest amount and the "Responder" accepts all offers) is *not* Pareto-inefficient: yet in evaluating economic allocations, other mechanisms (*e.g.*: other social norms, perhaps involving issues of welfare distribution) might arise; as a matter of fact, in the laboratory Proposers' offers average 40% of the money, while Responders reject offers of about 20% of the money half the time (Camerer [2003], Ch. 2).

In response to the shortcomings associated with the conventional economic model of rational self-centred agents, (relatively) recent developments in game theory improve the analysis of strategic interaction by allowing for diverse assumptions about players' beliefs, emotions, preferences, and rationality: most of these models have been designed to tackle experimental evidence, otherwise inexplicable, and address the positive facet of the problem.⁴⁸ One line of research explains experimental regularities about "other-regarding" behaviour by suggesting adjusted utility functions, which allow the individual's welfare to reflect a range of motivations besides narrow self-interest. Some of the *social preference* theories, namely the so-called models of "reciprocal fairness", seem to be most effective in accounting for other-regarding behaviour where intentions matter: think of Rabin [1993], Dufwenberg-Kirchsteiger [2004], Falk-Fischbacher [2006], Charness-Rabin [2002]; all such intention-based models – with the exception of Battigalli and Dufwenberg's [2007] framework – *explicitly* assume that players have a preference for a somehow specified equitable payoff (Rabin [1993], Dufwenberg-Kirchsteiger [2004]) or they are just (intention-based) inequity averse (Falk-Fischbacher [2006]⁴⁹) or they have a taste for both

⁴⁸ One exception being the body of literature rejecting the notion of individual rationality, which addresses the normative facet by hinging on alternative concepts of rationality (*e.g.*: collective rationality, rational commitment, *etc.*), thereby allowing groups of individuals to count as agents (Bacharach [1999]; Gold and Sugden [2007; 2008]; Sugden [2003]) or rely on moral arguments which prescribe a certain behaviour, even if it is not in one's self-interest to do so (Collard [1983]; Harsanyi [1980]; Laffont [1975]; Sugden [1984]).

⁴⁹ Falk and Fischbacher [2006] do not define "kindness" in terms of the feasible payoffs of Player *i* in relation to an equitable payoff, but directly in relation to the payoff that Player *j* gets: in this respect, their model can therefore be viewed as an *intention-based* inequity

fairness and efficiency (captured by quasi-maximin preferences in Charness-Rabin [2002]). Thus, the aforementioned models may be interpreted as more or less *implicitly* assuming that players have internalized a (variously defined) “social norm of fairness or reciprocity”.

In a different perspective, surveys from diverse disciplines – including cognitive psychology and neuroscience⁵⁰ – support the view that human conduct is often guided by “conformist preferences”, which thrive on behavioural expectations within a society or group, with conformity being the act of changing one’s behaviour to match the purported beliefs of others (Cialdini and Goldstein [2004]). To that end, the present essay takes the investigation of other-regarding motives in social dilemmas one step further: despite a growing body of literature considering preferences for a fair outcome allocation among players, most economic theories do not explain the underpinning *conditions for a social norm* to exist and to be in operation among *players with conformist motivations*. Therefore, inspired by Cristina Bicchieri’s [2006] philosophical account of social norms,⁵¹ here I develop an original model of conformist preferences in mixed-motive games, building on the guilt aversion framework of Battigalli and Dufwenberg [2007].

In what follows, I will define what are the mechanics of those informal norms regulating social behaviour; I will further maintain that *social* norms are

aversion theory (as opposed to a *simple* inequity aversion theory *à la* Fehr and Schmidt [1999] or Bolton and Ockenfels [2000]).

⁵⁰ See: Klucharev *et al.* [2009]; for a scientific review of studies that use experimental games in combination with either PET scans or functional magnetic resonance imaging (fMRI), refer to Montague and Lohrenz [2007].

⁵¹ For another book-length philosophical treatment of norms, see Ullmann-Margalit [1977].

brought about by a certain type of (conditionally) conformist preferences; I will finally claim that different social norms may be made salient in different environments (depending on the individuals' beliefs about the "currently-normal" behaviour of other group-members). Two points naturally arise.

- (i) On psychological games: consistent with most sociological publications holding that social norms are necessarily sustained by expectations (Hechter and Opp [2001]), a psychological game framework *à la* Geanakoplos *et al.* [1989] (recently extended by Battigalli and Dufwenberg [2009]) appears to be most suitable to grasp situations where informal rules may play a major role; in this respect, the current study departs from López-Pérez's [2008] model of norm compliance while is in line with Li's [2008] (although the latter restricts attention to normal form games).⁵²
- (ii) On inducing socially desirable outcomes: if social norms vary – according to the individuals' expectations – the aforementioned policy-modelling problem could be resolved through some finely tuned process of belief manipulation. To that end, a full understanding of the mechanics of social norms is required, thereby allowing a neutral theoretical framework which can account for different (conjectures about) norms; in this regard, the current study leads to a generalization of the above theories of reciprocal fairness.

That being said, other considerations are in order. First of all, a quick note on methodology: this essay aims at explaining in game-theoretic terms why

⁵² Both models are extensively reviewed in the appendix (section II.8.b. below).

social norms emerge, assuming that players are (fully) rational utility-maximizers. This implies that social norms emerge because they yield “benefits” for the agents themselves; therefore, the study at hand departs both from those theories relaxing the assumption of common knowledge of rationality (*e.g.*: Kreps *et al.* [1982]; Fudenberg and Maskin [1986]⁵³) and those rejecting the notion itself of individual rationality (see footnote 48).

Another important issue concerns the evolution of norms. This essay aims at providing a framework for the analysis of social norms, without recourse to evolutionary arguments: under the hypothesis that, in the short run, one can treat the biological or cultural aspects of human nature as fixed (save for changes in the players’ own beliefs, of which they are obviously fully aware and accordingly adjust their strategies), I will not consider the dynamics of evolutionary change.⁵⁴ Instead, the present approach hinges on the idea that a social norm will be enforced if the actions prescribed by that norm do not allow, on the part of any player, a positive incentive to unilaterally deviate; in other words, a social norm will be enforced if the actions prescribed by that norm are supported by a refinement of the sequential equilibrium (allowing for belief-dependent conformist preferences).

In shaping a theory of conformist preferences in mixed-motive games, I will draw on the widely maintained view (Sugden [2000]; Elster [1989], Ch.

⁵³ Such theories explain behaviour in terms of reputation-building by assuming incomplete information about rationality, although they only account for experimental regularities specific to a narrow domain, *i.e.*: repeated games.

⁵⁴ Notice that, conversely, evolutionary game theory often assumes that players have no control over the strategy they play, they do not need to know the structure of the game, and may not even realize they are playing a game at all.

6) that individuals may feel *guilt* at violating a social norm, and this painful emotion would generate conformity even in the absence of external sanctioning; therefore, Battigalli and Dufwenberg's [2007] model of guilt aversion naturally lends itself to depict players with (conditionally) conformist preferences. To sum up, in what follows: I define a "norm" as a rule that dictates a set of strategy profiles; I assume that players, conditional on each history of an extensive form game, hold a conjecture about the active player's norm-complying actions available at that history; I then model a norm-driven decision maker as a player with conformist preferences, whose utility function is a linear combination of her material payoff and a component representing the social cost of deviating, in the form of the sum of losses that other conformist players would suffer because of a norm violation. A "social norm" is said to exist and to be followed by a population if players have *conditionally* conformist preferences, they hold correct beliefs about the strategies in line with some "normally-expected behaviour", and are sensitive enough to the potential social cost of deviating.

The remainder of the essay is organized in this manner: II.2. reviews Bicchieri's [2006] account of norms; II.3. introduces some general notation on extensive form games, and conditional systems of beliefs; II.4. formally lays out the model; II.5. discusses an equilibrium solution; II.6. provides some illustrations, and II.7. concludes.

II.2. Bicchieri's account of norms

This section reviews Bicchieri's [2006] philosophical account of norms, which lays the foundations for the model of conformity to be introduced later. Bicchieri's analysis starts by distinguishing "social norms" from "moral norms" and "descriptive norms", where social norms (as well as descriptive norms)

are necessarily sustained by expectations about the others' compliance while moral norms are not; furthermore, while descriptive norms (involving fashions or fads) have no intrinsic value, may eventually evolve into standard "conventions" and are used to solve a (pre-existing) pure *coordination game*, social norms usually apply to *mixed-motive games* only. In a nutshell, an established social norm can be seen as an informal, non-binding rule for choosing in a mixed-motive game, (transforming *this* into a coordination game) so that members of a population prefer to follow such a norm depending on whether they expect sufficiently many in the population to follow it; the "conditions for a social norm to exist" are described as follows. Let R be a behavioural rule for situations of type S , where S is a strategic interaction that can be represented as a mixed-motive game. Then, R is a *social norm* in a population P , if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that for each individual $i \in P_{cf}$ the below properties hold:⁵⁵

1. (contingency) i knows that a rule R exists and applies to situations of type S ;
2. (conditional preference) i prefers to conform to R in situations of type S , if
 - 2.1. (empirical expectations) i believes that a sufficiently large subset of P conforms to R in situations of type S ;

and either of the following

 - 2.2. a. (normative expectations) i believes that a sufficiently large subset of P expects i to conform to R in situations of type S ,

⁵⁵ In Bicchieri's view, different individuals may have different beliefs about what "sufficiently large" means, yet what matters to conformity is that each individual believes that her threshold has (at least) been reached.

2.2. *b.* (normative expectations with sanctions) *i* believes that a sufficiently large subset of *P* expects *i* to conform to *R* in situations of type *S*, prefers *i* to conform, and may sanction behaviour.

A social norm *R* (exists and) is *followed* by population *P*, if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that for each individual $i \in P_f$ conditions 2.1 and either 2.2.a or 2.2.b are met for *i* and, as a result, *i* prefers to conform to *R* in situations of type *S*.

A few comments are in order. First, note that P_{cf} denotes the set of *conditional followers* of *R* (*i.e.*: “individuals who know about *R* and have a conditional preference for conforming to *R*”), while P_f denotes the set of *followers* of *R* (*i.e.*: “individuals who know about *R* and have a preference for conforming to *R*”, because they believe that the conditions for their conditional preference are fulfilled). Hence, in Bicchieri’s view, *R* can be a social norm for a population *P*, even though it is not currently being followed by *P*: indeed, *R* is a social norm if P_{cf} is sufficiently large; *R* is also followed if P_f is sufficiently large.⁵⁶

In the appendix to Chapter 1 Bicchieri develops a general utility function based on norms. Considering an *n*-player normal form game, let S_i denote the strategy set of Player *i* and $S_{-i} = \prod_{j \neq i} S_j$ be the set of strategy profiles of players other than *i*. A norm N_i is defined as a (set-valued) function from one’s expectation about the opponents’ (norm-complying)

⁵⁶ This implies that *not* every conditional follower of a norm *R* eventually decides to conform to it, as some (conditional followers) might think that their expectations have not been fulfilled.

strategies to one's own strategies, that is, $N_i: L_{-i} \rightarrow S_i$, with $L_{-i} \subseteq S_{-i}$.⁵⁷ A strategy profile $s = (s_1, \dots, s_n)$ is said to instantiate a norm for Player j if $s_{-j} \in L_{-j}$ (i.e.: if N_j is defined at s_{-j}), and to violate a norm if $s_j \neq N_j(s_{-j})$. Player i 's utility function is a linear combination of i 's material payoff $\pi_i(s)$ and a component that depends on norm compliance:

$$U_i(s) = \pi_i(s) - k_i \max_{s_{-j} \in L_{-j}} \max_{m \neq j} \left\{ \pi_m(s_{-j}, N_j(s_{-j})) - \pi_m(s), 0 \right\}, \quad (2.2.1)$$

where $k_i \geq 0$ shows i 's sensitivity to the norm and j refers to the norm violator. The norm-based component represents the maximum loss (suffered by players other than the norm violator j) resulting from all norm violations: the first maximum operator aims at taking care of the possibility that there might be multiple norm-complying strategy profiles (a situation which does not occur in the example below, where it degenerates); the second maximum operator ranges over all the players other than the norm violator j . To sum up, the player's utility equals her own material payoff, minus a quantity corresponding to the norm followers' maximum loss resulting from the norm violation, multiplied by the player's sensitivity parameter (notice that in braces is the difference between what the most negatively-affected norm follower

⁵⁷ For example, in an n -player Prisoner's Dilemma a shared norm may be to cooperate: in that case, L_{-i} includes the *cooperate* strategies of all players other than i . Note that in the case where – given the others' strategies – there is not a norm prescribing how Player i should behave, then N_i is not defined at L_{-i} .

would get in case of norm compliance and what she actually gets). In order to illustrate the above utility function, consider a 2-player Prisoner's Dilemma and suppose that a norm of reciprocal cooperation has been established: then, the norm dictating cooperation is defined at C (*cooperate*) and undefined at D (*defect*). Now, if Player 1 violates the norm by choosing D – while Player 2 follows the norm – Player 1's utility will be

$$U_1(D, C) = \pi_1(D, C) - k_1[\pi_2(C, C) - \pi_2(D, C)].$$

Notice that, according to Bicchieri's theory, Player 1 suffers even in the case in which she conforms to the norm but the opponent does not; in such a case Player 1's utility is in fact

$$U_1(C, D) = \pi_1(C, D) - k_1[\pi_1(C, C) - \pi_1(C, D)].$$

To conclude, Bicchieri uses the above utility function to explain experimental results from a variety of social dilemmas, reading the differences in the observed rate of cooperation as changes in the players' sensitivity or in the players' beliefs about the opponents' sensitivity or changes in the relevant norm itself: this provides an insightful starting point for a dynamic model of norm compliance, which I am now to introduce.

II.3. Preliminaries

II.3.a. Notation on extensive form games

An extensive form game (with perfect recall) is given by the structure $\langle N, H, P, (\mathcal{S}_i)_{i \in N} \rangle$, where: $N = \{1, \dots, n\}$ is the *set of players*, H is the finite set

of feasible histories, P is the *player function*, \mathcal{S}_i is the *information partition* of Player i .

Each element of H is a history, which is a (finite) *sequence of actions* taken by the players: let $h(a^l)$ denote a sequence (a^1, \dots, a^l) , with a^l being the l -th action chosen along the game tree.⁵⁸ Let the set of histories satisfy the usual properties, *i.e.*: (i) the empty sequence (of length 0) h^0 is a member of H ; (ii) if $(a^l)_{l=1, \dots, L} \in H$ and $M < L$, then $(a^l)_{l=1, \dots, M} \in H$. Further, let Z denote the *set of terminal histories* (leading to the leaves, or end-nodes, of the game tree), with $H \setminus Z$ being the set of non-terminal histories; given that, let $A_i(h)$ denote the *set of feasible actions* for Player i at history h ($A_i(h)$ is singleton if Player i is not active at h , while $A_i(h)$ is empty if and only if h is a terminal history).

The *player function* P is defined, in the usual way, as a function that assigns to each element of $H \setminus Z$ an element of N , with $P(h)$ being the player choosing an action after the history h . Then, for each player $i \in N$, \mathcal{S}_i denotes the *information partition* of Player i – and $I_i \in \mathcal{S}_i$ is an information set of Player i – where a partition \mathcal{S}_i of $H_i = \{h \in H : P(h) = i\}$ has the property that $A(h) = A(h')$ if h and h' are in the same cell of the partition (for some $I_i \in \mathcal{S}_i$, one may denote by $A(I_i)$ the set $A(h)$, and by $P(I_i)$ the player $P(h)$, for any $h \in I_i$).

The *material payoffs* of players' strategies are described by functions $m_i: Z \rightarrow \mathbb{R}$ for each player $i \in N$. Further, for each player $i \in N$ let S_i denote

⁵⁸ Note that, in what follows, a node of the game tree is identified with the history leading up to it (*i.e.*: a path in the game tree), as in Osborne and Rubinstein [1994].

the *set of pure strategies* of Player i : hence, $s_i = (s_{i,h})_{h \in H \setminus Z}$ is a strategy for Player i , that is, a plan specifying the action chosen at every history after which Player i moves (with $s_{i,h}$ being the action implemented by s_i if history h occurred). A strategy profile s is a tuple of strategies, with one strategy for each player of the game: let $S = \prod_{i \in N} S_i$ be the *set of strategy profiles*; similarly define $S_{-i} = \prod_{j \neq i} S_j$ for players j other than i . Finally denote the *set of Player i 's pure strategies allowing history h* (i.e.: strategies leading to – and succeeding – h) as $S_i(h)$; strategy profiles allowing history h are defined as $S(h) = \prod_{i \in N} S_i(h)$, and $S_{-i}(h) = \prod_{j \neq i} S_j(h)$ for all players j other than i . With a slight abuse of notation, let $z(s)$ indicate a terminal history induced by some strategy profile $s \in S$.

II.3.b. Conditional systems of beliefs

Battigalli and Dufwenberg [2009] provide a framework for the analysis of dynamic psychological games, where conditional higher-order systems of beliefs influence the players' motivation. As in their model, here behavioural strategies are used to describe Player j 's beliefs about Player i 's actions at each history after which i has to play: formally, a *behavioural strategy* of Player i is a collection of independent probability measures $\sigma_i = (\sigma_i(\cdot | h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} \Delta(A_i(h))$, where $\sigma_i(a|h)$ is the probability of action a at history h and $\Delta(A_i(h))$ denotes the set of probability measures over the set of Player i 's feasible actions at history h . Then, $\Pr_{\sigma_i}(\cdot | \hat{h}) \in \Delta(S_i(\hat{h}))$ is the probability measure over Player i 's strategies conditional on \hat{h} derived from σ_i and, therefore, for some strategy $s_i \in S_i(\hat{h})$ $\Pr_{\sigma_i}(s_i | \hat{h}) := \prod_{h \in H \setminus Z: h \succcurlyeq \hat{h}} \sigma_i(s_{i,h} | h)$ indicates the conditional probability of s_i , given that \hat{h} has occurred (note that $h \succcurlyeq \hat{h}$ is a history subsequent or equal to \hat{h} , and $s_{i,h}$ is the action selected by s_i if history h took place).

Now, every player $i \in N$ holds a *system of first-order beliefs* $\alpha_i = (\alpha_i(\cdot | h))_{h \in H_i}$ about the strategies of *all* the co-players, that is, at each $h \in H_i$ Player i holds an updated (*i.e.*: revised) belief $\alpha_i(\cdot | h) \in \Delta(S_{-i}(h))$. At each $h \in H_i$ Player i further holds a *second-order belief* about the first-order belief system of *each* of the opponents: yet, for simplicity, here for some $h \in H_i$, $\beta_i(h)$ merely indicates i 's belief about an arbitrary j 's first-order belief system (*i.e.*: $\beta_i(h)$ denotes i 's belief about $\alpha_{-i} \equiv [\alpha_j(\cdot | h')]_{j \neq i, h' \in H_j}$); given that, let $\beta_i^{S_i}(h) \in \Delta(S_i(h))$ denote i 's strategy-part of $\beta_i(h)$, which represents i 's belief about what every other player *unanimously* believes about i 's strategies. Finally it is assumed that players' beliefs at different information sets must satisfy Bayes' rule and common knowledge of Bayesian updating.⁵⁹

⁵⁹ Recalling that a behavioural strategy σ_i is used to describe the other players' beliefs about Player i 's behaviour, the reader can anticipate that (as it will be imposed later on) in equilibrium $\alpha_i(s_{-i} | \hat{h}) \equiv \prod_{j \neq i} \Pr_{\sigma_j}(s_j | \hat{h})$. Also, since in equilibrium α_i will be derived from the behavioural strategy profile $\sigma = (\sigma_i)_{i \in N}$, the beliefs of every player $j \neq i$ about Player i 's strategies will be the same, which will render the above simplifying assumption about $\beta_i(h)$ innocuous and (through an additional consistency requirement) will imply that, in equilibrium, $\beta_i(h) = \alpha_{-i} \equiv [\alpha_j(\cdot | h')]_{j \neq i, h' \in H_j}$. For a discussion of the consistency requirements in equilibrium, see section II.5. below.

II.4. A model of social norms

II.4.a. Norms and perfectly norm-driven decision makers

I can now turn to shape an original theory of conformity to social norms out of a dynamic game with belief-dependent motivations. In this sub-section a “norm” is defined as a rule that dictates a set of strategy profiles at each decision node of the game tree: thus, the dictated set of strategy profiles is intended as showing the behaviour most in accordance with a certain (exogenous) principle *after a given history*; this implies that if no player ever deviates from the prescripts dictated by a norm at the initial history, then the strategy profiles dictated by that very norm at all the successor nodes will be the same as those dictated at the root of the game tree. Further, it is assumed that all norms regulating human behaviour are contained in a universal set of norms, while each player is only aware of the norms contained in her personal subset of norms (as determined by a collection of attitudes, values, goals, and practices characterizing her group, organization or institution).

Definition II.1. Given an extensive form game G , a *norm* is a set-valued function r that assigns to every non-terminal history $h \in H \setminus Z$ one or more elements from the set $S(h)$ of strategy profiles allowing history h ; that is, a norm $r: H \setminus Z \rightarrow S$ is a rule dictating the strategy profile/s most “appropriate” – according to a certain principle – for each node of the given (mixed-motive) game.

Hence, denote by R the *set of norms* and for each $i \in N$ let R_i be the norm subset of Player i , with $R_i \subseteq R$. The interpretation is as follows: given a universal set of norms (R), the culture of each player i marks out a subset R_i ,

stored in i 's memory, which contains default rules of behaviour in accordance with set usage, procedure, discipline or principle (*e.g.*: Pareto optimality, “equitable” income distribution, *etc.*).⁶⁰ It is assumed that each player's norm subset may contain all or just part of the rules of the other players' norm subsets (depending on the extent to which players share the same culture), or may even be empty.

Now, given an extensive form game G and a certain norm \dot{r} , with $\dot{r} \in R$, let $\dot{r}(h^0)$ denote the *set of strategy profiles completely consistent with \dot{r}* , where the expression “completely consistent” refers to the fact that such a set contains the very strategy profiles the norm dictates at all subsequent histories (whenever no deviation occurs along the play). Further, let $\dot{r}(H)$ denote the *set of strategy profiles (partly) consistent with \dot{r}* , which depicts the set of all the strategy profiles dictated by the norm at each history $h \in H \setminus Z$ (including histories that would be impossible to reach if $\dot{r}(h^0)$ was played); that is, $\dot{r}(H)$ contains all the strategy profiles consistent with \dot{r} only from some history h onwards (for all $h \in H \setminus Z$, with $h \succcurlyeq h^0$), *i.e.*: $\dot{r}(H) = \{s \in S(h): \exists h \in H \setminus Z \text{ s.t. } s = \dot{r}(h)\}$. Similarly, given a norm subset $R_i \subseteq R$ for each $i \in N$, denote by $R_i(H)$ the *set of strategy profiles (partly) consistent with any $r \in R_i$* , *i.e.*: $R_i(H) = \bigcup_{r \in R_i: R_i \subseteq R} r(H)$. For each $h \in H \setminus Z$ denote by $R_i(h)$ the *set of norm-complying strategy profiles allowing history h* , which is defined as $R_i(h) = \{s \in S(h): \exists r \in R_i \text{ s.t. } s = r(h)\}$; also, note that $R_i(H) \equiv \bigcup_{h \in H \setminus Z} R_i(h)$. Given that, let $A_{i,h} \left(R_i(\hat{h}) \right)$ denote the *set of Player i 's norm-complying actions at history h* (for any $h \succcurlyeq \hat{h}$), which depicts the set of actions

⁶⁰ For some examples of norms, see section II.6. below.

prescribed to Player i at history h , as dictated at \hat{h} by any norm in R_i : so, if Player i – once at history h – takes an action being part of the norm-complying strategy profiles allowing \hat{h} , then $s_{i,h} \in A_{i,h}(R_i(\hat{h}))$. Finally, denote the set of Player i 's norm-complying strategies allowing history h as $S_i(R_i(h))$, which represents i 's strategy-component of the set of norm-complying strategy profiles allowing h ;⁶¹ notice that, in the event that $R_i \neq R_j$ (for some $i, j \in N$, with $j \neq i$) the set of Player i 's norm-complying strategies allowing h – according to i 's norm subset $R_i \subseteq R$ – may not be the same as the set of Player i 's norm-complying strategies according to j 's norm subset, in other terms it might well be that $S_i(R_i(\hat{h})) \neq S_i(R_j(\hat{h}))$ for some history \hat{h} .

Given that, it is assumed that players, conditional on each history of an extensive form game, hold a conjecture about the active player's norm-complying actions at that history.

Definition II.2. Given an extensive form game G and for each $i \in N$ a norm subset $R_i \subseteq R$, a *norm-conjecture* of Player i is a collection of independent probability measures $\rho_i = (\rho_i(\cdot | h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} \Delta(A_{P(h)}(h))$, with $\rho_i(a|h)$ being the probability of action a at history h , such that:

⁶¹ Obviously the set of norm-complying strategy profiles allowing history h^0 can be defined from actions as $R_i(h^0) \equiv \prod_{y \in N} (\prod_{h \in H \setminus Z} A_{y,h}(R_i(h)))$, that is, the Cartesian product of all players' (sequences of) norm-complying actions at all histories or, in other words, the Cartesian product of all players' norm-complying strategies allowing h^0 .

$$\text{supp } \rho_i = \text{supp } (\rho_i(\cdot | h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} A_{P(h),h}(R_i(h)),$$

where $\text{supp } \rho_i$ denotes the *support* of ρ_i , and $A_{P(h),h}(R_i(h))$ is the set of norm-complying actions of the active player (*e.g.*: $P(h) = y$) at history h .⁶²

Notice that the support of ρ_i is the set of the active player's (norm-complying) actions which are assigned positive probability by i 's norm-conjecture ρ_i at each history. In plain words – in order for (the potentially conformist) Player i to identify the actions being part of a strategy profile appearing to best describe some “normally-expected behaviour” (according to the standards set in i 's social group and reflected in R_i) – conditional on each $h \in H \setminus Z$, she holds a conjecture $\rho_i(\cdot | h)$ about the active player's norm-complying actions at history h .⁶³ It should be stressed that (possibly depending on the size of the social group) it may not be obvious to the members of a group that one norm is more adequate than another, so they have to form beliefs about what is expected (from the group as a whole) to do. It follows that *Player i 's belief about the strategy profiles consistent with some norm in R_i* is a probability measure over $R_i(H)$; therefore, $\text{Pr}_{\rho_i}(\cdot | \hat{h}) \in \Delta(R_i(\hat{h}))$ is the probability measure over norm-complying strategy profiles conditional on \hat{h}

⁶² Recall that, for each player $y \in N$ that takes an action after some history h , the value of the player function at h is y , *i.e.*: $P(h) = y$.

⁶³ Notice that, while the above system of first-order beliefs α_i is a probability measure over the strategies of all the co-players (*i.e.*: all players other than i), here $\rho_i = (\rho_i(\cdot | h))_{h \in H \setminus Z}$ is a collection of independent probability measures over the actions the active player ought to take at each history $h \in H \setminus Z$. Also, note that it is assumed that ρ_i is undefined if $R_i = \emptyset$.

and so, for some $s \in R_i(\hat{h})$, the *conditional probability of norm-complying strategy profile* s (given that \hat{h} has occurred) is computed by $\Pr_{\rho_i}(s|\hat{h}) = \prod_{h \in H \setminus Z: h \succ \hat{h}} \rho_i(s_{P(h),h}|h)$.

Before proceeding, notice that the above definition of norm reminds us of the one suggested by López-Pérez [2008], where a norm is defined as a correspondence mapping h into $A(h)$.⁶⁴ However, on a mere conceptual level, here it is assumed that defining a norm as a correspondence mapping non-terminal histories into *strategy profiles* allows to better capture the strategic complexity of such rules, in that a certain action of Player $P(h)$ might be considered most appropriate for some decision node (*i.e.*: for some history h) only in light of what is expected from all players to do at the successor nodes. In substance, on the one hand, assuming a unique possible norm in the society, like the E-norm of López-Pérez, the value of that norm at the initial history (*i.e.*: $s = \dot{r}(h^0)$ for some $s \in S$) would induce the same leaves as López-Pérez's fairmax paths; on the other hand, in the event of a player's deviation from the initially-recommended path, López-Pérez's norm selects the whole set $A(h)$ whereas the present theory allows the norm \dot{r} to determine the actions in accordance with a given principle also at histories that would be impossible to reach if $\dot{r}(h^0)$ was played. Also, disregarding the role of expectations in sustaining a social norm seems to be a conceptual drawback of López-Pérez's model, although that makes his framework a parsimonious one.

⁶⁴ For a detailed review of López-Pérez's [2008] and Li's [2008] models of norm compliance, refer to the appendix (section II.8.b. below).

The present definition of norm appears more similar to Li's [2008] convention, where "the right thing to do" depends also on what the co-player is (believed to be) doing. Indeed, Li models a norm as belief-dependent rankings over one's strategy space and normalizes them to the unit interval; more precisely, given a strategic form game and denoting Player i 's set of mixed strategies as Σ_i and i 's belief about j 's strategy as $b_j \in \Sigma_j$, Li defines a convention as a mapping $\omega: \Sigma_i \times \Sigma_j \rightarrow [0,1]^2$. Notice that, by applying her definition (which is given in the context of 2-player normal form games) to a 2-player extensive form game with no proper subgames, in some cases one would get a collection of values corresponding to what is here referred to as "Player i 's belief about the strategy profiles consistent with some norm in R_i " (i.e.: $\Pr_{\rho_i}(\cdot | h^0) \in \Delta(R_i(H))$); however, the two notions do not generally coincide since Li's convention is not a probability measure and so in her model it may well be that $\sum_{(\sigma_i, b_j) \in \Sigma_i \times \Sigma_j} \omega_i(\sigma_i, b_j) \neq 1$.

Now, I can move on to introduce a notion of "pure social response to a norm-conjecture" (which reminds of a *moral choice*, as intended in some philosophical literature), as follows.

Definition II.3. A strategy $s_i^{**} = (s_{i,h})_{h \in H \setminus Z}$ is a *pure social response to norm-conjecture* $\rho_i = (\rho_i(\cdot | h))_{h \in H \setminus Z}$ at history \hat{h} if the following condition holds for all $s_i \in S_i(\hat{h})$:

$$\Pr_{\rho_i}(s_i^{**} | \hat{h}) \geq \Pr_{\rho_i}(s_i | \hat{h}),$$

where $\Pr_{\rho_i}(s_i | \hat{h})$ is the conditional probability of a pure strategy of Player i at history \hat{h} – according to i 's own norm-conjecture ρ_i – and is computed by $\Pr_{\rho_i}(s_i | \hat{h}) = \prod_{h \in H \setminus Z: h \succ \hat{h}} \rho_i(s_{i,h} | h)$.

Notice that, by definition II.2, $\rho_i = (\rho_i(\cdot | h))_{h \in H \setminus Z}$ assigns positive probability only to (some) norm-complying actions at history h , which trivially implies that a pure social response s_i^{**} to norm-conjecture ρ_i at history \hat{h} must be a norm-complying strategy, *i.e.*: $s_i^{**} \in S_i(R_i(\hat{h}))$. Given that, one can define the set of pure social responses to ρ_i at \hat{h} as $C_i(\rho_i) := \arg \max_{s_i \in S_i(\hat{h})} \Pr_{\rho_i}(s_i | \hat{h})$.

Definition II.4. A *perfectly norm-driven decision maker* is an agent i such that – facing a decision problem defined by an extensive form game G and given a non-empty norm subset $R_i \subseteq R$ – i plays a pure social response $s_i^* \in S_i(R_i(h^0))$ to norm-conjecture ρ_i .

To sum up, definition II.3 determines the set (C_i) of Player i 's social responses to ρ_i at \hat{h} as a collection of pure strategies with the highest probability of adhering to some principle, as prescribed by some $r \in R_i$, from some history \hat{h} onwards. Definition II.4 suggests a notion of extremely socially-conscientious agent i by which i takes the actions selected by strategy s_i^* that, according to ρ_i , are most appropriate for all histories after which she moves (regardless of the material payoff to i). In the next subsection the utility function of a perfectly norm-driven decision maker will be devised as a special case of that of another category of socially-conscientious agents, namely the “*fairly* norm-driven decision makers”.

II.4.b. Belief-dependent conformist preferences

I model a (fairly) norm-driven decision maker i as a player with conformist preferences, whose utility function is a linear combination of her material payoff and a component representing the social cost of deviating, in the form

of the sum of losses that other conformist players j would suffer because of a norm violation. For that, one needs to define some player j 's expectation about her material payoff, given her strategy s_j and her initial belief $\alpha_j = (\cdot | h^0)$ about the strategies of the co-players (as it will be soon clear, j 's expectation affects i 's utility function): so, drawing on Battigalli and Dufwenberg's [2007] concept of *simple guilt*, such an expectation is given by $E_{s_j, \alpha_j} [m_j | h^0] = \sum_{s_{-j}} \alpha_j (s_{-j} | h^0) m_j (z(s_j, s_{-j}))$. Here, if Player j is a norm-driven decision maker – and presumes that her co-players are norm-driven too – she can form her belief α_j by assuming her co-players' behaviour to be consistent with some r , with $r \in R_{-j}$.

Now, the present theory assumes that players are naïve in the following way: if Player j presumes that her co-players are norm-driven, then she believes that they hold the same norm-conjecture as hers (*i.e.*: $\rho_j = (\rho_j(\cdot | h))_{h \in H \setminus Z}$); hence, Player j forms her first-order belief α_j by assuming her co-players' behaviour (at each history where they are active) to be consistent with her own norm-conjecture ρ_j . Notice that, here, her initial belief $\alpha_j = (\cdot | h^0)$ would still correspond to a probability measure over the strategies of all the opponents, but with the support of α_j containing only opponents' norm-complying strategies – according to j 's norm subset R_j – therefore the probability of a certain strategy profile of all players other than j is now given by:

$$\alpha_j (s_{-j} | h^0) \equiv \Pr_{\rho_j} (s_{-j} | h^0) = \prod_{h \in H \setminus Z: h \succcurlyeq h^0} \rho_j (s_{-j, h} | h). \quad (2.4.1)$$

Note that, for the sake of simplicity, the present theory assumes that *players cannot randomize*, yet randomized choices enter the analysis as an expression of the players' beliefs about the opponents' (norm-complying) strategies. Given that, a "fairly norm-driven decision maker" is defined as follows.

Definition II.5. A *fairly norm-driven decision maker* has conditionally-conformist preferences characterized by the following utility function

$$u_i^C(z, s_{-i}, \alpha_j) = m_i(z) - k_i d_i^C d_i^E \sum_{j \neq i} \max \{0, E_{\rho_i, s_j, \alpha_j} [m_j | h^0] - m_j(z)\},$$

with $s_{-i} \in S_{-i}(z)$, $k_i \in [0, \infty)$, and where:

- k_i is Player i 's *sensitivity to the norm*, which measures the agent's degree of conformity, thereby the "marginal cost" of a norm violation;
- d_i^C is a dummy variable equal to one if agent i is aware of one or more norms applicable to the given decision problem (*i.e.*: whenever $R_i \neq \emptyset$), equal to zero otherwise;
- d_i^E is a dummy variable equal to one if agent i believes that at least one $j \neq i$ will adhere to some r , with $r \in R_i$, depending on the leaf z^{t-1} reached in a previous instance of the same multi-person decision problem (whenever the game is repeated) as follows

$$d_i^E = \begin{cases} 1 : & \text{in period 1,} \\ 1 : & \text{in period } t \in T \text{ if } \exists s \in R_i(H) \text{ s.t. } z^{t-1} = z(s),^{65} \\ 0 : & \text{in period } t \in T \text{ if } \nexists s \in R_i(H) \text{ s.t. } z^{t-1} = z(s). \end{cases}$$

It is now clear that the psychological loss comes from any positive difference between the presumed expected payoff to j and the payoff j would get in the event of a norm violation; notice that i does not know what α_j is, yet she can estimate it in the same fashion as in formula (2.4.1) by presuming that every $j \neq i$ holds the same norm-conjecture as hers. Also notice that the dummy variable d_i^E contributes to conceiving of a motive of *conditionally* conformist behaviour in that – whenever the game is repeated – d_i^E takes on the value 1 if there exists at least one strategy profile consistent with some $r \in R_i$ (*i.e.*: $s \in R_i(H)$) such that it induces the terminal history realized in the previous period. To sum up, if $d_i^C = 1$, $d_i^E = 1$, and $k_i > 0$ Player i will exhibit conformist preferences;⁶⁶ besides, the larger her sensitivity k_i , the more will she experience some disutility from someone's not conforming to her (presumed) norm. In this respect, it should be stressed that the sensitivity parameter k_i sets the size of a hypothetical feeling of uneasiness of member i of a social group in which (because of someone's norm violation) some other member's welfare gets unjustly, or unexpectedly, reduced.

⁶⁵ Note that z^{t-1} denotes the terminal history realized in the previous period (*i.e.*: period $t - 1$), with the *set of periods* being $T = \{1, \dots, q\}$. Recall that $R_i(H)$ is the set of strategy profiles consistent with any $r \in R_i$ (as defined in section II.4.a. above).

⁶⁶ Obviously if $k_i = 0$ or $d_i^C = 0$ or $d_i^E = 0$ the utility function reduces to one of classical, non-conformist motivation.

Further, it should be stressed that such a utility function differs from the one Bicchieri presents in the appendix to Chapter 1 (Bicchieri [2006], Ch. 1) since, according to Bicchieri's motivation function, Player i would suffer a psychological loss even in the case in which she conforms to the norm but Player j does not and, by doing so, j gets a material payoff larger than the one implied by the norm (see the Prisoner's Dilemma example in section II.2. above). However, even though Bicchieri's formulation seems plausible, here it is assumed that the most prominent cause of conformity in social dilemmas is some painful emotion which an individual i may feel in the event that any *other* member of the social group gets an *outcome inferior to the "normally-expected" one* (according to ρ_i), regardless of i 's direct responsibility: in this respect, notice that here an agent with preferences represented by a utility function u_i^C , on the one hand, would not suffer a psychological loss in the case in which she conforms to her (presumed) norm but her *only* co-player does not and, as a consequence, i gets a material payoff *lower* than the one implied by her norm; on the other hand, i would suffer a psychological loss in the case in which she conforms to her presumed norm but Player g does not and, because of g 's strategy, a third player j gets a material payoff *lower* than the one implied by i 's norm.⁶⁷

⁶⁷ A justification of the present modelling of the psychological disutility (from someone's not conforming to a norm) comes from the following argument: if, on the one hand, it might seem more suitable to let Player i 's psychological loss equal the opponents' "disappointment" due to i 's behaviour only (and not due to someone else's strategy), on the other hand, it should be further stressed that such psychological loss has to be intended as the malaise of a member of a social group in which some peer's welfare gets unjustly reduced. Indeed, the current modelling accounts for Player i 's suffering a psychological loss in the case in which

Before proceeding, in light of definition II.5, one further observation is necessary.

Remark II.1. A *perfectly norm-driven decision maker* (see also definition II.4) has unconditionally-conformist preferences characterized by a utility function u_i^C , with:

- (i) $d_i^C = 1 [R_i \neq \emptyset]$;
- (ii) $d_i^E = 1$ [even in period $t \in T$ if $\exists s \in R_i(H)$ s.t. $z^{t-1} = z(s)$];
- (iii) $k_i \rightarrow \infty$.

A perfectly norm-driven decision maker represents an agent i who cares infinitely much about the social implications of her actions, even though – in previous occurrences (if any) of the same social dilemma – no other member of the population has complied with any of the norms contained in R_i . In other words, a perfectly norm-driven decision maker has unconditional preferences for conformity to some norm $r \in R_i$; that is, norm r constitutes a *moral* norm for i (in this respect, empirical expectations set the boundaries between *moral*, often referred to as “Kantian”, norms and *social* norms).

she conforms to her presumed norm but Player g does not and, because of g 's behaviour only, a third player j gets a material payoff *lower* than expected (by i) whereas Player i , along with g , gets a material payoff *higher* than the one implied by i 's norm. As a matter of fact, norm-enforcing instincts have been probed using neuroimaging: results show that humans have an automatic drive to react to social wrongs perpetrated on *themselves* as well as *others* (Montague and Lohrenz [2007]).

II.4.c. *Social norms*

Given the above apparatus, I shall introduce a set of conditions for a social norm to exist or, more explicitly, conditions for a norm r to constitute a “social norm for i ”. Before stating such conditions it should be highlighted that, as mentioned above, the present theory of conformity draws on guilt aversion in that guilt is a driver for conformity in mixed-motive games; however, this is not a social preference model proper because, as it will be soon clear, social norms rather serve as an equilibrium selection device, which here means that – when beliefs are correct – the guilt component of the utility function is always null in equilibrium.⁶⁸ In fact, the form of the current utility function incorporates a taste for *conditional* social preferences which characterize, for example, a scenario where people dislike vandalism or littering, although they are more likely to indulge in misbehaviour whenever evidence of vandalism or littering are present in the environment; more generally, the present theory well depicts the case of an agent having a *preference for some principle* – whatever the norm specifically prescribes to her – only on condition that others do not deviate from the precepts of that norm (which is in sharp contrast with the social preference models mentioned in the introduction).

Definition II.6. Let $r \in R$ be a norm applicable to a certain class \mathcal{C} of mixed-motive games, where each game is a structure $G = \langle N, H, P, (u_i^{\mathcal{C}})_{i \in N} \rangle$. r is a *social norm* for Player i of game G , if the following conditions hold for i .

⁶⁸ For a discussion of possible equilibrium scenarios, see section II.5. below.

1. (contingency) $r \in R_i \Rightarrow d_i^C = 1$.
2. (conditional preference) $\exists s^* \in r(h^0)$ s.t. $u_i^C(z(s_i^*, s_{-i}^*)) \geq u_i^C(z(s_i, s_{-i}^*))$ for $\forall s_i \in S_i$, where $r(h^0)$ is the set of strategy profiles completely consistent with r . That is, i prefers to adhere to r in a play of G if

2.1. (empirical expectations)

$$\left(\begin{array}{l} \alpha_i(s_{-i}|h^0) = \prod_{j \neq i} \Pr_{\rho_i}(s_j|h^0) \text{ for } \forall s_{-i} \in S_{-i}, \\ \text{with } \text{supp}(\rho_i(\cdot|h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} A_{P(h),h}(r(h^0)) \end{array} \right) \Rightarrow d_i^E = 1;$$

and either 2.2. a or 2.2. b

2.2. a. (normative expectations) $\beta_i^{s_i}(h^0) = \left(\Pr_{\rho_i}(s_i|h^0) \right)_{s_i \in S_i}$,

2.2. b. (normative expectations with “psychological sanctions”)

- i. $\beta_i^{s_i}(h^0) = \left(\Pr_{\rho_i}(s_i|h^0) \right)_{s_i \in S_i}$, and
- ii. $D_i^j = E_{\rho_i, \beta_i, \alpha_i}[m_j|h^0] - m_j(z(s_i, s_{-i}^*)) \geq 0$ for $\forall s_i \in S_i$,
 $\forall j \in N$ ($j \neq i$, with $D_i^j > 0$ for at least one j), and
- iii. $k_i > 0$ and sufficiently large.

The above set of conditions for a social norm to exist introduces a mathematically-precise definition of social norm, which in part formulizes Bicchieri’s [2006] verbalization.⁶⁹ Besides, the overall interpretation is not

⁶⁹ See section II.2. above, for Bicchieri’s own conditions: in this regard note that Bicchieri’s conditions for a social norm to exist differ from the conditions of the present theory (as stated in definition II.6), among the other issues, also in that such conditions are here defined *from*

dissimilar: indeed, for both theories, a social norm has to be intended as informal (it is not a legal rule) and not necessarily enforced (it is non-binding); yet, and most importantly, it is necessarily sustained by expectations in that it is not unconditional (it is not a moral norm). In this connection, a few comments are in order. Condition 1 (i.e.: $r \in R_i$) states that i is aware of norm r applicable to game G . Condition 2.1 (i.e.: $(\alpha_i(s_{-i}|h^0) = \prod_{j \neq i} \Pr_{\rho_i}(s_j|h^0)$ for $\forall s_{-i} \in S_{-i}$, with $\text{supp}(\rho_i(\cdot|h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} A_{P(h),h}(r(h^0))$) states that i believes that every $j \neq i$ adheres to $r \in R_i \cap R_j$; that is, i 's first-order belief is derived from i 's norm-conjecture ρ_i , with the support of ρ_i containing only norm-complying actions dictated by r .⁷⁰ Conditions 2.2.a and 2.2.b apply to alternative situations, that is: the former refers to the case where i believes that every $j \neq i$ expects her to behave according to i 's norm-conjecture ρ_i (which means that i 's second-order belief is derived from ρ_i), simply because j acknowledges the legitimacy of i 's norm-conjecture; the latter refers to the case where i believes that every $j \neq i$ expects her to behave according to ρ_i , and j also prefers i to conform.⁷¹ More precisely, condition 2.2.b holds

the viewpoint of Player i , irrespective of the correctness of her beliefs; a discussion of additional points of difference is provided below.

⁷⁰ Notice that $A_{P(h),h}(r(h^0))$ is the set of norm-complying actions of the active player at history h , as dictated at h^0 by r .

⁷¹ Condition 2.2.a depicts a situation in which i believes that every $j \neq i$ expects i to conform to the norm, yet i does not necessarily believe that j prefers i to conform: in other words, condition 2.2.a accounts for a situation where j conforms because j 's cost of a norm violation (i.e.: k_j) is high enough to make j 's deviation from s_j^* unprofitable, although there could be a terminal history induced by some strategy profile (s_{-j}, s_j^*) where j would be better off. An example of a case where condition 2.2.a is fulfilled is discussed in section II.6.a. below (see

whenever its three components hold at once: (i) the first expression (i.e.: $\beta_i^{s_i}(h^0) = \left(\Pr_{\rho_i}(s_i|h^0) \right)_{s_i \in S_i}$) states that i believes that every $j \neq i$ expects her to behave according to i 's norm-conjecture ρ_i (that is, i 's second-order belief is derived from ρ_i); (ii) the second expression (i.e.: $D_i^j = E_{\rho_i, \beta_i, \alpha_i}[m_j|h^0] - m_j(z(s_i, s_{-i}^*)) \geq 0$ for $\forall s_i \in S_i, \forall j \in N$) states that i believes that every $j \neq i$ prefers her to behave according to i 's norm-conjecture ρ_i (that is, i 's expectation of j 's disappointment D_i^j in the event of a norm violation is non-negative for each $z(s_i, s_{-i}^*) \neq z(s_i^*, s_{-i}^*)$); (iii) the final expression (i.e.: $k_i > 0$ and sufficiently large) states that i 's cost of a norm violation is psychologically hurting (whenever $k_i > 0$) and is high enough to make i 's deviation from s_i^* unprofitable.

Before introducing the conditions for a social norm to (exist and) be followed by every $i \in N$, I shall briefly illustrate how the above sufficiently-large- k_i requirement operates in a simple decision problem. Hence, it is required that $k_i \geq \max \{ \hat{k}_i^{s_i}, \dots, \hat{k}_i^{s_i^*} \}$, where each $\hat{k}_i^{s_i}$ is a sensitivity parameter such that $u_i^C(z(s_i, s_{-i}^*)) = u_i^C(z(s_i^*, s_{-i}^*))$ for some $s_i \in S_i$, with $s_i^* \in r(h^0)$. Consider the following Discrete Dictator Game.⁷²

second scenario in proposition II.2); note that Bicchieri's interpretation of condition 2.2.a is slightly different.

⁷² Notice that *DDG*, as represented in Figure II.1, is a peculiar case of Dictator Game in that it presents an inefficient option (i.e.: for $s_1 = c$).

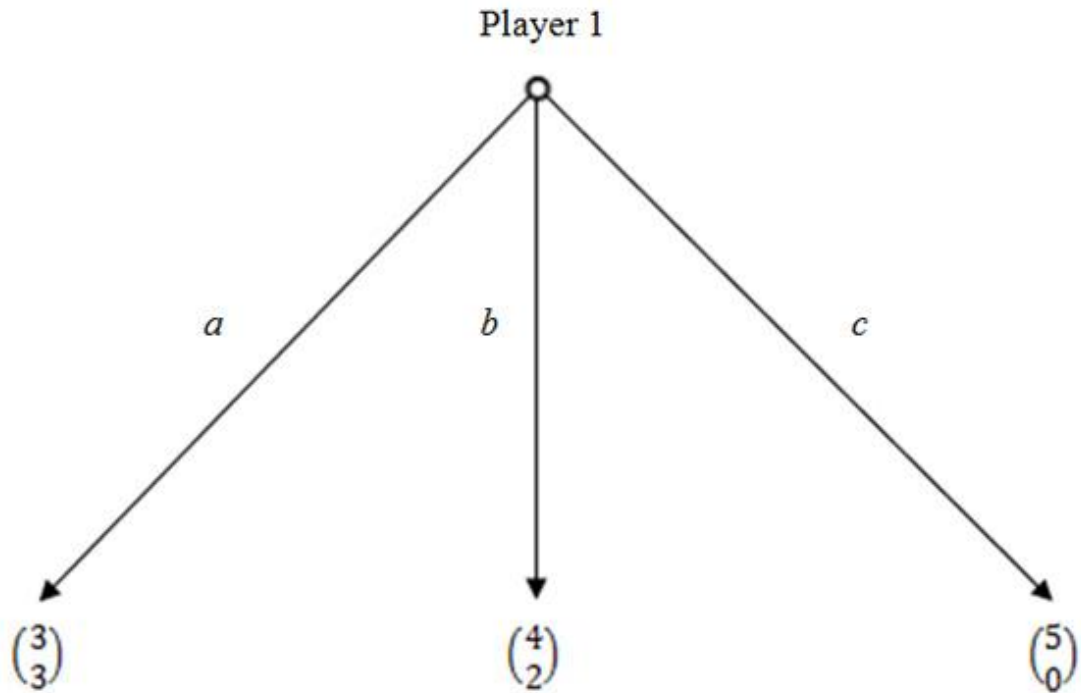


Figure II.1 - Discrete Dictator Game “*DDG*”

Now, suppose that a social norm based on some equitable principle has been established (*i.e.*: an “equitable” social norm r^* exists for Player 1),⁷³ then the set of strategy profiles completely consistent with r^* is singleton, that is, $r^*(h^0) = \{a\}$. Therefore – given that in this case Player 1 believes that $\rho_i(a|h^0) = 1$ for $i = 1, 2$, and that $\beta_1(s_1|h^0) = \Pr_{\rho_i}(s_1|h^0) = 1$ for $s_1 = a$ – it follows that Player 1’s expectation of Player 2’s (expected) material payoff at h^0 equals $E_{\rho_i, \beta_1}[m_2|h^0] = 1 \cdot 3 = 3$. Further, Player 1’s expectation of Player

⁷³ See section II.6. below for some precise definitions of norms based on equitable principles.

2's "disappointment" at $(b|h^0)$ is $E_{\rho_i, \beta_1}[m_2|h^0] - m_2(z(b)) = 3 - 2 = 1$; this implies that Player 1's utility at $(b|h^0)$ equals $u_1^c(z, \rho_i, \beta_1) = m_1(z(b)) - \hat{k}_1^b [E_{\rho_i, \beta_1}[m_2|h^0] - m_2(z(b))] = 4 - \hat{k}_1^b \cdot 1$. Similarly, Player 1's expectation of Player 2's "disappointment" at $(c|h^0)$ is $E_{\rho_i, \beta_1}[m_2|h^0] - m_2(z(c)) = 3 - 0 = 3$; this implies that Player 1's utility at $(c|h^0)$ equals $u_1^c(z, \rho_i, \beta_1) = m_1(z(c)) - \hat{k}_1^c [E_{\rho_i, \beta_1}[m_2|h^0] - m_2(z(c))] = 5 - \hat{k}_1^c \cdot 3$. On the other hand, Player 1's expected utility (=payoff) at $(a|h^0)$ is given by $u_1^c(z(a), \rho_i, \beta_1) = m_1(z(a)) = 1 \cdot 3 = 3$. Finally, Player 1's conformist preferences against b and c can be expressed, respectively, as: $u_1^c(z(b), \rho_i, \beta_1) = m_1(z(a)) \Rightarrow 4 - \hat{k}_1^b = 3 \Rightarrow \hat{k}_1^b = 1$; $u_1^c(z(c), \rho_i, \beta_1) = m_1(z(a)) \Rightarrow 5 - 3\hat{k}_1^c = 3 \Rightarrow \hat{k}_1^c = 2/3$. Consequently, the sufficiently-large- k_i requirement (for social norm r^* to exist for Player 1 of *DDG*) imposes that $k_1 \geq \max \{\hat{k}_1^b, \hat{k}_1^c\} = \max \{1, \frac{2}{3}\}$, that is, $k_1 \geq 1$.

Now, the above conditions for a social norm to exist are to be intended as those necessary for a norm r to be held in place: if fulfilled for every $i \in N$, at least one of the strategy profiles dictated by that norm r is an equilibrium – provided that all beliefs are correct and that players maximize expected utilities – as implied by the remarks II.2-3 below. Hence, definition II.6 results in a social norm (existing and) being "followed by population N " if the conditions in remark II.2 simultaneously hold.

Remark II.2. A social norm r^* (exists and) is followed by population N , if: every player $i \in N$ has conformist preferences represented by a utility function u_i^c , with $d_i^c = 1$, $d_i^E = 1$, and $k_i > 0$; every i maximizes her expectation of u_i^c ; every i holds correct beliefs about every j 's ($j \in N$, with $j \neq i$) first-order belief and behaviour; every player i 's behaviour is consistent

with at least one of the end-nodes yielded by $r^* \in R_i \cap R_j$ (according to norm-conjectures $\rho_j = \rho_i$, for $\forall j \in N$); k_i is sufficiently large for every $i \in N$.

Note that the expression “a social norm r^* is followed by population N ” (or “every player $i \in N$ conforms to r^* ”) implies that every player in the population plays her part of one of the strategy profiles contained in $r^*(H)$, which in turn implies that every player plays her part of one of the strategy profiles contained in $r^*(h^0)$.

II.5. Equilibrium concept

In this section an equilibrium concept for mixed-motive games with belief-dependent conformist preferences is discussed: by imposing the requirement that all beliefs (and norm-conjectures) are correct in equilibrium, I derive a “Social Sequential Equilibrium” as a special case of the sequential equilibrium notion of Kreps and Wilson [1982]. Kreps and Wilson’s definition of equilibrium for generic extensive form games consists of sequentially rational, consistent assessments, where: (i) an *assessment* is a profile of behavioural strategies and conditional first-order beliefs (along with higher-order beliefs in Battigalli and Dufwenberg’s [2009] specification); (ii) an assessment is *consistent* if the profile of first-order beliefs $\alpha = (\alpha_i)_{i \in N}$ is derived from the behavioural strategy profile $\sigma = (\sigma_i)_{i \in N}$, that is, for $\forall i \in N$, $\forall s_{-i} \in S_{-i}$, $\forall \hat{h} \in H_i$, it must be that $\alpha_i(s_{-i}|\hat{h}) = \prod_{j \neq i} \Pr_{\sigma_j}(s_j|\hat{h})$; notice that, since α_i is derived from $\sigma = (\sigma_i)_{i \in N}$, the beliefs of every player $j \neq i$ about Player i ’s strategies must be the same; given that, Battigalli and Dufwenberg’s [2009] specification of sequential equilibria for psychological games extends the consistency requirement by demanding that higher-order beliefs at each information set are correct for $\forall i \in N$, $\forall h \in H_i$, that is,

$\beta_i(h) = \alpha_{-i}$; (iii) finally, an assessment is *sequentially rational* if, for every player i and every information set $h \in H_i$, the strategy of i is a best response to the other players' strategies given i 's beliefs at h .

In the present framework I further extend the consistency requirement by imposing that every player i holds correct beliefs about every j 's first-order belief, which is derived from norm-conjectures $\rho_j = \rho_i$ (for $\forall j \in N$, with $j \neq i$). It follows the definition of a “socially consistent assessment”.

Definition II.7. A *socially consistent* assessment is a profile $(\sigma, \rho, \alpha, \beta) = (\sigma_i, \rho_i, \alpha_i, \beta_i)_{i \in N}$ that specifies behavioural strategies, norm-conjectures, first- and second-order beliefs, such that for $\forall i \in N, \forall s_{-i} \in S_{-i}, \forall \hat{h} \in H_i$:

$$(i) \quad \alpha_i(s_{-i}|\hat{h}) = \prod_{j \neq i} \Pr_{\sigma_j}(s_j|\hat{h});$$

$$(ii) \quad \beta_i(\hat{h}) = \alpha_{-i};$$

$$(iii) \quad \beta_i^{s_i}(\hat{h}) = \left(\Pr_{\rho_i}(s_i|\hat{h}) \right)_{s_i \in S_i} \text{ and } \left(\rho_j(\cdot|h) \right)_{j \neq i, h \in H \setminus Z} = \left(\rho_i(\cdot|h) \right)_{h \in H \setminus Z}.$$

Notice that condition (iii) in definition II.7 is the distinguishing feature of a socially consistent assessment in that it implies that (not only are beliefs derived from a behavioural strategy profile but also) a behavioural strategy profile $\sigma = (\sigma_i)_{i \in N}$ contains probability measures which equal those contained in norm-conjecture ρ_i , with $\rho_j = \rho_i$ for every $j \neq i$.

The core equilibrium concept for mixed-motive games with belief-dependent conformist preferences can now be presented.

Definition II.8. Given an extensive form game G and a norm subset $R_i \subseteq R$, for each $i \in N$, where $G = \langle N, H, P, (u_i^c)_{i \in N} \rangle$, a *Social Sequential Equilibrium*

(“SSE”) of G is a socially consistent assessment, such that for $\forall i \in N$, $\forall h \in H_i$, $\forall s_i^* \in S_i(h)$:

$$\Pr_{\sigma_i}(s_i^*|h) > 0 \Rightarrow s_i^* \in S_i(R_i(h)) \Rightarrow s_i^* \in \arg \max_{s_i \in S_i(h)} E_{s_i, \alpha_i, \beta_i, \rho_i}[u_i^C|h].$$

In plain words, a socially consistent assessment is a social sequential equilibrium if each probability measure $\Pr_{\sigma_i}(\cdot|h)$ assigns positive conditional probability only to conditional expected-payoff maximizing *norm-complying* strategies; that is, all players hold the same belief about the strategy profiles consistent with some norm in R_i and maximize the expectation of the utility function (given their correct beliefs). Note that, here, it is assumed common knowledge of the utility functions u_i^C , implying that the sensitivity parameters k_i are commonly known (and are, in effect, *sufficiently large*) as well as the fact that each player i knows that every $j \neq i$ adheres to some $r \in R_i \cap R_j$ (*i.e.*: resulting in $d_i^E = 1$), given that each player’s norm-subset is non-empty (*i.e.*: resulting in $d_i^C = 1$).

Further, note that (if $d_i^C = 1$ and $d_i^E = 1$) one could define a sequential equilibrium *à la* Battigalli and Dufwenberg [2007] by specifying a consistent assessment (σ, α, β) dropping condition (iii) of definition II.7 above: given that, their equilibrium notion can be obtained by dropping the requirement that each probability measure $\Pr_{\sigma_i}(\cdot|h)$ assigns positive conditional probability only to norm-complying strategies in definition II.8 above; this implies that every game with simple guilt has a *psychological sequential equilibrium à la* Battigalli and Dufwenberg, irrespective of the magnitude of the sensitivity parameter $(k_i)_{i \in N}$. Conversely, here, a fairly norm-driven decision maker with utility function u_i^C (as in definition II.5) has conditionally-conformist preferences such that if – for some player $g \neq i$ – g ’s cost of a norm violation is *not* high enough to make g ’s deviation from s_g^* unprofitable

(i.e.: k_g is not sufficiently large), then condition 2.1 of definition II.6 will not hold for i (i.e.: resulting in $d_i^E = 0$) and so the utility function u_i^C will reduce to one of classical (“non-psychological”) motivation, thereby implying a standard notion of equilibrium.⁷⁴

On the other hand, it should be stressed that the present theory considers the possibility that players may deviate from the precepts of a norm in the case where norm r constitutes a social norm only for some player j of game G , as due to the fact that j may hold incorrect beliefs about others’ belief and behaviour and, as a consequence, r might not be followed by all members of N : this is due to the definition of social norm being based on expectations and conditional preferences. Instead, if condition 1 of definition II.6 holds for every player $i, j \in N$ and every $i, j \in N$ holds correct norm-conjectures $\rho_j = \rho_i$ (as well as first- and second-order beliefs), then the following *equilibrium scenarios* are possible:

- (i) conditions 2.1-2.2 of definition II.6 hold for every player $i, j \in N \Rightarrow$ social norm r^* exists (for $\forall i, j \in N$) and is followed by population $N \Rightarrow$ a *social sequential equilibrium* of G occurs; *or else*
- (ii) conditions 2.1-2.2 of definition II.6 do not hold \Rightarrow social norm r does not exist for any player $i, j \in N$ (and it is not followed by population N) \Rightarrow a social sequential equilibrium of G does not occur (yet a *subgame perfect equilibrium* occurs if G is a game with observable actions and

⁷⁴ That is justified by the classical interpretation of equilibrium beliefs as the result of a transparent reasoning process.

no chance moves; a *standard sequential equilibrium à la* Kreps and Wilson occurs otherwise).

Note that in scenario (ii) the utility function reduces to one of classical, non-conformist motivation, which justifies the standard notions of equilibrium adopted. It should be stressed that – for a given extensive form game G , and a norm subset $R_i \subseteq R$ for each $i \in N$ – a social sequential equilibrium does not always exist: this perfectly captures the fragility of social norms in actual society.

Finally, the following result is a direct consequence of definition II.8.

Remark II.3. Given an extensive form game G , and a norm subset $R_i \subseteq R$ for each $i \in N$, if a social norm $r^* \in R_i$ (exists and) is followed by population N , then some social sequential equilibrium of G occurs.

Notice that the *converse is not necessarily true*, as a certain socially consistent, sequentially rational assessment (*i.e.*: a social sequential equilibrium) might be induced by multiple norms in R , some of which may not even belong to R_i for some $i \in N$. For example, consider a 2-player game and let each player's norm subset be defined as $R_i = \{r^E, r^M\}$ for $\forall i \in N$. Then, assume that at the initial node of the game, the norm r^E dictates the strategy profiles $r^E(h^0) = \{(b, c), (a, c)\}$ whereas the norm r^M dictates the strategy profiles $r^M(h^0) = \{(a, c), (a, d)\}$, with each pair of lower-case letters denoting a strategy profile. Further, assume that both players are fairly norm-driven decision makers with utility functions $(u_i^C)_{i \in N}$ and that, while holding correct beliefs, they play the strategy profile (a, c) . Now, while (a, c) is a social sequential equilibrium of the game, this does not necessarily imply that, say, r^E (rather than r^M) constitutes a social norm and is being followed

by the players of the game.⁷⁵ Interestingly, this well captures the case of a traveller who, once in a foreign country, observes some locals interacting (without taking part in the actual game herself): while the outcome of the interaction may turn out to be compatible with some of the norms stored in the observer's mind, she may not be able to tell which one has been held in place, especially if the foreign country is particularly culturally-different from hers; on a smaller case, a similar problem occurs the first time we happen to interact with members of a group, organization or institution whose social norms we do not yet know.

II.6. Illustrations

In this section I turn to analyse some specific dynamic interactions accounting for belief-dependent conformist preferences. Recalling that a norm (definition II.1 above) is a set-valued function $r \in R$ that assigns to every non-terminal history $h \in H \setminus Z$ one or more elements from the set $S(h)$ of strategy profiles allowing history h – and following my discussion about a pattern of belief formation relative to a *presumed* social norm – here I move on to define an illustrative set of norms, comprising the following principles, which are assumed to regulate behaviour in social dilemmas:

$R = \{r^E, r^F, r^M, r^W\}$, where each norm is defined as below.

⁷⁵ Obviously the players of the game know which social norm they are following.

- *Equity principle:*

$$r^E(H) = \{s \in S(h): h \in H \setminus Z; z(s) \text{ s.t. } m_1(z) = \dots = m_n(z)\}.$$

- *Inequity reducing principle:*

$$r^F(H) = \left\{s \in S(h): h \in H \setminus Z; z(s) \text{ s.t. } z \in \arg \min_{z \in Z} \left(\frac{1}{n} \sum_{i \in N} [m_i(z) - \bar{m}(z)]^2 \right) \right\}.^{76}$$

- *Classical-utilitarian welfare maximization principle:*

$$r^M(H) = \left\{s \in S(h): h \in H \setminus Z; z(s) \text{ s.t. } z \in \arg \max_{z \in Z} (\sum_{i \in N} m_i(z)) \right\}.^{77}$$

- *Rawlsian (minimax) welfare maximization principle:*

$$r^W(H) = \left\{s \in S(h): h \in H \setminus Z; z(s) \text{ s.t. } z \in \arg \max_{z \in Z} W(\mathbf{m}(z)) \right\},$$

where $W(\mathbf{m}(\hat{z}))$ denotes a *Rawlsian social welfare function* and is

defined as $W(m_1(\hat{z}), \dots, m_n(\hat{z})) = \min_{i \in N} \{m_1(\hat{z}), \dots, m_n(\hat{z})\}$.

It should be stressed that the above set does not aim at representing the whole range of social norms that may emerge in strategic interactions but is only meant to provide a useful (yet simple) illustration of the conditions under which conformity sets in, in mixed-motive games.

⁷⁶ Note that $\bar{m}(z)$ denotes the mean value of the players' material payoffs, for a given terminal node z .

⁷⁷ Obviously any allocation consistent with the classical-utilitarian welfare maximization principle is a Pareto-efficient solution.

II.6.a. Trust Games

Consider the following *trust game*: at the initial node h^0 , Player 1 (the “trustor”) chooses either “a” or “b” – when opting for “b”, the game terminates and material outcomes are allocated as shown in the vector of payoffs at the end-node $z(b)$ (with the number on top referring to Player 1’s payoff); on the other hand, if Player 1 opts for “a”, the choice passes to Player 2 (the “trustee”), who in turn can decide on “c” or “d”, the consequences of which are shown in the vector of payoffs at the end-nodes $z(c)$ and $z(d)$, respectively.

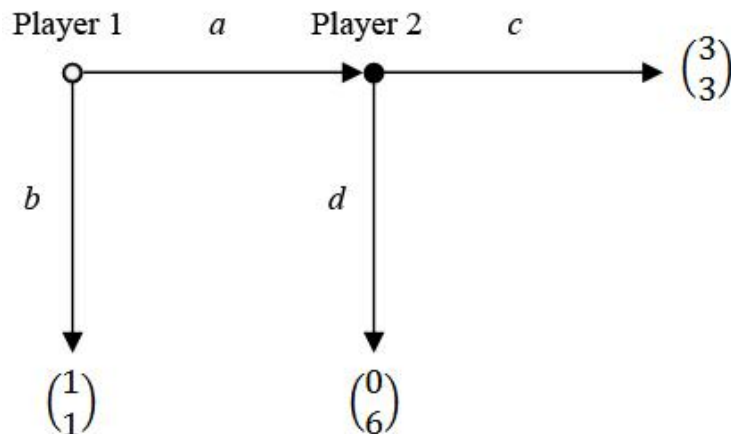


Figure II.2 - Trust Game “TG”

Now, assume that players believe that there may exist some social norm $r \in R_i$ for $\forall i \in N$: more precisely, assume that a norm dictating a strategy profile yielding an equal (material) payoff among players is the presumed social norm; thus, given the above set of norms, one may let $R_1 = R_2 = \{r^E\}$. It is clear that, at the initial node, the norm r^E dictates the following strategy profiles: $r^E(h^0) = \{(b, c), (a, c)\}$, i.e.: $r^E(h^0)$ is the set of strategy profiles

completely consistent with r^E . Recalling that a norm-conjecture of player i (definition II.2 above) is a collection of independent probability measures $\rho_i = (\rho_i(\cdot | h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} \Delta(A_{P(h)}(h))$ – with $\rho_i(a|h)$ being the probability of a at h , such that $\text{supp } \rho_i = \text{supp } (\rho_i(\cdot | h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} A_{P(h),h}(R_i(h))$ – the norm-conjecture induced by r^E , for $\forall i \in N$, can be represented by the following matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i(c|h(a)) & \rho_i(d|h(a)) \end{bmatrix} = \begin{bmatrix} \hat{\rho} & 1 - \hat{\rho} \\ 1 & 0 \end{bmatrix},$$

where $\hat{\rho} \in \{0,1\}$; that is, if $\hat{\rho} = 0$ then strategy profile (b, c) is implemented, whereas if $\hat{\rho} = 1$ then (a, c) is implemented.⁷⁸ Thus, recall that (if player i is a norm-driven decision maker and presumes that her co-player is norm-driven too) player i can form her belief α_i by assuming her co-player's behaviour to be consistent with her norm-conjecture ρ_i : i 's initial belief $\alpha_i = (\cdot | h^0)$ corresponds to a probability measure over the strategies of the opponent, with the support of α_i containing only the opponent's norm-complying strategies; for instance, here, the probability of a certain strategy of Player 2 is given by $\alpha_1(s_2|h^0) \equiv \Pr_{\rho_1}(s_2|h^0) = \rho_1(s_{2,h(a)}|h(a))$.

Then, Player 1 can calculate her expected payoff as well as the opponent's expected payoff and potential disutility from 1's not conforming to

⁷⁸ It is assumed that players cannot randomize, as per section II.4.b. above.

the norm (again assuming that $\rho_1 = \rho_2$). In brief, Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $E_{\rho_i, \alpha_1, \beta_1}[m_2|h^0] = 3\hat{\rho} + 1(1 - \hat{\rho}) = 2\hat{\rho} + 1$, for $i = 1, 2$. Further, Player 1's expectation of Player 2's disappointment at $(b|h^0)$ is $E_{\rho_i, \alpha_1, \beta_1}[m_2|h^0] - m_2(z(b)) = (2\hat{\rho} + 1) - 1 = 2\hat{\rho}$; this implies that Player 1's utility at $(b|h^0)$ equals $u_1^C(z, \rho_i, \alpha_1, \beta_1) = m_1(z(b)) - k_1[E_{\rho_i, \alpha_1, \beta_1}[m_2|h^0] - m_2(z(b))] = 1 - k_1(2\hat{\rho})$. On the other hand, Player 1's expected utility (=expected payoff) at $h(a)$ is given by $u_1^C(z, \rho_i, \alpha_1, \beta_1) = E_{\rho_i, \alpha_1}[m_1|a] = 1 \cdot 3 = 3$. Then, Player 1's conformist preferences can be expressed as $u_1^C(z(b), \rho_i, \alpha_1, \beta_1) \leq E_{\rho_i, \alpha_1}[m_1|a] \Rightarrow 1 - k_1(2\hat{\rho}) \leq 3 \Rightarrow k_1(\hat{\rho}) \geq -1$, which is always satisfied. Similarly, Player 2's expected utility at $(d|h(a))$ equals $u_2^C(z, \rho_i, \alpha_2, \beta_2) = m_2(z(a, d)) - k_2[E_{\rho_i, \alpha_2, \beta_2}[m_1|a] - m_1(z(a, d))] = 6 - 3k_2$, whereas Player 2's expected utility (=payoff) at $(c|h(a))$ is given by $u_2^C(z, \rho_i, \alpha_2, \beta_2) = m_2(z(a, c)) = 1 \cdot 3 = 3$. Then, Player 2's conformist preferences can be expressed as $u_2^C(z(a, d), \rho_i, \alpha_2, \beta_2) \leq u_2^C(z(a, c), \rho_i, \alpha_2, \beta_2) \Rightarrow 6 - 3k_2 \leq 3 \Rightarrow k_2 \geq 1$. To conclude – recalling that a socially consistent assessment is a social sequential equilibrium (definition II.8 above) if each probability measure $\Pr_{\sigma_i}(\cdot | h)$ assigns positive conditional probability only to conditional expected-payoff maximizing norm-complying strategies – it follows that the only norm-conjecture induced by r^E that yields an SSE is given by the matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i(c|h(a)) & \rho_i(d|h(a)) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

in which case (a, c) is an SSE for $k_2 \geq 1$; instead, if $k_2 < 1$, norm r^E is (not a social norm and is) not followed by population N (by remark II.2 above). It

should be further stressed that – given the norm subset $R_i = \{r^E\}$ for $\forall i \in N$ – in the case where $k_2 < 1$ no social sequential equilibrium of TG occurs (by remark II.3 above). Yet, for $k_2 < 1$, TG (with conformist preferences) has the same solution as the standard subgame perfect equilibrium, i.e.: (b, d) .

I shall now generalize the analysis by means of the following Standardized Trust Game, with parameters x and v such that $x > 0$ and $v > 1$.

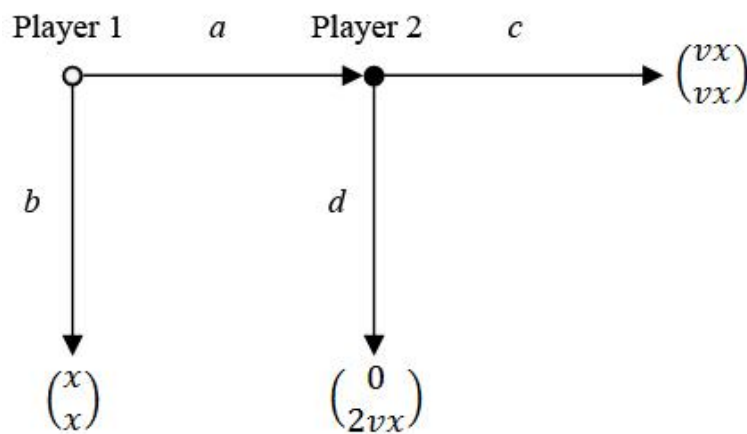


Figure II.3 - Standardized Trust Game “STG”

Proposition II.1. Given the norm subset $R_i = \{r^E, r^F\}$ for $\forall i \in N$, the only SSE of STG is (a, c) , whenever $k_2 \geq 1$.

Proof: the proof is analogous to that for TG (where $x = 1$, $v = 3$), and is therefore omitted.

The following results refer to alternative specifications of the norm subsets.

Proposition II.2. Given the norm subset $R_i = \{r^M\}$ for $\forall i \in N$, the following SSE of STG may occur:

- for $\rho_i(a|h^0) = 1, \rho_i(c|h(a)) = 1; (a, c)$, whenever $k_2 \geq 1$,
- for $\rho_i(a|h^0) = 1, \rho_i(c|h(a)) = 0; (a, d)$, whenever $k_1 \geq \frac{1}{2v-1}$.

Proof: see Appendix.

Proposition II.3. Given the norm subset $R_i = \{r^W\}$ for $\forall i \in N$, the only SSE of STG is (a, c) , whenever $k_2 \geq 1$.

Proof: see Appendix.

Notice that the second scenario in proposition II.2 (*i.e.*: SSE (a, d)) provides an example of an interaction where condition 2.2.a of definition II.6 above is fulfilled: in fact, Player 1 conforms to r^M because her cost of a norm violation is high enough to make 1's deviation from $s_1^* = a$ unprofitable (*i.e.*: if $k_1 \geq \frac{1}{2v-1}$), although there is one terminal history induced by a strategy profile $(s_2, s_1^*) = (c, a)$ – other than that implied by norm conjecture $\rho_i(a|h^0) = 1, \rho_i(c|h(a)) = 0$ – where Player 1 would be better off.⁷⁹ So from the viewpoint of Player 2, in the second scenario (*i.e.*: SSE (a, d)), Player 2 believes that Player 1 expects him to behave according to ρ_2 , simply because Player 2

⁷⁹ This implies that condition 2.2.b of definition II.6 above is not fulfilled in this case, because the second expression of condition 2.2.b, from the viewpoint of Player 2, does not hold here.

believes that Player 1 acknowledges the legitimacy of 2's norm-conjecture (*i.e.*: $\rho_2(a|h^0) = 1, \rho_2(c|h(a)) = 0$). Now, even though an equilibrium consisting of the strategy profile (a, d) may seem a controversial solution, this well captures many situations characterized by what I shall call “rationally-(mis)placed” trust: an example is given by a set of circumstances where a woman (marries and) brings a dowry to a man of dubious reputation; in effect, she (Player 1) may lucidly expect the man (Player 2) to use and invest the dowry, keeping all the proceeds for himself, and still, she may prefer to marry him if the local culture pushes women to get a husband; thus, if the cost of deviating is high (*i.e.*: if $k_1 = \frac{1}{2^{v-1}}$), the influence of culture (and its norms) is such that the woman is indifferent between remaining unmarried (*i.e.*: $(s_2, s_1) = (\cdot, b)$) and getting married-but-losing-everything (*i.e.*: $(s_2, s_1) = (d, a)$).

To sum up, the above exercise has shown that the range of equilibria obtained could possibly explain much of the experimental results (collected in Camerer [2003], pp. 83-100), based on the norm-conjectures induced by a variety of culture-dependent principles; indeed, different cultures may give prominence to different norms and, in turn, different conjectures about norms. The intuition is confirmed by a large cross-cultural experimental study undertaken in fifteen small-scale societies (Henrich *et al.* [2001]): the investigation directly addressed the question of whether the individual's social environments shape behaviour by implementing a study of behaviour in a set of social dilemma games; a number of field researchers, working in twelve countries on five continents, recruited subjects from small-scale

societies presenting a wide variety of economic and cultural conditions.⁸⁰ Their results show that *group*-level differences in the structure and organization of everyday economic activity explain a substantial part of the experimental variation observed across societies (the higher the degree of market integration and the higher the payoffs to cooperation of everyday life, the greater the level of cooperation in experimental games); moreover, and quite interestingly, *individual*-level economic and demographic variables do not explain observed behaviour either within or across groups.

II.7. Closing Remarks

This essay has presented an original theory of conformist preferences in mixed-motive games, building on Battigalli and Dufwenberg's [2007] model of guilt aversion; a notion of Social Sequential Equilibrium allowing for belief-dependent conformist motivations has been proposed by refining Battigalli and Dufwenberg's [2009] specification of the sequential equilibrium concept of Kreps and Wilson [1982]. Such a theory departs from the existing game-theoretic literature, since it explicitly defines (social) norms as rules that dictate a set of strategy profiles, where the norms considered here are informal, not necessarily enforced, and necessarily sustained by expectations. Although the motivational factors considered here are related to the much-investigated concepts of fairness and reciprocity, I shall stress

⁸⁰ Their sample comprised three foraging societies, six that practice slash-and-burn horticulture, four nomadic herding groups, and three (sedentary) small-scale agriculturalist societies.

that the focus of this study has been on a “mere” conformity motivation in social dilemmas, implying that the (presumed) behaviour of other members of a social group – be it fair or not – serves the individual as a means to guiding her own actions.

Further, the focus of this study has been on why people follow rules rather than on the specifics of what the rules are. This implies that the present theory can account for the reasons that have led to the establishment of a given norm, but not for the reasons that have led to the evolution of an individual’s norm subset (which is exogenously determined) and consequent norm-conjectures; notice that this is also due to the fact that the model partly relies on past compliance to explain future uniform behaviour.⁸¹ A justification for the present modelling approach comes from the assumption that – in the short run – one can treat the biological or cultural aspects of human nature as fixed; it also seems reasonable to assume that it is the players’ culture to mark out each player’s norm subset, with the norm subset containing rules of behaviour in accordance with set usage. Therefore, this theory implies a tendency for all agents (with conformist motivations) to conform to the “currently-normal” behaviour, which leads to the absence of evolution in the present setting.

The above considerations seem to undermine the predictive power of such a theory since it relies on an exogenous (culture-dependent) definition

⁸¹ In fact, the *empirical expectations* component of social norms is related to past compliance in that d_i^E equals one if the agent believes that at least one other agent will adhere to some norm in her norm subset, depending on the terminal history reached in a previous instance of the same multi-person decision problem.

of the norm subsets, implying that the system will not evolve away from its current position, *provided that* no exogenous variation in the beliefs about peers' behaviour occurs. On the other hand, the model suggests that – if social norms are based on beliefs and beliefs are, in effect, exogenously manipulated – it may be possible to induce pro-social behaviour at low cost. Indeed, a finely-tuned process of belief transmission could possibly favour the occurrence of the desired equilibrium.

For instance, social psychology research conducted at several U.S. universities shows that students hold exaggerated beliefs about the alcohol consumption habits of their peers (Berkowitz and Perkins [1986]). Such studies have concluded that students consume greater quantities of alcohol in order to fit in with their (biased) perceptions of acceptable social behaviour, that is, in order to comply with their presumed drinking norm in operation on campus. Research further shows that students that participate in a peer-oriented discussion focusing on correcting such inflated perceptions report drinking significantly less: in particular, a study from the *National Institute on Alcohol Abuse and Alcoholism*, an agency of the United States Department of Health and Human Services, reports that several educational institutions that persistently communicated actual student norms have experienced reductions of up to twenty percent in high-risk drinking over a relatively short period of time (NIAAA [2002]).

To conclude, future research should delve into the study of alternative systems of belief elicitation and transmission: only through a full understanding of the mechanics of social norms, an institution or policy-maker will be able to predict how external signals may alter beliefs and drive behaviour towards more socially desirable outcomes.

II.8. Appendix II

II.8.a. Proofs

Proof of Proposition II.2.

Given the norm subset $R_i = \{r^M\}$ for $\forall i \in N$, at the initial node the norm r^M dictates the following strategy profiles: $r^M(h^0) = \{(a, c), (a, d)\}$. Hence, the norm-conjecture induced by r^M , for $\forall i \in N$, can be represented by the following matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i(c|h(a)) & \rho_i(d|h(a)) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \hat{\rho} & 1 - \hat{\rho} \end{bmatrix},$$

where $\hat{\rho} \in \{0,1\}$; that is, if $\hat{\rho} = 0$ then strategy profile (a, d) is implemented, whereas if $\hat{\rho} = 1$ then (a, c) is implemented. Then, Player 1's expectation of Player 2's (expected) material payoff at h^0 equals $E_{\rho_i, \alpha_1, \beta_1}[m_2|h^0] = vx\hat{\rho} + 2vx(1 - \hat{\rho}) = 2vx - vx\hat{\rho}$, for $i = 1, 2$. Further, Player 1's expectation of Player 2's disappointment at $(b|h^0)$ is $E_{\rho_i, \alpha_1, \beta_1}[m_2|h^0] - m_2(z(b)) = (2vx - vx\hat{\rho}) - x$; this implies that Player 1's utility at $(b|h^0)$ equals $u_1^C(z, \rho_i, \alpha_1, \beta_1) = m_1(z(b)) - k_1[E_{\rho_i, \alpha_1, \beta_1}[m_2|h^0] - m_2(z(b))] = x - k_1(2vx - vx\hat{\rho} - x)$. On the other hand, Player 1's expected utility (=expected payoff) at $h(a)$ is given by $u_1^C(z, \rho_i, \alpha_1, \beta_1) = E_{\rho_i, \alpha_1}[m_1|a] = vx\hat{\rho}$. Then, Player 1's conformist preferences can be expressed as $u_1^C(z(b), \rho_i, \alpha_1, \beta_1) \leq E_{\rho_i, \alpha_1}[m_1|a] \Rightarrow x - k_1(2vx - vx\hat{\rho} - x) \leq vx\hat{\rho} \Rightarrow \begin{cases} k_1 \geq -1 & \text{if } \hat{\rho} = 1 \\ k_1 \geq 1/(2v - 1) & \text{if } \hat{\rho} = 0 \end{cases}$, where the first case is always satisfied. Similarly, Player 2's expected utility at $(d|h(a))$ equals $u_2^C(z, \rho_i, \alpha_2, \beta_2) = m_2(z(a, d)) - k_2[E_{\rho_i, \alpha_2, \beta_2}[m_1|a] - m_1(z(a, d))] = 2vx -$

$k_2 vx \hat{\rho}$, whereas Player 2's expected utility (=payoff) at $(c|h(a))$ is given by $u_2^c(z, \rho_i, \alpha_2, \beta_2) = m_2(z(a, c)) = vx$. Then, Player 2's conformist preferences can be expressed as $u_2^c(z(a, d), \rho_i, \alpha_2, \beta_2) \leq u_2^c(z(a, c), \rho_i, \alpha_2, \beta_2) \Rightarrow 2vx - k_2 vx \hat{\rho} \leq vx \Rightarrow \begin{cases} k_2 \geq 1 & \text{if } \hat{\rho} = 1 \\ \text{not defined} & \text{if } \hat{\rho} = 0 \end{cases}$. Therefore, the following SSE of *STG* may occur: (a, c) , whenever $k_2 \geq 1$, for $\hat{\rho} = 1$; (a, d) , whenever $k_1 \geq \frac{1}{2v-1}$, for $\hat{\rho} = 0$. ■

Proof of Proposition II.3.

A Rawlsian social welfare function is defined as $W(m_1(\hat{z}), \dots, m_n(\hat{z})) = \min_{i \in N} \{m_1(\hat{z}), \dots, m_n(\hat{z})\}$. Such a function has to be evaluated at each of the three leaves of the game tree, *i.e.*: $W(m_1(z(b)), m_2(z(b))) = \min_{i \in N} \{x, x\} = x$; $W(m_1(z(a, d)), m_2(z(a, d))) = \min_{i \in N} \{0, 2vx\} = 0$; $W(m_1(z(a, c)), m_2(z(a, c))) = \min_{i \in N} \{vx, vx\} = vx$. It follows that here the set of maximizers of W is singleton: hence, given the norm subset $R_i = \{r^W\}$ for $\forall i \in N$, at the initial node the norm r^W dictates the strategy profile $r^W(h^0) = \{(a, c)\}$ only; then, the norm-conjecture induced by r^W , for $\forall i \in N$, can be represented by the following matrix:

$$\rho_i = \begin{bmatrix} \rho_i(a|h^0) & \rho_i(b|h^0) \\ \rho_i(c|h(a)) & \rho_i(d|h(a)) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

The rest of the proof is trivial. ■

II.8.b. A review of alternative theories of norm compliance

López-Pérez [2008] proposes a model of norm compliance which builds on Charness and Rabin's [2002] in that it assumes players to have a taste for fairness and efficiency, and to be influenced by previous history as well. In brief, López-Pérez's model applies to any extensive form game of perfect recall, where $N = \{1, \dots, n\}$ is the set of players, $u(z)$ is a vector of the players' utility at terminal node z , and $x(z)$ is a vector of the players' (material) payoffs at z . The main feature here is the explicit introduction of a norm, defined as a non-empty correspondence $\psi: h \rightarrow A(h)$ that applies to any information set h : thus, $a \in A(h)$ is said to be consistent with norm ψ if $a \in \psi(h)$, otherwise a is a deviation from ψ ; the interpretation is that a norm is a prescription indicating how a player should move at a decision node. Given that, López-Pérez examines a specific norm, namely the "Efficiency and equity norm" (*E-norm*), on the assumption that the E-norm is the only norm in the society; he further assumes that there exist two types of agents, that is, "selfish" and "principled", where the former ignore the E-norm while the latter have internalized the E-norm (and suffer from violating it). The size of the emotional cost of a deviation is assumed to depend directly on the number of norm followers, so a principled player's utility function takes the form:

$$u_i(z) = \begin{cases} x_i(z) & \text{if } i \in R(z) \\ x_i(z) - \gamma \cdot r(z) & \text{if } i \notin R(z) \end{cases}$$

(2.8.1)

where $R(z)$ is the set of players that complied with the norm in the history leading to z ,⁸² $r(z)$ denotes the cardinality of $R(z)$, and γ is a positive parameter indicating how intensely principled types have internalized the norm. Given (an) initial decision node t_0 , and denoting by $X(t_0)$ the set of all $x(z)$ succeeding t_0 ,⁸³ López-Pérez defines a “fairmax distribution” (ε, δ) as an allocation $x = \{x_1, \dots, x_n\} \in X(t_0)$ that maximizes the function:

$$F_{\varepsilon\delta} = \varepsilon \cdot \sum_{i \in N} x_i - \delta \left(\max_{i \in N} x_i - \min_{i \in N} x_i \right), \tag{2.8.2}$$

over $X(t_0)$; a path connecting t_0 and one of its fairmax distributions (ε, δ) is said to be a “fairmax path”. It is assumed that if information set h has at least one node on a fairmax path, then the E-norm selects all actions of $A(h)$ that belong to a fairmax path; otherwise, any action becomes commendable, that is, the E-norm selects the whole set $A(h)$. The author then normalizes the efficiency parameter ε to one and keeps δ strictly positive but smaller than one, in order to show that social efficiency (captured by the sum of material payoffs) is relatively more important than equality. Finally, López-Pérez

⁸² More specifically, $R(z)$ includes all the players who acted consistently with the E-norm or made no choice at all in the history of z .

⁸³ In the event of an initial chance move, t_0 denotes any node immediately succeeding any such move.

assumes that each player's type (*i.e.*: selfish or principled) is private knowledge, while the objective probability μ of being a principled agent is common knowledge; he thus applies a standard perfect Bayesian equilibrium concept to explain experimental evidence coming from a variety of games. To conclude, the main advantage of this model is its tractability, in addition to having the merit of defining a norm explicitly; on the other hand, disregarding the role of expectations in sustaining a social norm seems to be its major drawback.

In the same line of research Li [2008] develops a more complex model of norm compliance, building on the psychological game framework of Geanakoplos *et al.* [1989]. The model is designed for 2-player normal form games only, where S_i and Σ_i are Player i 's sets of pure strategies and mixed strategies, respectively; $\pi_i(\sigma)$ indicates Player i 's material payoff when strategy profile $\sigma = (\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$ is played. Letting $b_i \in \Sigma_i$ and $c_i \in \Sigma_i$ denote j 's first-order belief and i 's second-order belief, respectively, Li models a norm (or "convention" in her terminology) as belief-dependent rankings over the players' strategy space, normalizing them to the interval $[0,1]$. In formulae, a norm is a mapping $\omega: \Sigma_1 \times \Sigma_2 \rightarrow [0,1]^2$ such that, for $i, j = 1,2$:

$$\begin{aligned}
(i) \quad & \text{for each } b_j \in S_j, \text{ either } \omega_i(s_i, b_j) = 1 \text{ for } \forall s_i \in S_i, \\
& \text{or } \max_{s_i \in S_i} \omega_i(s_i, b_j) = 1 \text{ and } \min_{s_i \in S_i} \omega_i(s_i, b_j) = 0; \\
(ii) \quad & \text{for } \forall \sigma_i \in \Sigma_i, \forall b_j \in \Sigma_j, \omega_i(\sigma_i, b_j) = \sum_{s_i} \sum_{s_j} \sigma_i(s_i) b_j(s_j) \omega_i(s_i, s_j).
\end{aligned}
\tag{2.8.3}$$

The author refers to the number $\omega_i(s_i, b_j)$ as the “social index” of i taking action s_i given belief b_j , and interprets it as a measure of the adequateness of i ’s action: condition (i) says that the function ω_i ranks Player i ’s pure actions given i ’s belief of j ’s action (with the first line depicting the case of a “trivial convention” where all the pure actions are equally adequate and assigned $\omega_i(s_i, b_j) = 1$); condition (ii) simply generalizes the norm specification to account for mixed strategies. Furthermore, given common knowledge of the norm, “ i ’s belief of the social index of j ’s action” is denoted by $\omega_j(b_j, c_i)$, whereas $f_i(\sigma_i, b_j, c_i) = \max\{0, \omega_i(\sigma_i, b_j) - \omega_j(b_j, c_i)\}$ indicates “ i ’s belief about how much more her own action conforms to the convention compared to her opponent’s”. Player i ’s utility function takes the form:

$$u_i(\sigma_i, b_j, c_i, \omega) = \pi_i(\sigma_i, b_j) + \theta_i \left[g_i(\omega_i(\sigma_i, b_j)) + h_i(f_i(\sigma_i, b_j, c_i)) \right], \quad (2.8.4)$$

where – $g_i, h_i: [0,1] \rightarrow \mathbb{R}$ are continuous and have the following properties – g_i is increasing in ω_i (“conformity effects”), h_i is decreasing in f_i (“interaction effects”),⁸⁴ $g_i + h_i$ is concave in ω_i . Hence, players’ utility depends both on material payoffs and social implications of their strategies (as captured by the expression in square brackets), with $\theta_i \in [0, \infty)$ indicating how salient the convention is to Player i ($\theta = 0$ representing the traditional agent). Li adapts

⁸⁴ The interaction effects of conventions capture the fact that a player prefers the opponent to comply with the norm.

the psychological-Nash equilibrium concept of Geanakoplos *et al.* [1989] to define a “social equilibrium”, that is, a Nash equilibrium satisfying an additional consistency condition that all beliefs correspond to actual strategies; the author then turns to examine a norm embodying both a principle of efficiency and fairness,⁸⁵ thereby applying it to some symmetric and public goods games. To sum up, Li’s theory proves to be an insightful model for the analysis of conformity in mixed-motive games: her approach draws on Geanakoplos *et al.*’s [1989] framework as well as on Charness and Rabin’s [2002] theory of quasi-maximin preferences, the result being a model more complicated than López-Pérez’s [2008] but also more realistic (as it crucially captures the conformity and interaction effects of conventions); on the other hand, its main drawback is to be designed for 2-player normal form games only.

⁸⁵ Measures for efficiency and fairness are defined in a way relatively similar to the distributional preferences of Charness and Rabin [2002].

III. A Test for Conformist Motivations in Experimental Games

III.1. Introduction

The first tests for conformity were conducted by Solomon E. Asch [1955, 1956], a pioneer in social psychology who undertook a series of small-group studies on the social pressures to conform: his experimental subjects were asked to answer a basic puzzle on the length of lines, while others provided an obviously incorrect answer – all but one of the participants in each session were confederates of the experimenter and had beforehand been instructed to give wrong answers in unanimity at certain points – as a result, many subjects (approximately 35%) felt under pressure to give the same incorrect answer as the misleading majority. The social psychology literature defines conformity as the act of changing one's behaviour to match the purported beliefs of others (Cialdini and Goldstein [2004]); yet, while the psychology literature offers plenty of experimental evidence of conformist behaviour, very few of their insights have been adopted by economists to describe the social pressures to conform in *problems of strategic interdependence*. In this respect, the present investigation sheds some light on conformity as a strategically-relevant motivation in social dilemmas: a motivation which implies that the peers' presumed behaviour serves the individual as a means of forming beliefs and taking actions (in games where one another's actions and beliefs enter one another's utility functions).

Hence, unlike the classical social psychology experiments, here conformity is put to test in a problem of strategic interdependence: what is being hypothesized is that a conformist player will behave as she thinks that other (conformist) players in the same role behave, in a social dilemma. Thus, the first hypothesis to test is that the experimenter can predict a conformist's behaviour from the conformist's guess about the behaviour of other players in the same role. Now, it should be noted that a false consensus effect hypothesis will predict an analogous correlation between

beliefs and behaviour, although with an inverse causal relationship: in fact, false consensus is usually referred to as an egocentric bias that occurs when people estimate consensus for their own behaviour (Ross *et al.* [1977]); so, in the case of false consensus effects, when forming a belief about the others' behaviour, a player will estimate the others' decisions based on her own decision.

In a nutshell, in order to disentangle consensus from conformity, one of the experimental treatments introduces an exogenous variation in beliefs by showing subjects some aggregate information about the others' beliefs. Indeed, if the experimenter can predict a subject's choice from the subject's guess (about the behaviour of other participants in the same role) in conjunction with the subsequently transmitted information about others' guesses, then one has effectively disentangled consensus from conformity and provided evidence in support of a conformity hypothesis. In fact, if false consensus is present, then there will be a causal relationship from behaviour to beliefs, and thus there will not be an effect of providing exogenous information; but if, on the other hand, conformity is present (in which case the causality runs from beliefs to behaviour), one will find that exogenously varying beliefs has an impact on behaviour.

More specifically, the experiment measures the impact of the beliefs of players in the same role, on behaviour, in a discrete Trust Game. Note that – in order to introduce an exogenous variation in beliefs, and to ensure that some subjects received information about an average belief of low cooperation and some others received information about an average belief of high cooperation – each subject was shown the average guess made by a specifically selected sample of participants: more precisely, the samples were selected in such a way that the beliefs transmitted to each participant were in the vicinity of either 25% or 75% (with 100% indicating the belief that all subjects in the same role will cooperate). Notice that the design does not

involve deception because the transmitted information about the others' guesses explicitly stated that such information referred to a *sample* of other participants.

In brief, the data show that the transmitted information can indeed influence one's behaviour, with the strength of the impact depending on one's prior (*i.e.*: stated) beliefs. By regressing a subject's choice on the stated belief, the transmitted information and their interaction, the probit model predicts a probability of cooperating equal to 0.55 for Trustees who received a low transmitted belief and 0.72 for those who received a high transmitted belief.

The remainder of the essay is organized in this manner: III.2. briefly reviews the "conventional" economic account of conformity, before exploring some related belief-dependent motivations (with a special focus on Trust Game experiments); III.3. discusses the implications of consensus and conformity in more detail, hence presents the experimental design, procedure and hypotheses; III.4. discusses the experimental results, and III.5. concludes.

III.2. Tests for conformity and related belief-dependent motivations

Conformity implies that the peers' presumed behaviour serves the individual as a means of forming beliefs and taking actions; hence, "conformist preferences" thrive on behavioural expectations within a society or group. Now, just as much of the social psychology literature revolves around variations of Asch's [1955, 1956] experiments, so have many economic studies on conformity drawn on the theory of *informational cascades*. An informational cascade (otherwise referred to as "herding behaviour") occurs

when individuals observe the predecessors' actions and then make the same choice that others have made, irrespective of their own private information signals: the idea, which is based on observational learning theory, was formally developed by Banerjee [1992] and Bikhchandani *et al.* [1992], and has been experimentally investigated in several studies since the seminal Anderson and Holt [1997]. Yet, while such studies are primarily focused on the effects of the transmission of information on conformist behaviour – in a “non-strategic” setting – the present essay follows a different approach in that, unlike informational cascade models, here an individual's payoff directly depends on what all preceding or subsequent players do (in other words, not only does an individual's action influence the others' choice of action, but also co-determines one another's payoff).⁸⁶

In this connection, substantial pieces of research have been delving into the effects of a different belief-dependent motivation, which nevertheless turns out to be related to conformity in mixed-motive games, that is, *guilt aversion*. In effect, Charness and Dufwenberg's [2006] investigation is fairly connected to the present test: the guilt aversion hypothesis implies that players may feel guilty if their behaviour falls short of the others' expectation; hence, in order to seek evidence of such a motivation, the experimenter asks subjects what they believe their opponents expect (thereby collecting

⁸⁶ The payoff structure of informational cascade models involves everyone who chooses the right option getting the same reward, irrespective of how many others have chosen that option before and after them. (An alternative specification of the payoff structure may allow for the total reward to be fixed or may grant extra rewards for being first or second to choose the correct option.)

second-order beliefs).⁸⁷ More specifically, after collecting the strategic choices of both the Trustor and the Trustee,⁸⁸ Charness and Dufwenberg have subjects make guesses about the choices of players in a different role (and offer to reward good guessers): Trustors are asked to guess the “proportion of Trustees who choose to cooperate”, while Trustees are asked to guess the “average guess made by Trustors who choose to cooperate”; notice that such guesses represent the experimenter’s measurement of the Trustor’s first-order beliefs and the Trustee’s second-order beliefs, respectively. To sum up their findings, the guilt aversion hypothesis is confirmed by a strong correlation between beliefs and behaviour: the Trustees who cooperated made significantly higher guesses about the Trustors’ guesses than did the Trustees who chose not to cooperate (*i.e.*: the Trustees who cooperated held significantly higher second-order beliefs regarding their choice to cooperate).

⁸⁷ Note that, while Charness and Dufwenberg’s study is primarily conceived to examine the role of pre-play communication (in the form of nonbinding promises, transmitted through written free-form messages) in shaping beliefs and enhancing cooperative behaviour, it does nevertheless disclose an interesting relationship between beliefs and choices; indeed, their results reveal significant correlations between second-order beliefs and actions, even in the treatments without pre-play communication (note that Dufwenberg and Gneezy [2000] provide an analogous test in a similar experimental game setting without pre-play communication).

⁸⁸ They examine one-shot non-simultaneous Trust Games with hidden action (*i.e.*: if the Trustee chooses to fulfil trust and cooperates, a chance move will determine whether the Trustor gets some material payoff with probability 5/6; the Trustor gets a payoff of zero otherwise); note that they use the *strategy method*.

Given that, of particular interest is Ellingsen *et al.*'s [2010] design: again, their aim is to test for guilt aversion but in a way that reduces the scope for consensus effects which, in their view, may have driven such a strong correlation between beliefs and behaviour. According to their argument, Charness and Dufwenberg's [2006] test is weak as the observed correlation does not prove that second-order beliefs affected behaviour (as postulated by guilt aversion); instead, Ellingsen *et al.*'s hypothesis is that those Trustees who made large back transfers would hold higher second-order beliefs (regarding their choice to cooperate), just because they thought that everyone would (expect to) behave like them.⁸⁹ Hence, in order to rule out that eventuality, Ellingsen *et al.* provide each Trustee with straightforward information about the co-paired Trustor's first-order belief.⁹⁰ (Notice that, from the viewpoint of the Trustee, this represents her induced second-order belief; also note that the experimenter omits to tell Trustors that Trustees will have access to their beliefs so as to avoid some strategic, untruthful reporting of beliefs.) More specifically, in Ellingsen *et al.*'s design, first each Trustor chooses whether to continue or to withdraw, then she guesses what fraction of the Trustees will choose to cooperate; after that, each Trustee is informed about the guess of the respective Trustor, and therefore decides whether to cooperate or not. In brief, their results are interesting, because – in contrast to Charness and Dufwenberg [2006] – Ellingsen *et al.* do not find a

⁸⁹ I shall return to Ellingsen *et al.*'s interpretation of Charness and Dufwenberg's results in the next section.

⁹⁰ Similarly, in their Dictator Game treatment, the Recipients' first-order beliefs are communicated to the respective Dictators.

correlation between (induced) second-order beliefs and behaviour, which seems to deny a guilt aversion motivation.

I conclude this section with studies testing for a related motivation, namely *trust responsiveness*, which is defined as a «tendency to fulfil trust because you believe it has been placed in you» (Bacharach *et al.* [2007], p. 350). Thus, in the treatments of Bacharach *et al.* [2007] and Guerra and Zizzo [2004], first beliefs from Trustors are elicited, and then the experimenter informs Trustees about the mean prediction made by the Trustors with whom they are not matched (*i.e.*: each Trustee receives a report about the mean value of the first-order beliefs of her non co-paired players). More specifically, in Bacharach *et al.*'s [2007] design, each Trustee is shown the mean value of the first-order beliefs of the Trustors with whom she is not matched, before the experimenter elicits each Trustee's second-order belief about her co-paired player. (Notice that, while this could provide an interesting exogenous variation in beliefs – since each Trustee may use the information about her non co-players' first-order beliefs, in order to update her second-order belief about the co-paired player – the authors do not investigate such an issue: in fact, to that end, each Trustee's second-order belief should rather be elicited before she is informed about the non co-players' guesses.) Lastly, Bacharach *et al.*'s results show that Trustees' cooperative behaviour varies with Trustees' second-order beliefs, in two of their experimental game variants, but unfortunately they do not further explore the relationship between Trustees' behaviour and the transmitted information, which could have given a good insight into the conformity motivation.

In the next section I shall further discuss how the above belief-dependent motivations relate to conformity: the discussion of the conformity hypothesis will lead the way to a novel experimental design.

III.3. Experimental design

III.3.a. False consensus effects vs. conformity to social norms

This study aims at disentangling conformity from consensus effects. In the theory of conformity to social norms that I have previously formalized, I assume that some subjects have conditional preferences for conformity to a norm, with the norm dictating a set of strategy profiles which depict the “currently-normal” or “appropriate” behaviour within a certain social group. Hence, in order for a potentially conformist player to identify the strategy profile/s that best describe(s) some normal behaviour, she forms a conjecture about the norm-complying actions at each history. To sum up, in such a theory, a conformist player is motivated by her beliefs about others’ beliefs and behaviour, and consequently adapts to her presumed social norm. Now, the present experiment does not provide a direct test for such a theory of conformity, because for that – before treating the subjects – one needs to isolate those with conformist preferences from those with classical, non-conformist motivations. However, the present experiment does shed some light on conformity as a motivation, which implies that the peers’ presumed behaviour serves the individual as a means of forming beliefs and taking actions.

A conformist player, in a social dilemma, would do her part of what is presumed to be the norm-complying strategy profile. It should be recalled that in the aforementioned theory, for a social norm to be in operation a certain set of conditions must hold, where such conditions involve *empirical* as well as *normative* expectations of conformity to the norm (see also Bicchieri [2006], Ch. 1). Now, in this study I shall focus on the empirical expectations, that is, a player’s beliefs about the others conforming to a given rule of behaviour; more specifically, the experiment will assess the

importance of the expectations regarding the behaviour of other participants in the same role.⁹¹

Thus, what is being hypothesized here is that a conformist player will behave as she thinks that other (conformist) players in the same role behave. Hence, the key hypothesis is that the experimenter should be able to predict a conformist's behaviour from the conformist's beliefs about the behaviour of other players in the same role: to that end, the first two treatments (*i.e.*: "T0" and "T1") are designed to test for any such relationship by asking subjects to guess how many of the other participants in the same role will choose one of two actions of a discrete Trust Game.

Now, it should be noted that a false consensus effect hypothesis will predict an analogous relation (although an *inverse* one). False consensus is usually referred to as an egocentric bias that occurs when people estimate consensus for their own behaviour: in that case, when forming a belief about the others' behaviour, a player will estimate the others' decisions based on her own decision (*i.e.*: according to the traditional definition of false consensus, a player who chooses a certain action – with a higher probability than some other player g does – will give a higher estimate of a third player j choosing that action than the estimate given by player g);⁹² or else – if some player g 's decision is observable – false consensus implies that an individual

⁹¹ Note that normative expectations are distinct from second-order empirical expectations in that the former embody an ought-to-do statement. See Bicchieri and Xiao [2009] for a test of both empirical and normative expectations of conformity to a norm, in a Dictator Game.

⁹² Dawes [1990] notes that the traditional definition does not justify the label "false": for this reason, Engelmann and Strobel [2000] provide the above alternative definition.

i , when forming a belief about the behaviour of a third player j , will estimate it by attaching more weight to her own decision than to that of g (Engelmann and Strobel [2000]). Given that, empirically conformity and consensus are related in that both imply that a subject's guess regarding the number of other people choosing a certain strategy is higher, if the guess is made by subjects choosing that very strategy (compared with the guess made by subjects choosing some other strategy). Therefore, on consideration the aforementioned results of Charness and Dufwenberg [2006] might have been driven by a conformity motivation rather than consensus (the latter being argued by Ellingsen *et al.* [2010]).⁹³

To sum up, both conformity and consensus imply that in a discrete Trust Game, say, a second-mover's guess regarding the fraction of other second-movers choosing a certain action is higher, if the guess is made by second-movers choosing that very action (compared with the guess made by second-movers choosing another action). Therefore, in order to disentangle consensus from conformity, a third treatment (*i.e.*: "T2") introduces an exogenous variation in beliefs by showing subjects some aggregate

⁹³ In this connection, it should be stressed that *guilt aversion* and *conformity in mixed-motive games* are related since – in the aforementioned theory of social norms – guilt is the feeling that generates conformity, but only if a certain set of conditions holds: in fact, a conformist player is usually motivated by *conditional* social preferences. Note that, according to such a theory, the expectations regarding the behaviour of all other participants in the experiment matter, whatever role they are assigned; instead, to a purely guilt-averse individual (in the sense of Battigalli and Dufwenberg [2007]) only the expectations of the (other-role) subject with which one is matched matter. In order to rule out any confounding element, the present experiment focuses on the (empirical) expectations regarding the behaviour of other participants in the same role.

information about the others' guesses (*i.e.*: the average guess made by a sample of other participants in the same role).⁹⁴ In a nutshell, if the experimenter can predict a subject's choice from the subject's guess (about the behaviour of other participants in the same role) in conjunction with the transmitted information about others' guesses, then one has effectively disentangled the two effects. More precisely, in a player affected by consensus one should find a relationship between her behaviour (and, underweighted, the observed behaviour of others) and her beliefs about some others' behaviour; yet, in a player affected by consensus, one should *not* find a relationship between her behaviour and some others' beliefs about others' behaviour. Again, in the case of conformity, the causal relationship runs from beliefs to behaviour; instead, in the case of consensus, one's action influences one's beliefs about some others' actions. It is therefore clear that if, in the lab, a subject's guess about some others' actions is elicited before the very subject is shown the average guess made by other players, then only the conformity hypothesis can be consistent with that subject using (also) the transmitted information to decide on the action to take.⁹⁵

⁹⁴ Notice that such an exogenous variation in beliefs – to a potentially conformist player – represents an exogenous variation in empirical expectations of conformity to a norm.

⁹⁵ In this respect, it should be noted that Ellingsen *et al.*'s design (while ruling out consensus) would not be apt to investigate how conformity works in social dilemmas because, in their Trust Game experiment, Ellingsen *et al.* transmit the sole information about each co-paired player's guess. Instead, social psychologists describe conformity as the act of matching beliefs, attitudes and behaviours to what individuals perceive is normal of their group or society, as opposed to what one other individual expects.

III.3.b. Game specification and treatments structure

Here I shall focus on the below discrete version of the Trust Game since a substantial amount of the prior evidence on suspected conformist behaviour comes from this family of games and it is therefore easier to compare results.

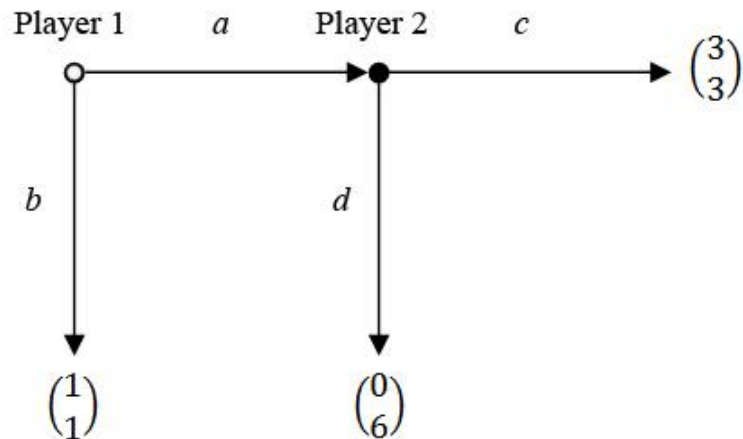


Figure III.1 - Trust Game “TG”

Note that in each experimental session players were referred to as “Participant A” or “Participant B”: at the initial node h^0 Player 1 (*i.e.*: Participant A, in the lab) chooses either a or b – when opting for b , the game terminates and material outcomes are allocated as shown in the vector of payoffs at the end-node $z(b)$ (with the number on top referring to Player 1’s payoff); on the other hand, if Player 1 opts for a , the choice passes to Player 2 (*i.e.*: Participant B, in the lab), who in turn can decide on c or d , the consequences of which are shown in the vector of payoffs at the end-nodes $z(c)$ and $z(d)$, respectively. Given that, each experimental session was comprised of one of three treatments: I shall refer to the treatments as “T0”, “T1”, and “T2”.

T0. Each instance of T0 is divided into three stages, as follows (*i.e.*: Introduction Stage; Play Stage I-II; Payment Stage).

Introduction Stage – subjects were randomly allocated to terminals and given the paper instructions; there, they were told that (in Part I) they would be assigned one of two roles (hence, randomly matched with a participant in a different role), and explained the decisions involved in each role; each subject was then asked to answer a set of control questions; a summary of the instructions was finally read aloud by the experimenter.

Play Stage, Part I – all plays were conducted using the “strategy method”. The order of subsequent tasks was as reported below:

- (i) subjects were assigned the role of Participant B;⁹⁶
- (ii) each subject was asked to guess how many of the other Participants B (in the same session) would choose either *c* or *d* (labelled as “share” and “keep”, respectively); subjects entered their guess by positioning a slider to the desired percentage;⁹⁷
- (iii) each subject was invited to wait until all participants had entered their guesses;
- (iv) each subject was asked to choose either *c* or *d* (*i.e.*: “share” and “keep”, respectively).

⁹⁶ Given that each subject was *privately* assigned a role, each subject did not know that every other person in Part I was assigned the role of Participant B.

⁹⁷ The slider was initially positioned at a value of 50%: subjects had to enter a guess by moving it towards a higher *share* rate (*i.e.*: towards a value of 100%) or towards a higher *keep* rate (*i.e.*: towards a value of 0%). Note that subjects could not leave the slider in the initial position.

Play Stage, Part II – subjects were told that Part II involved exactly the same steps as in Part I, although they would be assigned a different role and matched with a different participant; after they had been given a brief reminder of the instructions (both on-screen and orally), subjects were assigned the role of Participant A. Steps (i)-(iv) of Part II had the same structure as above, although each subject's decision and guess were about a or b (labelled as “in” and “out”, respectively).

Payment Stage – the payment mechanism consists of two parts:

- each subject received a £3 show-up fee;
- each subject was paid according to the outcome, as shown in the vector of payoffs, at the end-node realized in *Part I* as well as the end-node realized in *Part II* (payoffs were in pound sterling).

A few comments are now due. First, it should be noticed that the above order of the decisions (which is reversed with respect to the natural sequence Player 1, Player 2) is made possible by the adoption of the strategy method; also, it should be stressed that in Part II each subject was randomly matched with a participant other than that she was matched with in Part I; besides, subjects did not know about the tasks to be undertaken in Part II until the end of Part I. Obviously, notice that subjects did not know how much they had earned in Part I until the end of Part II (because every subject in each part is assigned the same role).

T1. Given that an important element of conformist preferences involves expectations regarding the behaviour of other subjects in the same role – in order to induce participants to state their true beliefs – a few sessions (*i.e.*: treatment T1) presented an incentivizing scheme as follows. As in Part I of T0, before entering their decision as Participant B, each subject was asked to guess how many of the other Participants B (in the same session) would choose to transfer half the money back: yet, in *Part I* of

T1 subjects were also told that they would receive an additional payment of £2 if their guess differed by no more than 5 percentage points from the realized value. Moreover, as in Part II of *T0*, before entering their decision as Participant A, each subject was asked to guess how many of the other Participants A (in the same session) would opt in: yet, in *Part II of T1* subjects were also told that they would receive an additional payment of £2 if their guess differed by no more than 5 percentage points from the realized value. To sum up, each instance of *T1* has exactly the same structure as *T0*, except for the incentivizing scheme for the elicitation of beliefs.⁹⁸

T2. Each instance of *T2* has exactly the same structure as *T0*, except for step *(iii)* of Play Stage I-II, which in *T2* was as follows:

- (iii)* each subject was invited to wait until all participants had entered their guesses, after which each subject was given feedback about other Participant B's guesses (*about other Participant A's guesses* in Part II).

Note that such feedback consisted of the average guess made by a sample of other participants in the same role (in the same session); also, when entering their guesses (at step *(ii)*), subjects did not know that those guesses would be pooled and transmitted to other participants (at step *(iii)*). Further, notice that such feedback was shown in the lower part of the same screen in which subjects were asked to enter their guesses: the message was phrased in such a way as to look like the outcome of an opinion survey; the font style

⁹⁸ If a subject's guess differed by more than 5 percentage points from the realized value, that subject would receive no additional belief-payment; note that each subject was not informed about the correctness of her guesses until the end of Part II.

and size were the same as those of the other messages, in order not to make the information too prominent. In Part I the message read: «A sample of other participants B in this session expects on average that $\langle x \rangle\%$ will transfer half the money, whereas $\langle 100-x \rangle\%$ will keep all the money»; similarly, in Part II the message read: «A sample of other participants A in this session expects on average that $\langle x \rangle\%$ will OPT IN, whereas $\langle 100-x \rangle\%$ will OPT OUT» (see Appendix III for a transcript of all the on-screen messages).

Lastly, given that (in Part I of T2) each subject was provided with an aggregate measure of the guesses made by other subjects in the role of Participant B, there might be interaction effects between Part I and Part II of T2: in effect, (in Part II of T2) some Participant A's choice might be affected by the fact that that very subject got some information in Part I. For this reason – and also because of the small difference in payoff between (b, \cdot) and (a, d) , which encourages Participants A to choose a – as in other studies (*e.g.*: Ellingsen *et al.* [2010]) the focus of this treatment will be on Trustees' rather than on Trustors' behaviour.

III.3.c. Hypotheses

Before stating the key hypotheses of the experiment, I shall recall that the conformity motivation implies that a conformist player will behave as other (conformist) players are thought to be behaving. To this end, T0 provides a preliminary measure of conformity: there, conformity predicts a positive correlation between one's behaviour and one's guesses about the behaviour of other participants in the same role. Then, given that an important element of conformist preferences involves expectations, in order to induce subjects to state their true beliefs, T1 presents an incentivizing scheme for the elicitation of beliefs. Further, T2 is motivated by the acknowledgment that consensus and conformity may concause the aforementioned correlation; for

this reason, T2 provides the definitive test since – by transmitting information about the others’ guesses – the experimenter should be able to disentangle consensus from conformity by regressing a subject’s choice on the *stated* and *transmitted* beliefs (or their interaction). In fact, if a subject’s guess about some others’ actions is elicited before the very subject is shown the average guess made by others, then only the conformity hypothesis is consistent with that subject using (also) the transmitted information to decide on the action to take.

In summary, of particular interest are the following hypotheses:

H1 – positive correlation between behaviour and beliefs about the behaviour of other participants in the same role;

H2 – neither the rate of cooperative behaviour nor the beliefs about the others’ behaviour vary when beliefs are incentivized;

H3 – behaviour is influenced by the transmitted information about the others’ beliefs.

The account of conformity outlined before leads one to expect to find support for H1 and H3, and to have open minds about H2.⁹⁹ In particular

⁹⁹ It should be noticed that T2 does not present an incentivizing scheme for the elicitation of beliefs, in order to avoid rewarding subjects based on the manipulated, transmitted information. In effect, the design of T2 was finalized after observing the results supporting both H1 and H2, in T0 and T1. As a consequence, if neither offering (in T1) nor not-offering (in T0) to pay for beliefs changes behaviour – when subjects are *not* shown an aggregate measure of the others’ beliefs – then one can reasonably assume that offering or not-offering to pay for beliefs will not change behaviour even when subjects *are* shown an aggregate

notice that, while H1 is consistent with both conformity and consensus, H3 is consistent with conformity only.

III.3.d. Procedure

The experiment was run with *zTree* (Fischbacher [2007]) in the Experimental Economics Lab at Royal Holloway, between February and May 2012; subjects were recruited via emails forwarded across all faculties at Royal Holloway. A total of 209 subjects participated in the experiment; each session consisted of one of the three treatments (no subject could participate in more than one session). Each session took around 45 minutes and average earnings were £8 (including a £3 show-up fee), with minimum and maximum payments being £4 and £14, respectively.

A crucial element of the current conformity test involves introducing an exogenous variation in conjectures about group norms, as expressed by beliefs about other players' behaviour in treatment T2. In order to collect enough data so as to conveniently test for the key hypothesis H3, a computerized sampling method was used for selecting the guesses (made at step (ii)) to be pooled and passed on to participants in the same session, at step (iii) of each part of T2. Before describing such a sampling method, first, let $\gamma_i(c)$ and $\gamma_i(d) = 1 - \gamma_i(c)$ denote *Participant B's stated guess about the fraction of the other Participants B* (in the same session) *that will choose c and d* (labelled as "share" and "keep", in the lab), respectively; similarly, let $\gamma_i(a)$ and $\gamma_i(b) = 1 - \gamma_i(a)$ denote *Participant A's stated guess about the*

measure of the others' beliefs (besides, in Part I of T2 subjects did not know that, after stating their guesses, these would be pooled and transmitted to other participants).

fraction of the other Participants A (in the same session) that will choose a and b (labelled as “in” and “out”, in the lab), respectively. Then, let $\bar{\gamma}_i(c)$ and $\bar{\gamma}_i(a)$ denote the average guess (made by a sample of other participants in the same role, in the same session) transmitted to subject i .

Now, in order to introduce an exogenous variation in beliefs, and to ensure that – in Part I of T2 – some subjects received information about an average belief of low cooperation (*i.e.*: $\bar{\gamma}_i(c) < 0.5$) and some others received information about an average belief of high cooperation (*i.e.*: $\bar{\gamma}_i(c) > 0.5$), each subject i was in fact shown the average guess made by a specifically selected sample of participants: more precisely, the guesses were selected in a way such that all $\bar{\gamma}_i(c)$ converged to the values of 0.25 and 0.75. Similarly, in Part II of T2 the guesses were selected in a way such that all $\bar{\gamma}_i(a)$ converged to the values of 0.25 and 0.75. It should be noted that, for a given subject i , the pieces of information transmitted in Part I and II formed one of the following combinations: $\bar{\gamma}_i(c) \sim \bar{\gamma}_i(a) \sim 0.25$, or $\bar{\gamma}_i(c) \sim \bar{\gamma}_i(a) \sim 0.75$, or $\bar{\gamma}_i(c) \sim 0.25$ and $\bar{\gamma}_i(a) \sim 0.75$, or $\bar{\gamma}_i(c) \sim 0.75$ and $\bar{\gamma}_i(a) \sim 0.25$; that is, some subjects received information about an average belief of low (or high) cooperation in both Part I and Part II, whereas some subjects received information about an average belief of low cooperation in Part I and high cooperation in Part II (or *vice versa*).

Before proceeding to the commentary on the data, a few observations are in order. First, the reason why the sampling algorithm has been devised in a way to select samples of subjects (*i.e.*: guesses) such that all $\bar{\gamma}_i(\cdot)$ converge to 0.25 and 0.75 is just to obtain two distributions of transmitted beliefs per each part of a session, that is, the “low transmitted belief” and the “high transmitted belief” distributions. In this respect, it should be noticed that one could have chosen any other value; on the other hand, 0.25 and 0.75 have been preferred for the only reason that they are unique in that each is the central value of a range of beliefs about low cooperation and high

cooperation, respectively. Finally, it should be stressed that the current design does not involve deception because – as mentioned above – the message shown on screen explicitly stated that the reported information referred to a sample of other participants.

III.4. Results

III.4.a. Analysis of treatments T0-T1

The analysis of treatments without belief manipulation, that is, the treatments with non-incentivized (T0) and incentivized (T1) elicitation of beliefs shows that there is a relationship between behaviour and expectations regarding the behaviour of other participants in the same role. To begin with, Table III-1 and Table III-2 summarize the data for the decision to cooperate and the stated belief (for each part) with reference to T0 and T1, respectively. Note that the decision to cooperate is dichotomous, taking on value 1 when a subject chooses c (*i.e.*: “shares” in Part I) or a (*i.e.*: “opts in” in Part II), and taking on value 0 otherwise; also note that the stated beliefs $\gamma_i(\cdot)$ are expressed as percentages.

Variable	Obs.	Mean	Std. Dev.	m	M
<i>share</i>	53	.509434	.5046949	0	1
$\gamma_i(\textit{share})$	53	38.62264	25.97945	0	99
<i>in</i>	53	.6415094	.4841463	0	1
$\gamma_i(\textit{in})$	53	50.41509	30.91765	0	100

Table III-1 - T0 summary statistics (m and M indicate the minimum and maximum values, respectively); $\gamma_i(\cdot)$ indicate stated beliefs

Variable	Obs.	Mean	Std. Dev.	m	M
<i>share</i>	46	.5869565	.4978213	0	1
$\gamma_i(\textit{share})$	46	45.69565	25.48888	0	100
<i>in</i>	46	.6086957	.4934352	0	1
$\gamma_i(\textit{in})$	46	62.54348	25.84458	5	100

Table III-2 - T1 summary statistics (m and M indicate the minimum and maximum values, respectively); $\gamma_i(\cdot)$ indicate stated beliefs

One can perform formal tests of the null hypothesis that there is no correlation between behaviour and expectations regarding the behaviour of other participants in the same role; the alternative hypothesis is that decision and belief are not independent. By using the data for T0, one gets a positive Spearman correlation coefficient of 0.5022 ($p = 0.0001$) for Part I, and a positive Spearman correlation coefficient of 0.6028 ($p = 0.0000$) for Part II; similarly, using the data for T1, one gets a positive Spearman correlation coefficient of 0.4994 ($p = 0.0004$) for Part I, and a positive Spearman

correlation coefficient of 0.4953 ($p = 0.0005$) for Part II.¹⁰⁰ Hence, in both cases there is strong evidence against the null hypothesis. Additionally, one could compare the difference in beliefs between those who chose to cooperate and those who chose not to: the following table presents the mean values of the beliefs for both cooperators and non-cooperators in each part, along with Z Statistics for the Wilcoxon rank-sum (Mann-Whitney) tests of the null hypotheses

$$\gamma_i(\text{share}) (\text{if share} \equiv 0) = \gamma_i(\text{share}) (\text{if share} \equiv 1),$$

$$\gamma_i(\text{in}) (\text{if in} \equiv 0) = \gamma_i(\text{in}) (\text{if in} \equiv 1).$$

Treat	Obs.	$\gamma_i(\text{share})$ if share \equiv 0	$\gamma_i(\text{share})$ if share \equiv 1	Z Stat	$\gamma_i(\text{in})$ if in \equiv 0	$\gamma_i(\text{in})$ if in \equiv 1	Z Stat
T0	53	25.88462	50.88889	-3.621 (***)	26.36842	63.85294	-4.347 (***)
T1	46	30.68421	56.25926	-3.350 (***)	46	73.17857	-3.323 (***)

Table III-3 - Mean values of belief variables; the Z Statistic reflects the Wilcoxon-Mann-Whitney test for the two populations compared (in brackets are significance levels for the tests above, with *** indicating $p < 0.01$); $\gamma_i(\cdot)$ indicate stated beliefs

¹⁰⁰ By using aggregate data for T0 and T1 (no. of obs.=99), one gets a Spearman correlation coefficient of 0.5064 ($p = 0.0000$) for Part I, whereas a Spearman correlation coefficient of 0.5440 ($p = 0.0000$) for Part II.

Again, there is strong evidence against the null hypothesis in all cases, thereby highlighting a relationship between behaviour and expectations regarding the behaviour of other participants in the same role. In other words, the tests fully support H1 as stated in section III.3.c. above, that is, there is strong evidence of a positive correlation between behaviour and expectations regarding the behaviour of other participants in the same role, across both treatments T0 and T1.

Given that, I shall move on to run a probit regression for each part of the above treatments so as to start presuming a causal hypothesis (in which beliefs determine a conformist's behaviour) that will be explored fully with the analysis of T2 in the next subsection. So, for now I will simply present the coefficients of probit regressions: (i) with *share* as the dependent variable and $\gamma_i(\textit{share})$ as the predictor, for Part I of each treatment; (ii) with *in* as the dependent variable and $\gamma_i(\textit{in})$ as the predictor for Part II of each treatment.

Treat	Obs.	<i>share</i>		<i>in</i>	
		$\gamma_i(\textit{share})$	<i>constant</i>	$\gamma_i(\textit{in})$	<i>constant</i>
T0	53	.0273249*** (.008943)	-1.007842*** (.3702524)	.0290061*** (.0069194)	-.984093*** (.3742122)
T1	46	.0292078*** (.0099164)	-1.060547** (.4369656)	.0293176*** (.0094252)	-1.542668*** (.5710146)

Table III-4 - T0 and T1 Probit regression coefficients; in brackets are robust standard errors (** and *** indicate $p < 0.05$ and $p < 0.01$, respectively, for the relevant Z Statistic); $\gamma_i(\cdot)$ indicate stated beliefs

Clearly, a positive coefficient means that an increase in the explanatory variable will lead to an increase in the predicted probability of the rate of cooperative behaviour; hence, Table III-4 shows that (in Part I of each treatment) an increase in $\gamma_i(\textit{share})$ leads to an increase in the predicted probability of *share*, whereas (in Part II of each treatment) an increase in $\gamma_i(\textit{in})$ leads to an increase in the predicted probability of *in*. Again, notice that such results may admit a reverse causality interpretation (in line with a false consensus effect hypothesis); for this reason, the analysis of T2 will shed more light on our conformity motivation.

On a different note – given that T2 does not present an incentivizing scheme for the elicitation of beliefs – it is important to check whether the rate of cooperative behaviour or the beliefs about the others' behaviour vary when beliefs are incentivized. I shall consider *beliefs* first: for Part I, the null hypothesis is that the mean for $\gamma_i(\textit{share})$ is the same for T0 and T1; for Part II, the null hypothesis is that the mean for $\gamma_i(\textit{in})$ is the same for T0 and T1. Thus, as for Part I, the Wilcoxon rank-sum (Mann-Whitney) test suggests that there is not a statistically significant difference between the underlying distributions of $\gamma_i(\textit{share})$ for T0 and T1 ($z = -1.296$, $p = 0.1950$). As regards Part II, the Wilcoxon rank-sum (Mann-Whitney) test suggests that there is mild evidence against the null hypothesis ($z = -1.827$, $p = 0.0677$): in light of this result, the analysis of the *decision to cooperate* (*i.e.*: a test of whether offering to pay for beliefs affected behaviour or not) becomes critical. In brief: for Part I, the null hypothesis is that the mean for *share* is the same for T0 and T1; for Part II, the null hypothesis is that the mean for *in* is the same for T0 and T1. Thus, as for Part I, the Wilcoxon-Mann-Whitney test suggests that there is not a statistically significant difference between the underlying distributions of *share* for T0 and T1 ($z = -0.769$, $p = 0.4421$): this is not surprising, given the above analysis of beliefs, and the established correlation between beliefs and behaviour. More interestingly, as regards

Part II, the Wilcoxon-Mann-Whitney test suggests that there is no evidence against the null hypothesis ($z = 0.335$, $p = 0.7377$). Hence, one can conclude that the tests almost fully support H2 as stated in section III.3.c. above, that is, offering or not-offering to pay for beliefs is very unlikely to affect behaviour or beliefs (incentivizing beliefs may, at most, induce players to overestimate the fraction of the other participants that will choose to opt in, in Part II).

Before proceeding to the analysis of T2, it should be highlighted that subjects' beliefs were very unequally distributed across the range $[0, 0.5) \cup (0.5, 1]$, with a stated belief below 50% being most frequent in Part I of both T0 and T1 (and with a stated belief above 50% being most frequent in Part II of both T0 and T1):¹⁰¹ more specifically, in Part I of T0 about 68% of subjects stated a belief such that $\gamma_i(c) < 0.5$ whereas in Part II of T0 about 45% of subjects stated a belief such that $\gamma_i(a) < 0.5$; similarly, in Part I of T1 about 59% of subjects stated a belief such that $\gamma_i(c) < 0.5$ whereas in Part II of T1 about 33% of subjects stated a belief such that $\gamma_i(a) < 0.5$. Given that, one should interpret the above findings bearing in mind that (the slope of) any relationship between behaviour and beliefs may vary when analyzing subjects who stated a belief below 50% (*i.e.*: subjects with an expectation of predominant defection) and those who stated a belief above 50% (*i.e.*: subjects with an expectation of predominant cooperation): I shall return to this issue in the next subsection.

¹⁰¹ Recall that it was not possible for subjects to enter a stated belief of 50% (see footnote 97 above).

III.4.b. Analysis of treatment T2

I will now proceed to the analysis of the treatment with belief manipulation. Table III-5 summarizes the data for the decision to cooperate and the stated belief (for each part) with reference to T2.

Variable	Obs.	Mean	Std. Dev.	m	M
<i>share</i>	110	.5818182	.4955179	0	1
$\gamma_i(\textit{share})$	110	44.43636	24.91104	0	99
<i>in</i>	110	.6818182	.4679022	0	1
$\gamma_i(\textit{in})$	110	59.3	27.35078	0	100

Table III-5 - T2 summary statistics (m and M indicate the minimum and maximum values, respectively); $\gamma_i(\cdot)$ indicate stated beliefs

Again, one can perform formal tests of the null hypothesis that there is no correlation between behaviour and expectations regarding the behaviour of other participants in the same role: by using the data for T2, one gets a positive Spearman correlation coefficient of 0.4060 ($p = 0.0000$) for Part I, and a positive Spearman correlation coefficient of 0.5176 ($p = 0.0000$) for Part II. In light of these results, one can conclude that the tests fully support H1 as stated in section III.3.c. above, that is, there is strong evidence of a positive correlation between behaviour and expectations regarding the behaviour of other participants in the same role, across all treatments. Yet, as discussed in section III.3.a. above, such a strong correlation might be due to consensus effects (which imply an inverse causal relationship): further

tests are therefore needed to establish whether beliefs affect behaviour, a point which would support the conformity hypothesis.

It should now be recalled that the defining element of treatment T2 involves introducing an exogenous variation in beliefs: to this end, T2 was implemented with a sampling algorithm such that subjects received information about an average belief of low or high cooperation; as explained before, for a given subject i , the pieces of information transmitted in Part I and II formed one of the following combinations: $\bar{\gamma}_i(c) \sim \bar{\gamma}_i(a) \sim 0.25$, or $\bar{\gamma}_i(c) \sim \bar{\gamma}_i(a) \sim 0.75$, or $\bar{\gamma}_i(c) \sim 0.25$ and $\bar{\gamma}_i(a) \sim 0.75$, or $\bar{\gamma}_i(c) \sim 0.75$ and $\bar{\gamma}_i(a) \sim 0.25$. (More specifically, after collecting the beliefs stated by participants, the programme computed and transmitted the beliefs as summarized in Table III-8 and Table III-9: see Appendix III.) In this respect, as discussed above, it should be recalled that (in Part II of T2) some Participant A's choice might be affected by the fact that that very subject got some information in Part I. For this reason – and also because of the small difference in payoff between (b, \cdot) and (a, d) , which encourages Participants A to choose a – as in other studies (*e.g.*: Ellingsen *et al.* [2010]) the focus of this treatment will be on Trustees' rather than on Trustors' behaviour.

Now, as noted during the analysis of treatments T0-T1, also in T2 subjects' stated beliefs were very unequally distributed across the range $[0, 0.5) \cup (0.5, 1]$, with a stated belief below 50% being most frequent in Part I and with a stated belief above 50% being most frequent in Part II: more specifically, in Part I of T2 almost 60% of subjects stated a belief such that $\gamma_i(c) < 0.5$ whereas in Part II of T2 about 36% of subjects stated a belief such that $\gamma_i(a) < 0.5$. Conversely, as just mentioned, the information transmitted was a percentage in the vicinity of 0.25 or in the vicinity of 0.75, and it was *assigned randomly to each subject regardless of the belief she stated at step (ii)*: in particular, the ratio of subjects assigned a transmitted belief of low cooperation (that is, $\bar{\gamma}_i(\cdot) < 0.5$, *i.e.*: $\bar{\gamma}_i(\cdot) \sim 0.25$) to those

assigned a transmitted belief of high cooperation (that is, $\bar{\gamma}_i(\cdot) > 0.5$, *i.e.*: $\bar{\gamma}_i(\cdot) \sim 0.75$) was of about 1:1, in each part. Given that – and also in light of possible interaction effects between low/high stated beliefs and low/high transmitted beliefs – I shall now proceed to analyze first subjects who stated a belief below 50%, and then those who stated a belief above 50%: in this way one can reasonably assume that each group being treated is indeed homogeneous.

Hence, I will start by running a probit regression for each group of stated beliefs (*i.e.*: below/above 50%) using the data from treatment T2, in order to check whether a model with *stated belief* as the only explanatory variable is significant.¹⁰² That is, I will simply present the coefficients of probit regressions: (i) with *share* as the dependent variable and $\gamma_i(\textit{share})$ as the predictor, for Part I of each group of stated beliefs (and for the whole sample); (ii) with *in* as the dependent variable and $\gamma_i(\textit{in})$ as the predictor for Part II of each group of stated beliefs (and for the whole sample).

¹⁰² I did not perform a similar regression using the data from treatments T0-T1, because the number of observations of each group of stated beliefs (*i.e.*: below/above 50%) for each of those treatments was too small to produce meaningful estimations.

Stated beliefs	<i>share</i>			<i>in</i>		
	Obs. Part I	$\gamma_i(\textit{share})$	<i>constant</i>	Obs. Part II	$\gamma_i(\textit{in})$	<i>constant</i>
$0 \leq \gamma_i(\cdot) < 0.5$	65	.0298773* (.0153424)	-.9243365** (.4434949)	40	.0253571 (.0164127)	-1.024315** (.4905198)
$0.5 < \gamma_i(\cdot) \leq 1$	45	.0315747 (.0208432)	-1.417182 (1.422782)	70	.034035** (.016576)	-1.483769 (1.24247)
all	110	.0230871*** (.0056345)	-.7770586*** (.2652993)	110	.0294296*** (.0052826)	-1.142494*** (.3148897)

Table III-6 - T2 Probit regression coefficients; in brackets are robust standard errors (*, ** and *** indicate $p < 0.10$, $p < 0.05$ and $p < 0.01$, respectively, for the relevant Z Statistic); $\gamma_i(\cdot)$ indicate stated beliefs

Table III-6 shows that (in Part I of T2) an increase in $\gamma_i(\textit{share})$ leads to an increase in the predicted probability of *share*, whereas (in Part II of T2) an increase in $\gamma_i(\textit{in})$ leads to an increase in the predicted probability of *in*. However, it should be noted that – while the coefficients for the stated beliefs are strongly significant when considering the whole sample (as shown in the last row of the table) – the significance of the coefficients considerably decreases (with many cases of insignificance being present) when analyzing each group of stated beliefs in turn (*i.e.*: below/above 50%). Although that may be due to a variety of factors, it is reasonable to suspect that adding another explanatory variable may increase the significance of the model, at least for one of the groups of stated beliefs if interaction effects between low/high stated beliefs and low/high transmitted beliefs are present.

Thus, I shall turn to run a probit regression for each group of stated beliefs (*i.e.*: below/above 50%) so as to check whether a model with both *stated belief* and *transmitted belief* as explanatory variables is significant. (As explained before I shall focus on Part I, though I will return to Part II later on.) The following table presents the coefficients of probit regressions: (*i*) with *share* as the dependent variable and with $\gamma_i(\textit{share})$, $\bar{\gamma}_i(\textit{share})$ as predictors, for Part I of each group of stated beliefs (and for the whole sample).

Stated beliefs	<i>share</i>			
	Obs. Part I	$\gamma_i(\textit{share})$	$\bar{\gamma}_i(\textit{share})$	<i>constant</i>
$0 \leq \gamma_i(\textit{share}) < 0.5$	65	.0298451* (.0153815)	.0005991 (.0066462)	-.9550853* (.5712344)
$0.5 < \gamma_i(\textit{share}) \leq 1$	45	.0496499* (.0257273)	.0190708** (.0088884)	-3.494586* (1.904342)
all	110	.0240949*** (.005691)	.005425 (.0052484)	-1.092166*** (.4135265)

Table III-7 - T2 Part I Probit regression coefficients; in brackets are robust standard errors (*, ** and *** indicate $p < 0.10$, $p < 0.05$ and $p < 0.01$, respectively, for the relevant Z Statistic); $\gamma_i(\cdot)$ indicate stated beliefs, $\bar{\gamma}_i(\cdot)$ indicate transmitted beliefs

Interestingly, Table III-7 shows that – while the coefficient for the transmitted belief $\bar{\gamma}_i(\textit{share})$ is not significant when considering the whole sample (as shown in the last row of the table) – an increase in the transmitted belief has a significant, positive effect on the group of subjects who stated a belief

above 50%: hence, for that group of subjects, an increase in the transmitted belief $\bar{\gamma}_i(\text{share})$ leads to an increase in the predicted probability of *share*.

The above results would seem to support the suspicion that interaction effects between stated beliefs and transmitted beliefs are present: as a matter of fact, that is easily confirmed (using the data from the whole sample, *i.e.*: 110 observations) by regressing the decision to cooperate on: the stated belief ($\gamma_i(\text{share})$), a dummy for low/high transmitted belief ($d\bar{\gamma}_i(\text{share})$), and an interaction term between stated belief and low/high transmitted belief; not surprisingly, performing such a regression shows that the interaction is significant.¹⁰³ For Part I, the probit regression in question gives:

$$\Phi(-.5867 + .016 * \gamma_i(\text{share}) - .8305 * d\bar{\gamma}_i(\text{share}) + .0290 * \gamma_i(\text{share}) * d\bar{\gamma}_i(\text{share})),$$

(.3613)
(.0066)
(.5578)
(.0130)

with robust standard errors being shown in brackets below each coefficient. The regression shows that, while the dummy for low/high transmitted belief is not significant ($p = 0.137$), the interaction term between stated belief and low/high transmitted belief is significant ($p = 0.026$), thereby confirming that the transmitted information has an effect on Trustees' behaviour, depending on the level of their stated beliefs.¹⁰⁴ Now, it is well known that – when using

¹⁰³ By using the actual transmitted belief ($\bar{\gamma}_i(\text{share})$) in place of the dummy for low/high transmitted belief ($d\bar{\gamma}_i(\text{share})$) one obtains similar results, confirming a significant interaction between stated beliefs and transmitted beliefs.

¹⁰⁴ For Part II, running a similar regression, one obtains relatively similar results. The probit regression in this case gives:

$$\Phi(-1.8211 + .0398 * \gamma_i(\text{in}) + 1.2573 * d\bar{\gamma}_i(\text{in}) - .0193 * \gamma_i(\text{in}) * d\bar{\gamma}_i(\text{in})),$$

(.5145)
(.0083)
(.6725)
(.011)

a non-linear model – one cannot infer that the probability (of cooperating) increases or decreases from the sign of the coefficients, if the model has an interaction term. Hence, to give an idea of the extent of the impact of the transmitted beliefs, I shall compute the predicted probability that Participant B cooperates (*i.e.*: $share = 1$), when the stated belief is held at its mean value (*i.e.*: 44.44) and the transmitted belief is either low or high (*i.e.*: ~ 0.25 or ~ 0.75 , respectively): in brief, the predicted probability of cooperating is 0.55 for those who get a low transmitted belief and 0.72 for those who get a high transmitted belief. (Notice that the impact of the transmitted beliefs is mitigated by the fact that subjects who stated a low belief are most frequent in the sample.)

In light of these findings, one can conclude that the tests provide support for H3 as stated in section III.3.c. above or, more precisely, the tests show that the transmitted information about the others' belief influence one's behaviour, depending on one's stated beliefs. It should now be recalled that, in the case of conformity, the causal relationship runs from beliefs to behaviour – whereas in the case of consensus one's action influences one's

with robust standard errors being shown in brackets below each coefficient. Unlike Part I, here the regression shows that *both* the dummy for low/high transmitted belief *and* the interaction term are mild significant (with p being 0.062 and 0.079, respectively), thereby confirming that the transmitted information may have an effect also on Trustors' behaviour, depending on the level of their stated beliefs. Now, as conjectured before, this difference in results (between Part I and Part II) could be due to interaction effects between Part I and Part II of T2: in effect, in Part II of T2 some Participant A's choice might be affected by the fact that that very subject got some information in Part I; running the above regression (for Part II) with the addition of a dummy for low/high transmitted belief in Part I (*i.e.*: $d\bar{y}_i(share)$) confirms that $d\bar{y}_i(share)$ is indeed significant in Part II.

beliefs about some others' actions – it is therefore clear that only conformity is consistent with a subject using (also) the transmitted information to decide on the action to take; hence, there is evidence of conformity being present in our data.

III.5. Concluding remarks

This essay has presented a test for conformist motivations in mixed-motive games. The data show that the conformity motivation is indeed present: in T2 an effect of (exogenously varying) beliefs on behaviour suggests that conformity is at least part of what has driven the correlation between beliefs and behaviour in T0-T1 (and, possibly, in previous Trust Game experiments). Interestingly, the results show that an increase in the transmitted belief has a significant, positive effect on the group of subjects who stated a higher belief (*i.e.*: subjects with an expectation of predominant cooperation), but not on those who stated a lower belief (*i.e.*: subjects with an expectation of predominant defection).

Now, it should be recalled that in the theory of social norms that I have previously formalized, a conformist player in a social dilemma would do her part of what is presumed to be the norm-complying strategy profile, provided that a certain set of conditions holds, where such conditions involve *both* empirical *and* normative expectations of conformity to the norm. Just for argument's sake, here is a very simplified version of such conditions (see also Bicchieri [2006], Ch. 1): (*i*) subjects must be aware of the existence of a norm; (*ii*) they must believe that the others will conform to the norm ("empirical expectations condition"); (*iii*) they must believe that the others expect them to conform to the norm ("normative expectations condition"). Given that, it should be noted that transmitting information regarding the

others' beliefs about some others' behaviour – to a potentially conformist player – constitutes an exogenous variation in empirical expectations of conformity to a norm of cooperation; hence, it should be stressed once again that in this experimental study I have focused on the empirical expectation side (*i.e.*: a player's belief about the others conforming to a given rule of behaviour), while disregarding the normative expectation side (*i.e.*: a player's belief about the others expecting her to conform to a given rule of behaviour).

Then, given that the transmitted information had a significant impact only on players who stated a higher belief, the data seem to suggest that only those players were conformist. What does it mean? That may suggest that those who stated a higher belief are exactly those who had a normative expectation such that they believed that the others would expect them to cooperate. Hence, while an exogenous variation in empirical expectations of conformity to a norm (*i.e.*: the transmitted information) may have an effect on the group of subjects for whom the normative expectations condition is fulfilled, it may not have an effect on the group of subjects for whom the normative expectations condition is not fulfilled (where the latter group may possibly comprise those subjects who stated a lower belief). To conclude, future research should delve into the empirical/normative expectation distinction which is crucial to conformity to social norms.

III.6. Appendix III

III.6.a. Additional data

\bar{y}_i (share) [Part I]	<i>high belief</i>		Total
	no	yes	
22	5	0	5
23	3	0	3
24	7	0	7
25	21	0	21
26	12	0	12
27	5	0	5
72	0	1	1
73	0	8	8
74	0	26	26
75	0	19	19
76	0	1	1
78	0	2	2
Total	53	57	110

Table III-8 - Beliefs transmitted in Part I of T2 (the last column shows the number of subjects being shown each belief)

$\bar{y}_i(in)$ [Part II]	<i>high belief</i>		Total
	no	yes	
0	1	0	1 ¹⁰⁵
20	2	0	2
21	3	0	3
23	7	0	7
24	10	0	10
25	13	0	13
26	7	0	7
27	6	0	6
28	2	0	2
29	5	0	5
39	1	0	1
72	0	1	1
73	0	5	5
74	0	12	12
75	0	24	24
76	0	7	7
77	0	4	4
Total	57	53	110

Table III-9 - Beliefs transmitted in Part II of T2 (the last column shows the number of subjects being shown each belief)

¹⁰⁵ One subject received a belief of 0, because no other belief lower than 50 had been stated by participants in that session.

III.6.b. Experimental instructions and screenshots

Paper instructions and transcripts of zTree screenshots are shown below.

General instructions for participants

Thank you for participating in this study.

Please note that it is prohibited to communicate with other participants during the experiment. If you have a question once the experiment has begun, please raise your hand and an assistant will come to your desk to answer it. Violation of this rule leads to immediate exclusion from the study and from all payments.

You will never learn the identity of the other participants, neither before nor after the study; and not one of the other participants will learn anything about your identity. Also, no other participant will learn what you earn during the experiment: upon completion of the session, the amount of money you will have earned will be paid out individually and privately. Hence, no other participant will know your choices and how much money you earn in this experiment.

You will receive £3 for participating in this session; additionally you also receive money depending on the decisions made (as described in the next paragraphs).

The experiment consists of two parts ("Part I" and "Part II"), each involving one simple decision task; your payment at the end of the session will be calculated as follows.

Your payment

= £3 (show-up fee) + any amount earned in Part I + any amount earned in Part II

In what follows we describe the procedure for Part I.

Part I

There are two types of participants, participants "A" and participants "B".

You will be assigned a type and paired with **one other participant** who was assigned another type than you.

This part consists of two steps, which you will perform with the particular participant you are paired with.

Step 1: Participant A must choose between the following two options. The first option ("OUT") gives a payout of £1 to both participants. The second option ("IN") is to instead transfer both pounds (i.e. £2 in total) to participant B and leave further decisions to him/her. If participant A transfers the 2 pounds to participant B, they will be tripled and participant B will receive $3 \times 2 = 6$ pounds.

Step 2: Only if participant A chooses the second option ("IN"), participant B will then decide if he/she transfers £3 back to participant A and keeps £3 for himself/herself OR if participant B keeps all the £6 for himself/herself.

Procedure for the two steps

Step one: Decision of participant A

It is up to participant A to choose one of the 2 options (OUT or IN): EITHER both participants receive £1 each OR the money and further decisions are transferred to participant B.

If participant A chooses the option OUT, both of you will receive £1. In this case participant B cannot change the payout allocation and the first part ends.

As a result

At the end of step one, there are two possible situations.

- If participant A has transferred the £2 to participant B (option IN), participant B has £6 and participant A has nothing.
- If participant A has chosen the option OUT, both of you have £1.

Step two: Decision of participant B

If participant A has transferred the money to participant B (option IN), then B receives £6 and it is now up to participant B to decide about the distribution of the £6 between the two participants. Participant B can EITHER:

- transfer £3 back to participant A and keep £3 for himself/herself

OR

- keep all the £6 for himself/herself and leave nothing to participant A.

After participant B's decision this part is completed and the earnings for both participants will be determined according to B's decision.

The above information is summarised in the following table:

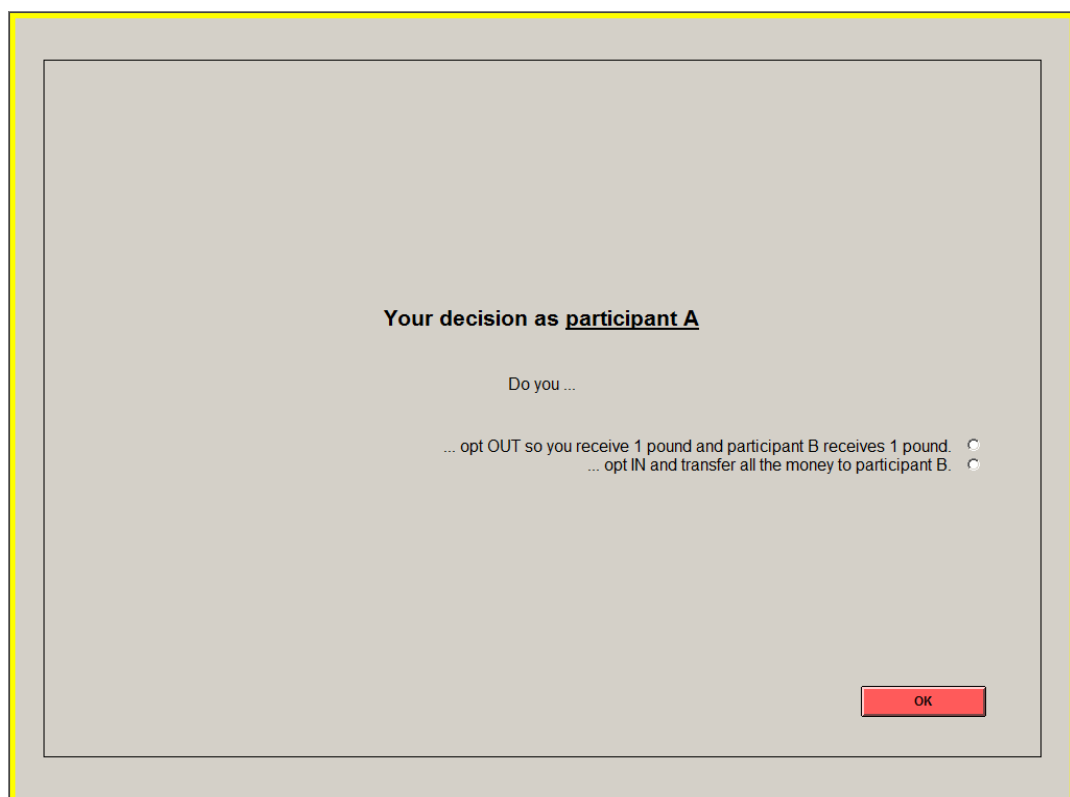
		A's income	B's income
A chose OUT		£1	£1
A chose IN	B keeps all	£0	£6
	B transfers half	£3	£3

Specific procedure and on-screen instructions for Part I

You are assigned the role of participant B

Note that you will complete the above-described two steps only once.

Step 1: Participant A decides by entering his/her choice on the screen shown below.



The screenshot shows a decision screen for Participant A. The screen has a light gray background and is enclosed in a yellow border. The text on the screen reads:

Your decision as participant A

Do you ...

... opt OUT so you receive 1 pound and participant B receives 1 pound.

... opt IN and transfer all the money to participant B.

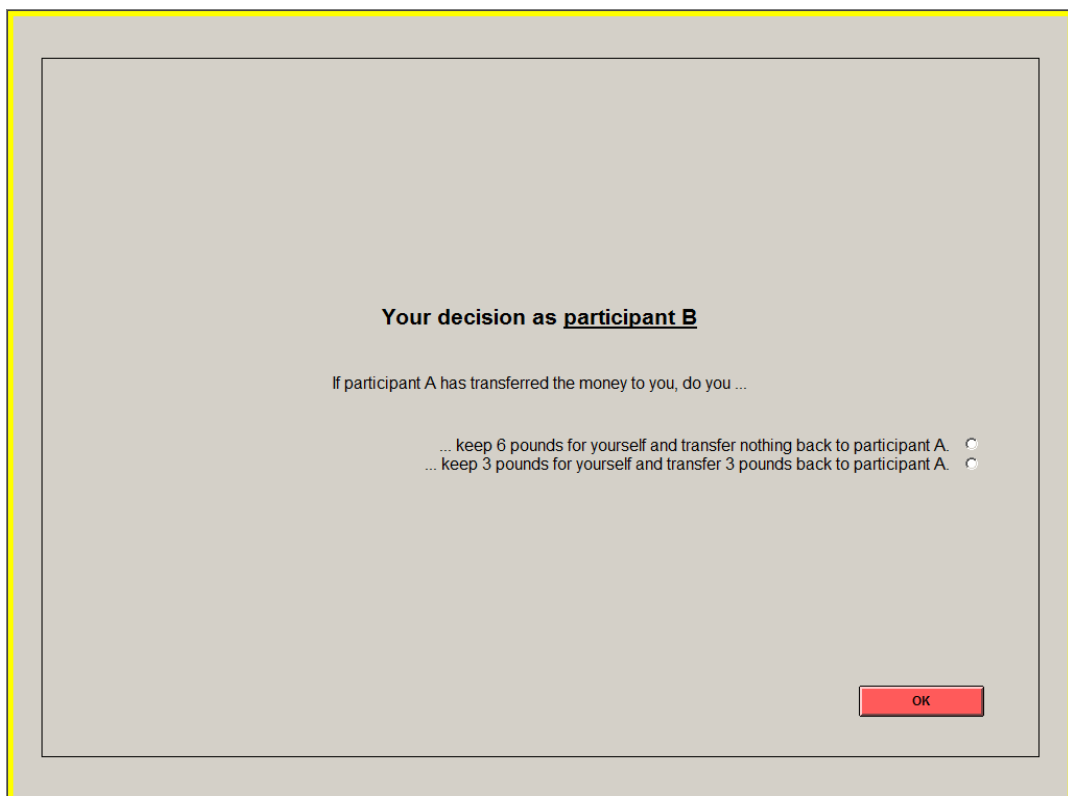
At the bottom right of the screen, there is a red button labeled "OK".

Step 2: We will ask you (participant B) how you would like to divide the £6 between participant A and yourself. Note that your answer will have an effect only if participant A does choose to transfer the money to you (option IN).

Participant A will not know your decision when he/she submits his/her own decision.

As explained above, **you decide on whether to transfer half the money to participant A or keep all the £6 for yourself.**

You will enter your choice on the following screen:



The screenshot shows a grey rectangular window with a yellow border. Inside the window, the text reads: "Your decision as participant B" followed by "If participant A has transferred the money to you, do you ...". Below this, there are two radio button options: "... keep 6 pounds for yourself and transfer nothing back to participant A." and "... keep 3 pounds for yourself and transfer 3 pounds back to participant A.". A red "OK" button is located in the bottom right corner of the window.

Control questions

Please answer the following control questions. Please contact the study organizer if you have any questions.

1. Participant A has chosen IN. You then choose to transfer half the money back to participant A.

What is the income of participant A?

What is the income of participant B (yourself)?.....

2. Participant A has chosen IN. You then choose to keep all the money for yourself.

What is the income of participant A?

What is the income of participant B (yourself)?.....

3. Participant A has chosen OUT.

What is the income of participant A?

What is the income of participant B (yourself)?.....

Please feel free to ask questions at any point if you feel you need some clarification. Please do so by raising your hand.

We will start with Part I once the instructions are clear to everyone. Are there any questions?

Part II

We are now ready to undergo the last part of the study. This part has exactly the **same two-step procedure as in Part I.**

The payouts are the same as before and are summarised in the following table:

		A's income	B's income
A chose OUT		£1	£1
A chose IN	B keeps all	£0	£6
	B transfers half	£3	£3

The only difference is that you are assigned a different type in this part than in the previous part.

You are now assigned the role of participant A.

Again, you will be paired with one other participant. **This other participant will be a different person than the one you were paired with in Part I.**

Please refer to your paper handout or ask an assistant if you need reminding of the procedure.

[Transcript of on-screen messages]

Treatments T0-T1

Screen 1 (Part I)

You are assigned the role of participant B

Prior to entering your decision as participant B, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants B).

In other words, we ask you to guess how many of today's participants B (excluding yourself) will choose to transfer half the money back, and how many of today's participants B will keep all the money for themselves.

Please enter your guess by positioning the below slider to the desired percentage.

[The below line is only for treatment T1.]

Note: You can earn some additional income if your guess is correct. If your guess differs by no more than 5 percentage points from the realized value, at the end of the study you will receive an additional payment of £2. Otherwise, you do not receive an additional income.

Screen 2 (Part I)

Enter 2nd mover decision.

Screen 3 (Part II)

Insert instructions for Part II here.

Screen 4 (Part II)

You are assigned the role of participant A

Prior to entering your decision as participant A, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants A).

In other words, we ask you to guess how many of today's participants A (excluding yourself) will choose IN, and how many of today's participants A will choose OUT.

Please enter your guess by positioning the below slider to the desired percentage.

[The below line is only for treatment T1.]

Note: You can earn some additional income if your guess is correct. If your guess differs by no more than 5 percentage points from the realized value, at the end of the study you will receive an additional payment of £2. Otherwise, you do not receive an additional income.

Screen 5 (Part II)

Enter 1st mover decision.

Screen 6

Outcome.

Treatment T2

Screen 1 (Part I)

You are assigned the role of participant B

Prior to entering your decision as participant B, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants B).

In other words, we ask you to guess how many of today's participants B (excluding yourself) will choose to transfer half the money back, and how many of today's participants B will keep all the money for themselves.

*****first lower part of screen 1*****

Please enter your guess by positioning the below slider to the desired percentage.

*****second lower part of screen 1 [to appear after subjects have entered their guesses]*****

A sample of other participants B in this session expects on average that $\langle x \rangle\%$ will transfer half the money, whereas $\langle 100-x \rangle\%$ will keep all the money.

TRANSFER HALF: $x\%$

KEEP: $(100-x)\%$

Screen 2 (Part I)

Enter 2nd mover decision.

Screen 3 (Part II)

Insert instructions for part II here.

Screen 4 (Part II)

You are assigned the role of participant A

Prior to entering your decision as participant A, we would like to know what you think of the other participants who have been assigned the same type as you (i.e. participants A).

In other words, we ask you to guess how many of today's participants A (excluding yourself) will choose IN, and how many of today's participants A will choose OUT.

*****first lower part of screen 4*****

Please enter your guess by positioning the below slider to the desired percentage.

*****second lower part of screen 4[to appear after subjects have entered their guesses]*****

A sample of other participants A in this session expects on average that $<x>\%$ will OPT IN, whereas $<100-x>\%$ will OPT OUT.

IN: $x\%$

OUT: $(100-x)\%$

Screen 5 (Part II)

Enter 1st mover decision.

Screen 6

Outcome.

Conclusions

The present thesis contributes to our understanding of some of the informal norms regulating human behaviour, namely conventions and social norms. It should be stressed that, although the present definitions of conventions and social norms differ from one another, in both cases they imply belief-based solutions to problems of strategic interdependence. Given the role played by such norms in concerting expectations, it is evident that here – when a convention or a social norm is in operation – it is the case that players are “reasoning together”. Now, unlike other models of conventions or social norms, the theories presented here aim at capturing the way people reason together by means of *neutral* frameworks, which can account for a wide range of belief-based solutions: more explicitly, in the case of coordination games, a *salience relation* is devised in such a way that need not reflect payoff or risk dominance, but may well be based on aesthetics, precedent, geometry, and so on; similarly, in the case of mixed-motive games, a *social norm* need not necessarily be egalitarian or Pareto-efficient (in effect, all is needed for a social norm to be followed is that players have conditionally conformist preferences, hold correct beliefs, and are sensitive enough to the social cost of deviating). Indeed, only by making use of a neutral framework the analyst can allow for these equilibrium selection devices to adequately reflect the mechanisms of a very wide range of economically-relevant rule-based phenomena, such as fashions, bargains and contracts.

This thesis has further developed the idea that informal norms vary because of changes in objective circumstances as well as because of subjective changes in perceptions or expectations. In this connection, the experiment has provided evidence of conformist motivations being present in

mixed-motive games, showing that exogenously varying expectations has an impact on behaviour. In this respect – given that here perceptions and expectations play such a crucial role – it is important to stress that in order for a game-theoretic account of norms to provide meaningful insights, one necessarily needs to combine deduction with empirical observation. In effect, only thanks to the empirical approach one may be able to observe and formalize which strategic principles players are likely to use: as observed by Camerer [2003], «[i]t is unlikely that a purely mathematical theory of rational play will ever fully identify which of many equilibria are likely to emerge because history, shared background, and the way strategies are described or made psychologically prominent surely matter. As a result, experiments and observation of the sort that naturalists do in biology can potentially do what mathematical analysis cannot – predict what will happen» (Camerer [2003], p. 337).

References

- Akerlof, George A.** 1980. "A Theory of Social Custom, of Which Unemployment May be One Consequence" *The Quarterly Journal of Economics*, 94(4): 749-775.
- Alós-Ferrer, Carlos and Christoph Kuzmics.** 2012. "Hidden Symmetries and Focal Points", mimeo.
- Anderson, Lisa R. and Charles A. Holt.** 1997. "Information Cascades in the Laboratory" *The American Economic Review*, 87(5): 847-862.
- Arrow, Kenneth J.** 1972. "Models of Job Discrimination" in *Racial Discrimination in Economic Life*, ed. Anthony H. Pascal. Lexington, Mass.: Heath.
- Asch, Solomon E.** 1955. "Opinions and Social Pressure" *Scientific American*, 193(5): 1-7.
- 1956. "Studies of Independence and Conformity: A Minority of One Against a Unanimous Majority" *Psychological Monographs*, 70(9): 1-70.
- Aumann, Robert J.** 1974. "Subjectivity and Correlation in Randomized Strategies" *Journal of Mathematical Economics*, 1(1): 67-96.
- 1976. "Agreeing to Disagree" *Annals of Statistics*, 4(6): 1236-1239.
- 1987. "Correlated Equilibrium as an Expression of Bayesian Rationality" *Econometrica*, 55(1): 1-18.
- Bacharach, Michael.** 1993. "Variable Universe Games" in *Frontiers of Game Theory*, ed. K. Binmore *et al.*. Cambridge, Mass.: MIT Press.
- 1999. "Interactive Team Reasoning: A Contribution to the Theory of Cooperation" *Research in Economics*, 53(2): 117-147.

- Bacharach, Michael and Michele Bernasconi.** 1997. "The Variable Frame Theory of Focal Points: An Experimental Study" *Games and Economic Behavior*, 19(1): 1-45.
- Bacharach, Michael, Gerardo Guerra and Daniel J. Zizzo.** 2007. "The Self-Fulfilling Property of Trust: An Experimental Study" *Theory and Decision*, 63(4): 349-388.
- Bacharach, Michael and Dale O. Stahl.** 2000. "Variable-Frame Level- n Theory" *Games and Economic Behavior*, 32(2): 220-246.
- Banerjee, Abhijit V.** 1992. "A Simple Model of Herd Behavior" *The Quarterly Journal of Economics*, 107(3): 797-817.
- Battigalli, Pierpaolo and Martin Dufwenberg.** 2007. "Guilt in Games" *The American Economic Review*, 97(2): 170-176.
- , 2009. "Dynamic Psychological Games" *Journal of Economic Theory*, 144(1): 1-35.
- Berkowitz, Alan D. and H. Wesley Perkins.** 1986. "Problem Drinking Among College Students: A Review of Recent Research" *Journal of American College Health*, 35: 21-28.
- Bernheim, B. Douglas.** 1994. "A Theory of Conformity" *The Journal of Political Economy*, 102(5): 841-877.
- Bicchieri, Cristina.** 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
- Bicchieri, Cristina and Erte Xiao.** 2009. "Do the Right Thing: But Only if Others Do So" *Journal of Behavioral Decision Making*, 22(2): 191-208.
- Bikhchandani, Sushil, David Hirshleifer and Ivo Welch.** 1992. "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades" *Journal of Political Economy*, 100(5): 992-1026.

- Binmore, Ken.** 2007. *Playing for Real: A Text on Game Theory*. Oxford: Oxford University Press.
- Bolton, Gary E. and Axel Ockenfels.** 2000. "ERC: A Theory of Equity, Reciprocity, and Competition" *The American Economic Review*, 90(1): 166-193.
- Camerer, Colin F.** 2003. *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton, N.J.: Princeton University Press.
- Casajus, André.** 2000. "Focal Points in Framed Strategic Forms" *Games and Economic Behavior*, 32(2): 263-291.
- Charness, Gary and Martin Dufwenberg.** 2006. "Promises and Partnership" *Econometrica*, 74(6): 1579-1601.
- Charness, Gary and Matthew Rabin.** 2002. "Understanding Social Preferences with Simple Tests" *The Quarterly Journal of Economics*, 117(3): 817-869.
- Cialdini, Robert B. and Noah J. Goldstein.** 2004. "Social Influence: Compliance and Conformity" *Annual Review of Psychology*, 55(1): 591-621.
- Collard, David A.** 1983. "Economics of Philanthropy: A Comment" *The Economic Journal*, 93(371): 637-638.
- Colman, Andrew M.** 1995. *Game Theory and its Applications in the Social and Biological Sciences*. Oxford: Butterworth-Heinemann.
- Cooper, Russell, Douglas V. DeJong, Robert Forsythe and Thomas W. Ross.** 1996. "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games" *Games and Economic Behavior*, 12(2): 187-218.
- Crawford, Vincent P. and Hans Haller.** 1990. "Learning How to Cooperate: Optimal Play in Repeated Coordination Games" *Econometrica*, 58(3): 571-595.

- Cubitt, Robin and Robert Sugden.** 2003. "Common Knowledge, Salience and Convention: A Reconstruction of David Lewis's Game Theory" *Economics and Philosophy*, 19(2): 175-210.
- Davis, Douglas D. and Charles A. Holt.** 1993. *Experimental Economics*. Princeton, N.J.: Princeton University Press.
- Dawes, Robyn M..** 1990. "The Potential Nonfalsity of the False Consensus Effect" in *Insights in Decision Making: A tribute to Hillel J. Einhorn*, ed. Robin M. Hogarth, Chicago, IL: University of Chicago Press.
- Dekel, Eddie, Barton L. Lipman and Aldo Rustichini.** 1998. "Standard State-Space Models Preclude Unawareness" *Econometrica*, 66(1): 159-173.
- Dufwenberg, Martin and Uri Gneezy.** 2000. "Measuring Beliefs in an Experimental Lost Wallet Game" *Games and Economic Behavior*, 30(2): 163-182.
- Dufwenberg, Martin and Georg Kirchsteiger.** 2004. "A Theory of Sequential Reciprocity" *Games and Economic Behavior*, 47(2): 268-298.
- Ellingsen, Tore, Magnus Johannesson, Sigve Tjøtta and Gaute Torsvik.** 2010. "Testing Guilt Aversion" *Games and Economic Behavior*, 68(1): 95-107.
- Elster, Jon.** 1989. *The Cement of Society*. Cambridge: Cambridge University Press.
- Engelmann, Dirk and Martin Strobel.** 2000. "The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given" *Experimental Economics*, 3(3): 241-260.
- Falk, Armin and Urs Fischbacher.** 2006. "A Theory of Reciprocity" *Games and Economic Behavior*, 54(2): 293-315.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation" *The Quarterly Journal of Economics*, 114(3): 817-868.

-----, 2006. "The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories" in *Handbook of the Economics of Giving, Altruism and Reciprocity: Vol. 1*, ed. Serge-Christophe Kolm and Jean Mercier Ythier. Amsterdam: North-Holland/Elsevier.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments" *Experimental Economics*, 10(2): 171-178.

Fudenberg, Drew and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information" *Econometrica*, 54(3): 533-544.

Gauthier, David. 1975. "Coordination" *Dialogue*, 14(2): 195-221.

Geanakoplos, John. 1989. "Game Theory without Partitions, and Applications to Speculation and Consensus" *Cowles Foundation Discussion Paper no. 914*.

Geanakoplos, John, David Pearce and Ennio Stacchetti. 1989. "Psychological Games and Sequential Rationality" *Games and Economic Behavior*, 1(1): 60-79.

Goeree, Jacob K. and Charles A. Holt. 2001. "Ten Little Treasures of Game Theory and Ten Intuitive Contradictions" *The American Economic Review*, 91(5): 1402-1422.

Gold, Natalie and Robert Sugden. 2007. "Collective Intentions and Team Agency" *The Journal of Philosophy*, 104(3): 109-137.

-----, 2008. "Theories of Team Agency" in *Rationality and Commitment*, ed. Fabienne Peter and Hans B. Schmid. Oxford: Oxford University Press.

- Guerra, Gerardo and Daniel J. Zizzo.** 2004. "Trust Responsiveness and Beliefs" *Journal of Economic Behavior & Organization*, 55(1): 25-30.
- Harsanyi, John C.** 1980. "Rule Utilitarianism, Rights, Obligations and the Theory of Rational Behavior" *Theory and Decision*, 12(2): 115-133.
- Harsanyi, John C. and Reinhard Selten.** 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press.
- Hawking, Stephen W. and Leonard Mlodinow.** 2010. *The Grand Design*. London: Bantam.
- Hechter, Michael and Karl-Dieter Opp.** 2001. *Social Norms*. New York, N.Y.: Russell Sage Foundation.
- Heifetz, Aviad, Martin Meier and Burkhard C. Schipper.** 2006. "Interactive Unawareness" *Journal of Economic Theory*, 130(1): 78-94.
- , 2008. "A Canonical Model for Interactive Unawareness" *Games and Economic Behavior*, 62(1): 304-324.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis and Richard McElreath.** 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies" *The American Economic Review*, 91(2): 73-78.
- Hintikka, Jaakko.** 1962. *Knowledge and Belief*. Ithaca, N.Y.: Cornell University Press.
- Hume, David.** 1740. *A Treatise of Human Nature*, ed. Lewis Amherst Selby-Bigge (2nd edition, 1978). Oxford: Clarendon Press.
- Janssen, Maarten C. W.** 2001. "Rationalizing Focal Points" *Theory and Decision*, 50(2): 119-148.

- Klucharev, Vasily, Kaisa Hytönen, Mark Rijpkema, Ale Smidts and Guillén Fernández.** 2009. "Reinforcement Learning Signal Predicts Social Conformity" *Neuron*, 61(1): 140-151.
- Kreps, David M., Paul Milgrom, John Roberts and Robert Wilson.** 1982. "Rational Cooperation in the Finitely-Repeated Prisoners' Dilemma" *Journal of Economic Theory*, 27(2): 245-252.
- Kreps, David M. and Robert Wilson.** 1982. "Sequential Equilibria" *Econometrica*, 50(4): 863-894.
- Laffont, Jean-Jacques.** 1975. "Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics" *Economica*, 42(168): 430-37.
- Lave, Lester B.** 1962. "An Empirical Approach to the Prisoners' Dilemma Game" *The Quarterly Journal of Economics*, 76(3): 424-436.
- Ledyard, John.** 1995. "Public Goods Experiments" in *The Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth. Princeton, N.J.: Princeton University Press.
- Lewis, David K.** 1969. *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.
- Li, Jing.** 2008. "The Power of Conventions: A Theory of Social Preferences" *Journal of Economic Behavior & Organization*, 65(3-4): 489-505.
- Locke, John.** 1689. *Two Treatises of Government*, ed. Peter Laslett (1988). Cambridge: Cambridge University Press.
- López-Pérez, Raúl.** 2008. "Aversion to Norm-Breaking: A Model" *Games and Economic Behavior*, 64(1): 237-267.
- Mehta, Judith, Chris Starmer and Robert Sugden.** 1992. "An Experimental Investigation of Focal Points in Coordination and Bargaining" in *Decision*

Making Under Risk and Uncertainty. New Models and Empirical Findings, ed. John Geweke. Dordrecht: Kluwer Academic Publishers.

Montague, P. Read and Terry Lohrenz. 2007. "To Detect and Correct: Norm Violations and Their Enforcement" *Neuron*, 56(1): 14-18.

NIAAA National Advisory Council on Alcohol Abuse and Alcoholism. 2002. "How to Reduce High-Risk College Drinking: Use Proven Strategies, Fill Research Gaps" *NIAAA College Materials*. National Institutes of Health: U. S. Department of Health and Human Services.
Available: <http://www.collegedrinkingprevention.gov/media/FINALPanel2.pdf>
(Accessed: 2011, December 03).

Osborne, Martin J. and Ariel Rubinstein. 1994. *A Course in Game Theory*. Cambridge, Mass.: MIT Press.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics" *The American Economic Review*, 83(5): 1281-1302.

Rapoport, Anatol and Albert M. Chammah. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, Mich.: University of Michigan Press.

Ross, Lee, David Greene and Pamela House. 1977. "The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Processes" *Journal of Experimental Social Psychology*, 13(3): 279-301.

Sacco, Pier Luigi, Paolo Vanin and Stefano Zamagni. 2006. "The Economics of Human Relationships" in *Handbook of the Economics of Giving, Altruism and Reciprocity: Vol. 1*, ed. Serge-Christophe Kolm and Jean Mercier Ythier. Amsterdam: North-Holland/Elsevier.

Sally, David. 1995. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992" *Rationality and Society*, 7(1): 58-92.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Sillari, Giacomo. 2005. "A Logical Framework for Convention" *Synthese*, 147(2): 379-400.

Sugden, Robert. 1984. "The Supply of Public Goods through Voluntary Contributions" *The Economic Journal*, 94(376): 772-787.

-----, 1995. "A Theory of Focal Points" *The Economic Journal*, 105(430): 533-550.

-----, 2000. "The Motivating Power of Expectations" in *Rationality, Rule and Structure*, ed. Julian Nida-Rümelin and Wolfgang Spohn. Amsterdam: Kluwer.

-----, 2003. "The Logic of Team Reasoning" *Philosophical Explorations*, 6(3): 165-181.

Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.

Vanderschraaf, Peter. 1998. "Knowledge, Equilibrium and Convention" *Erkenntnis*, 49(3): 337-369.