Actions, beliefs and feelings: an experimental study on psychological game theory

Giuseppe Attanasi, Rosemary Nagel

A Survey of Psychological Games: Theoretical Findings and Experimental Evidence*

Giuseppe Attanasi† and Rosemarie Nagel‡ §

1. Introduction: the role of psychological games

Most economic models assume that agents maximize their expected material payoff. However, subjects in the lab exhibit persistent and significant deviations from this self-interested maximizing behavior. A reasonable explanation for this behavior is that players can be motivated not only by *material* (monetary) payoffs but also by what are sometimes referred to as '*psychological*' utilities. These are related to preferences that are in some degree 'other regarding': they take others into account. Traditional game theory does not provide enough tools to adequately describe many of these preferences: the traditional approach assumes that utilities only depend on the actions that are chosen by the players. By contrast, when players are emotional or motivated by reciprocity or social respect, their utilities may *also directly* depend on the beliefs (about choices,

*Published in "*Games, Rationality and Behaviour. Essays on Behavioural Game Theory and Experiments*", A. Innocenti and P. Sbriglia (eds.), Palgrave McMillan, Houndmills, February 2008 (Ch. 9, pp. 204-232).

†Toulouse School of Economics (Georges Meyer Chair in Mathematical Economics, LERNA) and Bocconi University (Department of Economics).

‡Universitat Pompeu Fabra (Department of Economics).

§We are grateful to Pierpaolo Battigalli for the useful suggestions encouraging this survey and to Martin Dufwenberg for the helpful starting points we found in his doctoral dissertation. We are also grateful to Nikolaos Georgantzis and the participants in several seminars for helpful discussions and to Michel Serafinelli for useful research assistance. Off course the authors are responsible for any error in the chapter. Giuseppe Attanasi acknowledges financial support from Bocconi University and thanks Universitat Jaume I of Castelló for its hospitality during part of this project. Rosemarie Nagel acknowledges financial support from the Spanish Ministry of Education and Science under grant SEC2002-03403, and thanks the Barcelona Economics Program of CREA for support. Both authors thank for the hospitality of HSS in Caltech where part of the chapter was written.

beliefs, or information) they hold. This is not to say that traditional game theory is not able to analyze the influence of feelings, emotions and social norms on the players' behavior. Distribution-dependent preferences *à la* Fehr and Schmidt (1999), for example, can be addressed by the traditional game theory. But when we deal with *intention-based* feelings, emotions and social norms, i.e. *belief-dependent* motivations, we need to turn to psychological game theory. This new framework focuses on strategic settings where at least one player has belief-dependent motivations or believes, with a certain probability, that one of his opponents has belief-dependent motivations. Nonetheless, it allows for every other kind of social preferences. In that sense, it can be interpreted as a generalization of the traditional game theory.

In games with belief-dependent motivations there are clearly two channels through which beliefs and information affect behavior: the direct (psychological) impact of beliefs on preferences over terminal histories, and the (traditional) impact of (updated) beliefs about the opponents on the preferences over own strategies. Geanakoplos, Pearce and Stacchetti (1989) is the seminal paper that shows the inadequacy of traditional methods in representing the involved preferences, and develops extensions of the traditional game theory in order to deal with the matter. Battigalli and Dufwenberg (2005) generalize and extend Geanakoplos, Pearce and Stacchetti (1989), thus providing the framework we use to analyze games with belief-dependent motivations both from a theoretical and from an experimental point of view.

Experimental evidence gives support to the theories of belief-dependent motivations. Even when not explicitly designed to test such motivations, several experimental works provide results that are in line with psychological game theoretical predictions. Some of these experiments can be seen as an indirect proof of the relevance of psychological games in explaining certain forms of strategic interaction. Some others have been explicitly designed to test the relevance of psychological game theory.

The plan of the chapter is as follows. In section 2, we provide the main theoretical insights of the general psychological game framework, through the example of a simple trust game in which belief-dependent motivations are involved. In section 3, we describe and discuss the main experimental papers that directly or indirectly refer to psychological game-theoretic explanations as being able to rationalize their experimental results. Throughout the chapter, we try to clarify ideas and to dispel misconceptions emerged among economists about this new field, in order to stress the role and the importance of psychological games both for the strategic interaction analysis and for the related experimental applications.

2. Main Theoretical Findings

2.1 Theoretical Literature

Geanakoplos, Pearce and Stacchetti (1989; henceforth GPS) introduce belief-dependent motivation into strategic decision making. They develop a new analytical framework centred on the concept of *psychological game* (or *game with belief-dependent motivations*), i.e. a strategic interactive situation in which players' utilities do not only depend on terminal nodes but also on the beliefs (about choices, beliefs, or information) they hold.[1] GPS also present several examples that illustrate the inadequacy of traditional methods in representing preferences that reflect various forms of belief-dependent motivation.

GPS framework may be seen as a generalization of a traditional game able to model only some of the specific belief-dependent motivations in strategic settings: as stated by Battigalli and Dufwenberg (2005; henceforth BD), 'GPS's toolbox of psychological games incorporates several restrictions that rule out many plausible forms of belief-dependent motivation' (p. 41). In particular, GPS only allow initial beliefs to enter the domain of a player's utility; however, many seemingly important forms of belief-dependent motivations require the introduction of *updated beliefs*.

BD generalize and extend GPS, by allowing updated higher-order beliefs, beliefs of others, planned strategies, and incomplete information to influence motivation. Among other advances, BD address the issue of how beliefs about others' beliefs are revised as the play unfolds: they are able to model *dynamic psychological effects* that are ruled out when epistemic types are identified with hierarchies of initial beliefs. They also define a notion of *Psychological Sequential Equilibrium*, which generalizes the sequential equilibrium notion for traditional games, for which they prove existence under mild assumptions. In the next paragraph we will use their notion of psychological sequential equilibrium to solve a simple two-stage psychological game.

The most well-known example of a psychological-game based application is Rabin's (1993) model of intention-based reciprocity, according to which players wish to act kindly (unkindly) in response to kind (unkind) actions. The key notion of kindness depends on beliefs in such a way that reciprocal motivation can only be described using psychological games.

However, the range of topics that have been explored in models of belief dependent

[1] As Dufwenberg (2006) correctly points out, 'the term *game with belief-dependent motivation* would be more descriptive than the term *psychological game*, but we stick with the latter which has become established' (p. 2).

motivation is still limited. BD suggest that there is a variety of interesting forms of belief-dependent motivations waiting to be analytically explored. In his survey paper on 'Emotions and Economic Theory', Elster (1998) argues that a key characteristic of emotions is that 'they are triggered by beliefs' (p. 49). He discusses, *inter alia*, anger, hatred, guilt, shame, pride, admiration, regret, rejoicing, disappointment, elation, fear, hope, joy, grief, envy, malice, indignation, jealousy, surprise, boredom, sexual desire, enjoyment, worry, and frustration.

2.2. An example: trust games as psychological games

Let us show and analyze some of the more important features of BD's psychological game-theoretic framework by means of a simple trust game where some belief-dependent motivations can emerge.

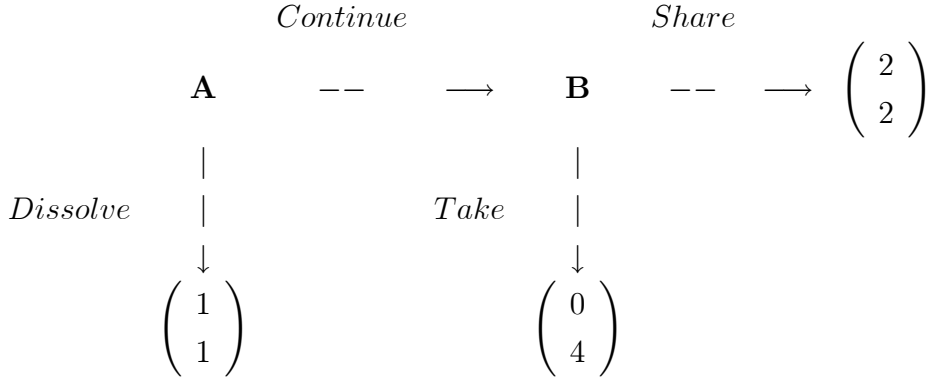

Figure 1 Trust Game with material payoffs

We build on a simple two-stage game representing the following economic situation of strategic interaction. Player A (the truster, 'he') and B (the trustee, 'she') are partners on a project that has thus far yielded total profits of 2€. Player A has to decide whether to withdraw from the project or not. If player A dissolves the partnership, the contract dictates that the players split the profits fifty-fifty. If player A leaves his resources in the project, total profits would be higher (4€); however, according to the contract, in that case player B has the right to share or not the profits after the project is completed. So, player A must decide whether to *Dissolve* or to *Continue* the partnership, without knowing if there will be profit sharing in case he continues. *After* knowing player A's

choice and only in case player A has chosen to *Continue* the partnership, player B has to decide whether to *Take* or *Share* the higher profits. The game tree with the *material* payoffs is represented in Figure 1. Payoffs are in euros and do not necessarily represent preferences. For this reason we call them 'material payoffs'.

We know from traditional game theory that the unique subgame perfect equilibrium of the two-stage trust game (with material payoffs) in Figure 1 is player A choosing *Dissolve* and player B choosing *Take* if A would *Continue*.

Now, suppose that we represent this game in a laboratory and let participants play it in pairs. According to the experimental literature on this subject,[2] one should expect quite a high percentage of pairs with outcomes $(Continue, Share)$.

A reason for this could be the bounded rationality of some A or B players (or both). However, it seems very difficult to support that in a so simple and clear game a quite high number of players (or pairs) are not able to understand the rules of the game or to calculate their profit-maximizing action given their expectations (first-order beliefs) about their opponent's choice.[3]

Another reason could be that player A or player B are motivated not only by self-interest, but also, at least some of them, by 'social preferences'.

Let us first concentrate on distribution-dependent preferences *à la* Fehr and Schmidt (1999). As stated above, this kind of preferences can be addressed by traditional game theory. More specifically, let us suppose that B is inequity averse (i.e. motivated by fairness) and that this is common knowledge among players. Formally, B's preferences are represented by the utility function

$$\begin{aligned} u_B(s_A, s_B) &= \pi_A(s_A, s_B) \\ &\quad -k \max\{0, \pi_B(s_A, s_B) - \pi_A(s_A, s_B)\} - h \max\{0, \pi_A(s_A, s_B) - \pi_B(s_A, s_B)\} \end{aligned}$$

where s_i is the strategy of player $i = A, B$, $\pi_i(s_A, s_B)$ is the material payoff of player $i = A, B$ and k, h are positive parameters such that $k \in [0, 1]$ and $h \in [0, k]$. A is a self-interested player and knows $u_B(s_A, s_B)$. Given the payoff structure of our simple trust game, the utility function of B reduces to $u_B(s_A, s_B) = \pi_A(s_A, s_B) -$

[2] Among others, Charness and Dufwenberg (2006). See section 3 for a complete review of the experimental literature on this family of trust games.

[3] A reasonable explanation for outcomes that partially differ from the subgame perfect equilibrium one could be the 'uncorrectness of beliefs' of some players. For example, suppose that A and B are both self-interest and rational, i.e. they maximize their (expected) material payoff. Suppose also that A believes that B is not rational and so he chooses *Continue*, being quite sure that B will choose *Share* after *Continue*. Being rational and self-interested, B instead chooses *Take* after *Continue* and so the outcome $(Continue, Take)$ takes place.

$h \max\{0, \pi_A(s_A, s_B) - \pi_B(s_A, s_B)\}$ and so $(Continue, Share)$ is the (unique) equilibrium outcome of the trust game in case B is highly inequity averse, i.e. $k \in \left(\frac{1}{2}, 1\right]$ and $h \in \left(\frac{1}{2}, k\right)$. In that case, $(Continue, Share)$ outcomes should emerge experimentally independently of B's expectation of A's expectation on B choosing *Share* after *Continue*.

However, as we shall report below, experimental evidence on trust games shows that there is a certain correlation between B's expectation of A's trust on her and trust fulfilment; that can be explained by a particular kind of social preferences: those expressed as belief-dependent motivations. As said above, traditional game theory is ill-equipped to address such preferences. For that reason, we need to introduce some psychological game tools.

Consider the game with material payoffs in Figure 1.

Let us first suppose that A is a self-interested player, while B is (also) a guilt-averse player.

Guilt aversion can be defined as follows: people suffer from guilt if they inflict harm on others; although guilt could have a variety of sources, one preeminent way to inflict harm is to let others down (see Tangney, 1995). Battigalli and Dufwenberg (2007) develop a general theory of guilt aversion and show how to solve for sequential equilibria. Their approach can be easily applied to our simple trust game, whenever we assume that B is affected by guilt. Moreover, Charness and Dufwenberg (2006) suggest that sensitivity to guilt aversion imposes a specific behavior to the trustee (B): suppose that in an experiment it is possible to truthfullly elicit B's expectation of A's expectation that B would *Share* after *Continue* (B's second-order beliefs of *Share*); given that the $(Continue, Share)$ outcome indicates trust fulfilment, as we shall report below, experimental evidence on trust games shows positive correlation between B's second-order beliefs of *Share* (after *Continue*) and trust fulfilment and thus higher expected mean guess of Bs who choose *Share* after *Continue* than the mean guess of Bs who choose *Take* after *Continue*.

We first consider the psychological utility of player A. Since A is only motivated by self-interest, his feeling sensitivity to each possible belief-dependent motivation is equal to zero. Therefore, A's total utility function reduces to his material payoff.

Next, let us concentrate on the psychological utility of player B. Since B is motivated by guilt aversion, she takes into account the disappointment of player A when the material payoff he expects to receive after *Continue* does not match the one he actually receives. The material payoff A expects to receive after *Continue* depends on his

first-order beliefs on B's strategy.

Let us analyze the situation from a formal point of view: define A's *initial* first-order belief that B will *Share* if A chooses *Continue* as $\alpha_A = \Pr_A[Share \text{ if } Continue]$. Hence, α_A measures A's trust on B *before* the game starts. Define also B's *conditional* 2nd-order belief that she would *Share* if A would *Continue* as $\beta_B = \mathrm{E}_B[\alpha_A|Continue]$. Hence, β_B measures B's expectation of A's trust on her *given that* B knows that A has chosen *Continue*.[4]

Given the notation we introduced, we can express A's expected material payoff after *Continue* as $E_{Cont,\alpha_A}[\pi_A] = 2 \cdot \alpha_A + 0 \cdot (1 - \alpha_A) = 2\alpha_A$. How much would A feel 'let down' after $(Continue, Take)$? According to BD, the amount of his disappointment is exactly $-2\alpha_A$, i.e. the difference between the payoff he receives after $(Continue, Take)$, which is zero, and the payoff he would have received after $(Continue, Share)$.

B's guilt is given by her expectation of A's disappointment, given that A has chosen *Continue*, i.e. B's expectation of $(-2\alpha_A)$**,** given *Continue*. This amount, $-2\beta_B$, multiplied by B's sensitivity to guilt aversion, $\theta_B^g \geq 0$, is exactly B's psychological utility when A chooses *Continue* and she chooses *Take*. B's total utility after $(Continue, Take)$ is given by $4 - \theta_B^g 2\beta_B$, i.e. the sum of her material payoff and her psychological utility.

The psychological game representing the trust game with a self-interested truster and a guilt-averse trustee is depicted in Figure 2. What appears at the terminal histories should be thought of as utilities, not as material payoffs, although the two notions coincide for all but one of the terminal histories.

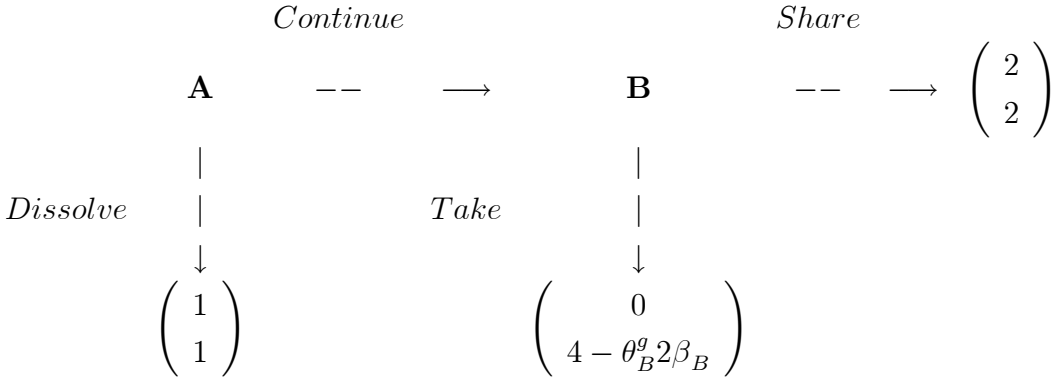

Figure 2 Trust Game with guilt aversion.

[4]In general, we use the Greek letter α to refer to *first*-order beliefs, and the letter β to refer to *second*-order beliefs.

Looking at Figure 2 we can conclude that in the simple trust game we are analyzing, B exhibits *guilt aversion* if her expected utility from playing *Take* after *Continue* depends negatively on her expectation of α_A, conditional on A choosing *Continue*.

Let us now suppose that A is again a self-interested player, while B is (also) a reciprocity-concerned player.

Reciprocity has two sides: *positive* reciprocity, where a player is kind in return to another one's kind choice, and *negative* reciprocity, where a player is unkind in return to another one's unkind choice. Rabin's (1993) theory of reciprocity, in which players reciprocate belief dependent (un)kindness with (un)kindness, is probably the most well-known application of GPS's psychological game theory. Rabin works with the normal form version of GPS's theory. His goal is to highlight certain key qualitative features of reciprocity, and he does not address issues of dynamic decision making, although he points out that this is important for applied work (p. 1296). Dufwenberg and Kirchsteiger (2004), while building on the same framework as Rabin in defining reciprocity, depart from it, developing a theory of reciprocity for extensive games. As BD, in dealing with sequential reciprocity they argue that it is necessary to deviate from GPS's extensive form framework: GPS only allow *initial* beliefs to enter the domain of a player's utility, while the modeling of reciprocal responses at various nodes of a game tree requires kindness to be re-evaluated using updated belief.

We build on the same intuition as Rabin (1993) according to which modeling reciprocity may require belief-dependent utilities, since kindness and perceived kindness depend on beliefs. And we also take into account the suggestion of Dufwenberg and Kirchsteiger (2004) that reciprocity in a dynamic setting is a 'conditional' belief-dependent motivation, in the sense that updated belief matters when evaluating the kindness of a player. Nonetheless, following Battigalli (2007), we model reciprocity in an easier and more direct way compared to the formal definition of Rabin (1993), although we build on the same form of belief-dependency.

Let us start by defining A's kindness (and B's perceived kindness) through a simple question: if A chooses *Continue* can we safely postulate that he has been kind to B? The answer is 'possibly not', because, in the case A is quite certain that B would choose *Share* after *Continue* (say, $\alpha_A > \frac{1}{2}$), he is just maximizing his material payoff. Thus, when A chooses *Continue*, he is (and is perceived to be) more kind, the less he thinks that B would choose *Share* after *Continue*. Obviously, in case A chooses *Dissolve*, he is not kind to B and we can intuitively state, without loss of generality, that A's kindness to B is negative (see the exact calculation below). Hence, *Continue* may be deemed

'kind', if α_A is low. *Continue* is perceived as kind by B if β_B is low. If β_B is low and B is highly motivated by reciprocity considerations, she reciprocates and chooses *Share* after *Continue*.

Formally, we define the 'kindness' of player i as a function of his strategy and beliefs:[5] i is kind (unkind) to j if he intends to make j get more (less) money than a context-dependent '*equitable payoff*' $\pi_j^{e_i}$, with $i, j = A, B$, $i \neq j$. For example, the equitable payoff ascribed by B to A, given B's belief, can be expressed with the formula

$$\pi_A^{e_B}(\alpha_B) = \frac{1}{2}\left(\max_{s_B} E_B[\pi_A; \alpha_B, s_B] + \min_{s_B} E_B[\pi_A; \alpha_B, s_B]\right)$$

where $\alpha_B = \Pr_B[Continue]$, i.e. it measures B's initial first-order belief that A would choose *Continue*. The equitable payoff is defined in this case as the average between the highest and the lowest expected payoff that B can allow to A, given her initial belief on the action chosen by A. Therefore, we can define 'Kindness' of B towards A as the difference between the expected payoff B would allow to A and the equitable payoff, i.e.

$$K_B(\alpha_B, s_B) = E_B[\pi_A; \alpha_B, s_B] - \pi_A^{e_B}(\alpha_B)$$

Notice that B's kindness depends on B's intentions.

In the same way, we can define $E_B[K_A; \beta_B]$, i.e. player B's perceived kindness of A. Since A's kindness depends on the first-order belief of A (his belief about s_B), then player B's perceived kindness depends on the second-order belief of B (belief of B about belief of A).

Given B's sensitivity to reciprocity, $\theta_B^r \geq 0$, we express the psychological part of B's (total) utility function when she is sensitive to reciprocity as in Battigalli (2007), i.e. as the product of B's sensitivity to reciprocity, her perceived A's kindness and A's material payoff.[6] Then, the total expected utility function of B when she is (also) concerned with

[5] We point out a subtle issue. Here we follow BD and assume that the kindness of a player depends on his beliefs and his actual *strategy*. This means that observing a player's behavior may force a change in the perception of his kindness. Battigalli (2007) note that kindness should be more appropriately modeled as a function of a player's beliefs about the *other*'s *behavior* and about *his own behavior*. When coupled with the trembling hand logic of sequential equilibrium (according to which deviations are unintentional and players never change their beliefs about the beliefs of others), this would imply that the perception of the co-player's kindness is always the same, on and off the equilibrium path, thus trivializing the equilibrium analysis of sequential reciprocity.

[6] Battigalli (2007) himself acknowledges that several functional forms could be able to adequately represent the psychological preferences of a player concerned with intention-based reciprocity. Among other reasons, we chose to use his formulation because of its simplicity and tractability.

reciprocity is:

$$u_B((\alpha_B, s_B); \beta_B) = \pi_B(\alpha_B, s_B) + \theta_B^r \cdot E_B[K_A; \beta_B] \cdot \pi_A(\alpha_B, s_B)$$

Given that we assumed A being a self-interested player, we have $\theta_A^r = 0$, hence again A's total utility function reduces to his material payoff.

Next, let us concentrate on the psychological utility of player B. According to Battigalli's (2007) formulation, A's Kindness when he chooses *Continue* is

$$K_A(Continue, \alpha_A) = 2\alpha_A + 4(1 - \alpha_A) - \frac{1}{2}\left(2\alpha_A + 4(1 - \alpha_A) + 1\right) = \frac{3}{2} - \alpha_A$$

and so B's perceived Kindness when A chooses *Continue* is

$$E_B(Continue, \beta_B) = \frac{3}{2} - \beta_B$$

Therefore,[7]

- B's total utility after $(Continue, \mathbf{Share})$ is given by

$$u_B((Continue, \mathbf{Share}); \beta_B) = 2 + \theta_B^r \cdot \left(\frac{3}{2} - \beta_B\right) \cdot 2 = 2 + \theta_B^r \left(3 - 2\beta_B\right)$$

- B's total utility after $(Continue, \mathbf{Take})$ is given by

$$u_B((Continue, \mathbf{Take}); \beta_B) = 4 + \theta_B^r \cdot \left(\frac{3}{2} - \beta_B\right) \cdot 0 = 4$$

Using the same procedure, we can calculate B's total utility after *Dissolve*, i.e.

$$u_B((Dissolve); \beta_B) = 1 + \theta_B^r \left(\beta_B - \frac{3}{2}\right)$$

which is no greater than 1, given that $\theta_B^r \geq 0$.

The psychological game representing the trust game with a self-interested truster and a reciprocity concerned trustee is depicted in Figure 3. Again, what appears at the terminal histories should be thought of as utilities, not as material payoffs, although the

[7] From now on, we denote with the 'bold' labels **Take** and **Share** player B's strategies '*Take* if *Continue*' and '*Share* if *Continue*' respectively.

two notions coincide for one of the terminal histories.

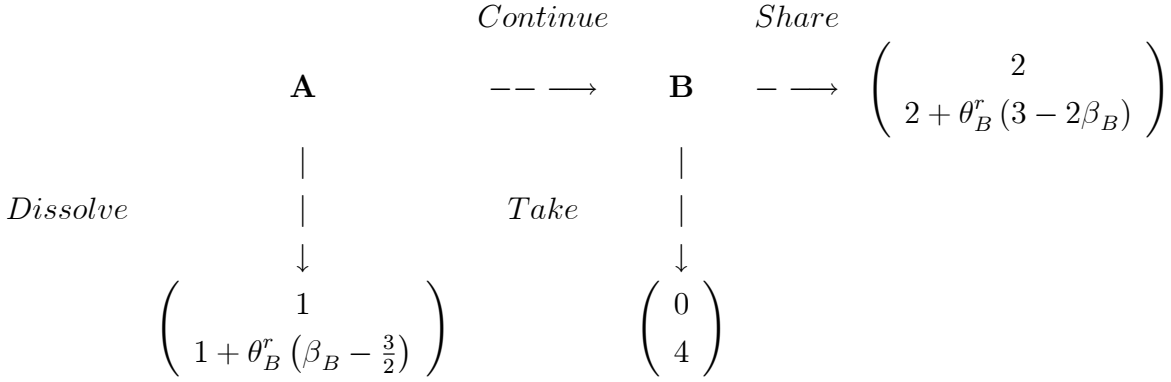

Figure 3 Trust Game with reciprocity.

2.3. Psychological games: some misconception dispelled

In this subsection we want to analyze more deeply some of the features of our psychological game framework in order to make the main concepts clearer and to dispel some misconceptions that could emerge while comparing psychological game theoretical tools to the well-known features of traditional game theory. We try to answer the questions that could be asked by a reader with a good background in game theory that approaches for the first time the games depicted in Figure 2 and in Figure 3. We bear in mind the main assumptions of the general framework of psychological games as described in BD.

Question 1. In Figure 2, why do we suppose that player B cares about A's disappointment even though A is just an expected material payoff maximizer?

Answer. Our analysis is *descriptive*, not *prescriptive*. Utility functions are meant to help analyzing players' behavior. They describe hypothetical preferences (i.e. conditioned to some hypothesis). In that sense, utilities in Figure 2 are only 'instruments'. We do not use them to represent players' happiness at the end of the game. It is possible that A feels disappointment but his behavior is not affected by the anticipation of such disappointment. Nonetheless, even though A's feelings do not influence his own behavior, it is possible that B cares about them.

Moreover, for the trust game it can be shown:

- that the qualitative analysis does not change if we put disappointment in A's utility function.

- that the set of pure strategy equilibria of the psychological game is the same both in case we add A's disappointment in his utility function and in case we do not add it.

Hence, we choose not to consider it in order to simplify both the equilibrium calculations and the analysis of the utility functions of the game.

Question 2. Given the high number of unknown parameters in a psychological game, under which conditions can we say that we are in a situation of *complete* information?

Answer. If we assume that the material game form in Figure 1 is common knowledge, and that players are expected material payoff maximizers and this is also common knowledge, then the trust game in Figure 1 has *complete* information. If we assume that players have psychological (belief-dependent) preferences, then it is more reasonable to allow for *incomplete* information, which comes from two sources:

(i) player i does not know which feeling j is sensitive to or, even if i knows what is the feeling, j does not know i knows that and so on. In our trust game, this happens for example if A does not know whether B is only self-interested, or (also) sensitive to guilt, or to reciprocity or to both. This means that A does not know if he is playing the trust game in Figure 1, that one in Figure 2, that one in Figure 3 or a trust game in which B's total utility after $(Dissolve)$ is $1+\theta_B^r\left(\beta_B-\frac{3}{2}\right)$, after $(Continue, Take)$ is $4-\theta_B^g 2\beta_B$ and after $(Continue, Share)$ is $2+\theta_B^r\left(3-2\beta_B\right)$, i.e. a psychological game which includes both guilt and reciprocity of player B.[8]

(ii) even if i knows which feeling j is sensitive to, i does not know j's sensitivity to that specific feeling. This happens, for example, in case A and B know they are playing the psychological game in Figure 2, but A does not know the value of θ_B^g.[9]

In strategic settings in which the two players do not know each other very well, condition (i) applies. In all these cases, we deal with a Psychological game with *Incomplete* Information.

Now suppose that we are in a setting in which condition (i) does not apply, i.e. the set of belief-dependent motivations to which each player is sensitive to is common knowledge among players. This is the case, for example, in which the game structure in Figure 2

[8] Another case would be that A knows with certainty that B is sensitive to, say, guilt aversion, but B does not know A knows that.

[9] Another case would be that in which A and B know they are playing the psychological game in Figure 2, A knows the value of θ_B^g, but B does not know that A knows it.

is common knowledge. Then, we say that the trust game in Figure 2 is a Psychological Game with *Complete* Information when θ_B^g is *commonly known* among players. In case θ_B^g is *not commonly known* among players, we deal again with a Psychological game with *Incomplete* Information.[10]

In both cases, β_B is *unknown* to A and is *endogenous*. Therefore, despite information is complete, in a psychological game there is still a source of uncertainty, that is not solved even after the end of the game. This is because A will not know the realized value of B's conditional second-order belief in any of the three end nodes of the game of Figure 2.

Question 3. According to BD, an easier way to represent and analyze the psychological game in Figure 2 is to write B's utility after $(Continue, Take)$ as $4 - \theta_B^g 2\alpha_A$ and, given that B does not know A's first-order beliefs about her strategy, suppose that she maximizes the *expected value* of the previous expression. In other words, B maximizes the expected value of a *state-dependent* utility function. It may be argued that it does not make any sense to insert a *state* variable in a *utility* function if that *state* is not revealed ex post or if that *utility* is not 'experienced' ex-post. This is because even after the end of the game, B would not know the exact value of α_A and so she would not know her exact total utility after $(Continue, Take)$. Therefore, a reasonable question could be: does it make sense to express B's belief-dependent motivations by inserting α_A into B's utility function?

Answer. As emphasized above, the utility functions are analytical tools used to explain players' behavior, not to measure their happiness: u_B does not represent the utility 'experienced' by B. In the simple trust game we analyze, we just want to describe the behavior of player B, who maximizes the expected value of her total utility (material payoff and psychological utility). Therefore, it is not necessary for the state variable or the induced utility to be indeed 'experienced' ex-post. Their only role is to help us analyzing players' motivations. Doing that by using (initial) first-order beliefs is much simpler than using (conditional) second-order ones.

Question 4. What is the difference between psychological games (with complete or incomplete information) and traditional games with incomplete information?

Answer. The payoffs of a (traditional) game with incomplete information depend on players' actions and on an *exogenous* parameter representing the state of nature, about which players are asymmetrically informed. In psychological games, players' beliefs in

[10] The same definitions can be extended to the trust game in Figure 3, given that θ_B^r is common knowledge or not, respectively.

the utility functions are not (exogenous) parameters, but rather endogenous variables. Consider the psychological game in Figure 2: β_B is endogenous because α_A is an endogenous variable and any belief on an endogenous variable is endogenous. β_B is a conditional second-order belief (conditional on A choosing *Continue*) and, at the same time, a terminal belief. Moreover, what B chooses to do is endogenous, but, by construction, β_B is not influenced by B's choice: B's beliefs on her opponent's beliefs and actions do not depend on what she decides to do after A has chosen *Continue*.

2.4. Solving a Psychological Game with Complete Information

In this subsection, we show how to solve a *Dynamic Psychological Game* with *Complete Information*. More specifically, we will solve the two psychological games in Figure 2 and in Figure 3, supposing respectively that θ_B^g and θ_B^r are commonly known among players. Although it could seem unrealistic to suppose that each others' feelings sensitivity is common knowledge, this assumption is needed as a first step in order to predict players' behavior when they hold belief-dependent motivations.

In general, the equilibrium concepts we use in order to solve psychological games are not 'different' from those used in traditional game theory. Starting from GPS, BD simply apply the same logic as Nash Equilibrium to solve static psychological games and the standard logic of backward induction and of the (traditional) sequential equilibrium to solve dynamic psychological games. They are somehow 'forced' to generalize the previous concepts in order to consider the fact that in psychological games beliefs enter players' utility function. That creates two additional problems that need to be taken into account when dealing with equilibrium behavior:

(a) the condition of *correctness of beliefs* requires also *correctness of stated utilities.* Given that different orders of beliefs are involved in the analysis, we have to explicitly impose that they are correct in equilibrium (for example, in each of the game in Figure 1, 2 and 3, it must be $\alpha_A = \beta_B$ in equilibrium). Nonetheless, when checking that in equilibrium players maximize their total utilities at all decision nodes given their correct beliefs about one another's actions, we must take into account that some of these beliefs (independently of the order) are part of (some of) the utilities of (some) players. Hence, in equilibrium, players maximize their total utilities measured according to the correct beliefs they hold on actions and beliefs of their opponents and, at the same time, (correct) beliefs of different orders match players' best replies calculated according to the 'correctly' stated total util-

ities. This issue is addressed by GPS and generalized by BD, who allow for both *own* beliefs and beliefs of *others* in the (total) utility function; they also provide a general framework able to account for conditional beliefs of whatever order in players' (total) utility function.

(b) as the play unfolds, beliefs about the beliefs of others are revised; but players' *beliefs updating* leads also to players' *utilities updating*.
Building on the theory of hierarchies of conditional beliefs due to Battigalli and Siniscalchi (1999), BD model a universal belief space that accounts for updating beliefs about others' beliefs. The idea is that, in a psychological game, in order to decide on the best course of action, player i may need to form (conditional) beliefs about the infinite hierarchies of (conditional) beliefs of other players, because they enter her psychological utility function. For example, in the trust games in Figure 2 and Figure 3, since β_B enters B's psychological utility function, in order to decide if she prefers to *Take* or to *Share* after *Continue*, B may need to form conditional beliefs about the first-order beliefs of A. But when B updates her second-order beliefs after A has chosen *Continue*, her utility from choosing *Take* after *Continue* is also updated.

GPS propose a notion of psychological equilibrium that takes into account (a), but does not solve (b), since they allow only initial pre-play beliefs to enter players' utility functions. Hence, GPS's notion of equilibrium cannot encompass belief-dependent motivations as those considered here.

Building on the previous considerations, BD propose a notion of *Psychological Sequential Equilibrium* (*PSE* henceforth), which generalizes the sequential equilibrium concept of Kreps and Wilson (1982). As stated in BD, 'Kreps and Wilson (1982) argue that an appropriate definition of equilibrium in extensive form games must refer to *assessments*, that is, profiles of (behavior) strategies and conditional (first-order) beliefs. They formulate a definition of sequential equilibrium in two steps: first, they put forward a *consistency* condition for assessments, and then they stipulate that an assessment is a sequential equilibrium if it is consistent and satisfies sequential rationality' (p. 18). BD follow a similar two-step approach, adding to it a third requirement concerning the higher-order beliefs that need to be specified in psychological games. This third condition is analogous to a condition used by GPS. Essentially, it requires that players hold at each node common, correct beliefs about each others' beliefs. This implies that in

equilibrium players do not change their beliefs about the beliefs of the opponents.[11]

BD's *PSE* is the equilibrium concept that we use to solve (static and) dynamic psychological games; it reduces to GPS's equilibrium when dealing with their restricted class of psychological preferences, i.e. when the utility of a terminal node for a player depends only on the hierarchy of *initial* beliefs of *that* player. BD also show that the equivalence can be extended to some dynamic psychological games not covered by GPS, i.e. those where the *initial* (but not the updated) beliefs of *others* enter the utility function. Nonetheless, this extended equivalence result only concerns sequential equilibria, and BD argue that the non-equilibrium analysis of psychological games is important. Moreover, their definition of sequential equilibrium makes essential use of conditional higher-order beliefs whereas GPS have to take an indirect route whereby they look at the subgame perfect equilibria of a fictitious standard game associated to initial belief hierarchies that satisfy a fixed point condition.[12]

In the next subsections, we will use the *PSE* as solution concept for our simple psychological game of trust, pointing out, whenever possible, the differences with the *Psychological Subgame Perfect Equilibrium à la* GPS.

2.4.1. PSE of the Trust Game with Guilt Aversion

Let us first find the set of *PSE* for the psychological game in Figure 2, according to all possible values of the positive parameter θ_B^g. We will focus on the equilibrium values of the key beliefs α_A, β_B (already defined above) and $\alpha_B = \Pr_B[Continue]$. Since in equilibrium beliefs are correct, the first-order belief α_A may be interpreted as the mixed strategy of B, and the first-order belief α_B may be interpreted as the mixed strategy of A. Furthermore, the correctness condition also imply that in equilibrium $\alpha_A = \beta_B$.

For $\theta_B^g \in (0,1)$, it is

$$u_B((Continue, \mathbf{Take}), \beta_B) > u_B((Continue, \mathbf{Share}), \beta_B) \qquad \forall \beta_B \in [0,1]$$

[11] BD argue that this feature is implicit in the sequential equilibrium concept as defined for standard games, although they find it objectionable.

[12] The GPS equilibrium concept for dynamic psychological games requires correctness of initial pre-play beliefs and that players optimize their beliefs and choices at all given decision nodes. In both games in Figure 1 and Figure 2, B's payoff depends directly on β_B, and it is necessary to impose explicitly that the beliefs α_B and β_B are correct in equilibrium, i.e. they match the (behavior) strategy of player B. To this end, note that *given* β_B, both psychological games have real numbers characterizing payoffs at each end-node. In this sense, they reduce to a 'standard game'. Therefore players must play a subgame perfect equilibrium in this (correct) belief-reparameterized game.

In games with imperfect information about past moves GPS look at the *sequential* equilibria of the fictitious standard game.

Thus, B always chooses *Take* and, going backward, A chooses *Dissolve*; hence, the only *PSE* profile of beliefs is $\alpha_B = 0, \alpha_A = \beta_B = 0$.

For $\theta_B^g \in [1, +\infty)$, it could be $u_B((Continue, \mathbf{Take}), \beta_B) \lesseqgtr u_B((Continue, \mathbf{Share}), \beta_B)$ depending on the value of β_B.

Let us first look at 'pure strategy'[13] equilibria.

For $\beta_B = 0$, it is $u_B((Continue, \mathbf{Take}), \beta_B) > u_B((Continue, \mathbf{Share}), \beta_B)$, hence B chooses *Take* and, since A holds correct beliefs in equilibrium ($\alpha_A = 0$), he chooses *Dissolve.* Hence, $\alpha_B = 0$, $\alpha_A = \beta_B = 0$ is a *PSE in beliefs.*

For $\beta_B = 1$, it is $u_B((Continue, \mathbf{Take}), \beta_B) < u_B((Continue, \mathbf{Share}), \beta_B)$, hence B chooses *Share* and, since A holds correct beliefs in equilibrium ($\alpha_A = 1$), he chooses *Continue.* Hence, $\alpha_B = 1$, $\alpha_A = \beta_B = 1$ is a *PSE in beliefs.*

Let us now look for mixed equilibria.

B is indifferent and can optimally play a mixed strategy if and only if $4 - \theta_B^g 2\beta_B = 2$, i.e. $\beta_B = \frac{1}{\theta_B^g}$. By the requirement of correctness of first and second-order beliefs, in every mixed strategy equilibrium it must be $\alpha_A = \beta_B = \frac{1}{\theta_B^g}$.[14] What about A? Given that $\alpha_A = \frac{1}{\theta_B^g}$, A chooses *Continue* if $2 \cdot \frac{1}{\theta_B^g} + 0 \cdot \frac{\theta_B^g - 1}{\theta_B^g} > 1$, i.e. $\theta_B^g \in (1, 2)$, he chooses *Dissolve* if $\theta_B^g \in (2, +\infty)$ and he randomizes if and only if $\theta_B^g = 2$. By the requirement of correctness of first-order beliefs in equilibrium, it is $\alpha_B = 1$ if $\theta_B^g \in (1, 2)$, $\alpha_B = 0$ if $\theta_B^g \in (2, +\infty)$ and $\alpha_B \in [0, 1]$ if $\theta_B^g = 2$.

Keeping in mind that in every equilibrium we require (by definition) $\alpha_A = \beta_B$, we summarize in the following table all the possible *PSE* of the psychological game in Figure 2, for all values of the positive parameter θ_B^g.

Player B's *g*uilt sens.	Pure Equilibrium $(Dissolve, \mathbf{Take})$	Pure Equilibrium $(Continue, \mathbf{Share})$	Mixed Equilibrium
$\theta_B^g \in (0,\ 1)$	Yes	No	No
$\theta_B^g = 1$	Yes	Yes	No
$\theta_B^g \in (1,\ 2)$	Yes	Yes	$\alpha_B = 1$, $\alpha_A = \frac{1}{\theta_B^g}$
$\theta_B^g = 2$	Yes	Yes	$\alpha_B \in [0, 1]$, $\alpha_A = \frac{1}{2}$
$\theta_B^g \in (2, +\infty)$	Yes	Yes	$\alpha_B = 0$, $\alpha_A = \frac{1}{\theta_B^g}$

Table 1 *P*sychological *S*equential *E*quilibria, given a guilt averse B

[13] We use the word *strategy* in an improper way, since we are looking for equilibria in beliefs. The word *pure*, in fact, refers not only to strategies, but also to beliefs, in the sense that $\alpha_A, \alpha_B, \beta_B \in \{0, 1\}$.

[14] In order for the equilibrium to be in mixed strategies it must be $\theta_B^g \neq 1$, otherwise B would play the pure strategy **Share**. Thus, we look for mixed strategy equilibria only for $\theta_B^g \in (1, +\infty)$.

Note that (*Dissolve*, **Take**) is always an equilibrium, but if B is sufficiently sensitive to guilt ($\theta_B^g \geq 1$) there are multiple equilibria. Is there a reasonable selection criterion? Like in Dufwenberg (1995, 2002), consider the following intuitive 'psychological forward induction' argument: it is *rational* for A to trust B only if he assigns at least 50% probability to strategy **Share**. In other words, A's optimal decision depends on B's choice: he has to compare the expected payoff of choosing *Continue*, that is $2\alpha_A$, with the (certain) payoff of choosing *Dissolve*, that is 1. Therefore, a rational A chooses to trust B and so to *Continue* the partnership only if it is $\alpha_A \geq \frac{1}{2}$. Then, whenever B observes the action *Continue*, and if he believes A is rational, then it must be the case that $\beta_B \geq \frac{1}{2}$. The logic of this argument depends on belief β_B being conditional on A choosing *Continue*. The argument cannot be recast using GPS's theory since it allows only initial beliefs, but it can be captured in BD's framework, since it allows all hierarchies of conditional beliefs[15] and considers a slightly different concept of sequential rationality.

The forward induction story involves somehow a change of beliefs, and, above all, the beliefs entering the utility functions are those held by B when his node is reached (thus not only the initial ones, as it was in GPS). Such a criterion, in which players strategically affect their opponents' beliefs, has an even more pervasive strength in psychological games than it has in traditional games, because of the *direct* effect of beliefs on utilities.

We conclude this subsection with an important remark: even if A has the opportunity to *signal* his own beliefs and to force B to hold certain second-order beliefs, it would be incorrect to infer that A chooses his own beliefs along with his own strategy. A affects β_B (a conditional second-order belief of B) and *in equilibrium* it must be the case that $\alpha_A = \beta_B$, but A takes α_A as given.[16]

2.4.2. PSE of the Trust Game with Reciprocity

Next we compute the set of *PSE* for the psychological game in Figure 3, for all possible values of the positive parameter θ_B^r.

[15] Battigalli and Siniscalchi (2002) characterize the *forward induction* logic in term of *strong belief in rationality*: the intuitive idea is that when players update their beliefs, they do this in a way that is consistent with the assumption of rationality of their opponents. They try to 'rationalize' the opponents' actions, that is to infer, from their choices, the beliefs of their opponents. These beliefs are such that the observed moves are best responses to them. BD show how this forward induction logic can be extended to psychological games.

[16] Similarly, in the perfect equilibrium of a leader-follower game, the leader affects the action of the follower, which is a best response, but the equilibrium choice of the leader takes as *given* the best-response *strategy* (contingent choice) of the follower.

For $\theta_B^r \in \left(0, \frac{2}{3}\right)$, it is

$$u_B((Continue, \mathbf{Take}), \beta_B) > u_B((Continue, \mathbf{Share}), \beta_B) \qquad \forall \beta_B \in [0,1]$$

Thus, B always chooses *Take* and, going backward, A chooses *Dissolve*; hence, the only *PSE in beliefs* is $\alpha_B = 0, \alpha_A = \beta_B = 0$.

For $\theta_B^r = \frac{2}{3}$, it is

$$\begin{cases} u_B((Continue, \mathbf{Take}), \beta_B) > u_B((Continue, \mathbf{Share}), \beta_B) & \text{for } \beta_B \in (0,1] \\ u_B((Continue, \mathbf{Take}), \beta_B) = u_B((Continue, \mathbf{Share}), \beta_B) & \text{for } \beta_B = 0 \end{cases}$$

Therefore, $\alpha_B = 0, \alpha_A = \beta_B = 0$ is again the only *PSE in beliefs*.

For $\theta_B^r \in \left(\frac{2}{3}, 2\right)$, no 'pure' *PSE* exists. The intuitive reason is that the more B thinks that A expected him to *Share*, the less B perceives A as kind, the higher B's incentive to *Take*. Recall that in any *PSE* we have $\alpha_A = \beta_B$. The indifference condition for B, given that A chooses *Continue*, yields $\beta_B = \frac{3\theta_B^r - 2}{2\theta_B^r}$. If $\beta_B < \frac{3\theta_B^r - 2}{2\theta_B^r}$, B's perception of A's kindness, $(\frac{3}{2} - \beta_B)$, is high and so she prefers to *Share*. Since α_A has to be correct ($\alpha_A = \beta_B$), this would imply $\alpha_A = 1$, which contradicts $\alpha_A = \beta_B < \frac{3\theta_B^r - 2}{2\theta_B^r}$. If $\beta_B > \frac{3\theta_B^r - 2}{2\theta_B^r}$, then B's perception of A's kindness is low and she prefers to *Take*, leading to a similar contradiction. The mixed equilibrium is easily obtained as a function of θ_B^r from the indifference condition $\beta_B = \frac{3\theta_B^r - 2}{2\theta_B^r}$ and the equilibrium condition $\alpha_A = \beta_B$, taking the best response of A to α_A.

For $\theta_B^r \in (2, +\infty)$, it is $\frac{3\theta_B^r - 2}{2\theta_B^r} > 1$. Then, B prefers to *Share*, independently of her perception of A's kindness, Therefore, the only *PSE in beliefs* is $\alpha_B = 1$, $\alpha_A = \beta_B = 1$. In order for $\alpha_B = 0, \alpha_A = \beta_B = 0$ to be a *PSE in beliefs* it must be $\alpha_A = \beta_B > \frac{3\theta_B^r - 2}{2\theta_B^r}$, which is impossible, given that $\alpha_A, \beta_B \in [0,1]$.[17]

Therefore, if B is sufficiently sensitive to reciprocity ($\theta_B^r \in [2, +\infty)$), $(Continue, \mathbf{Share})$ is the unique equilibrium. This is because, according to Dufwenberg and Kirchsteiger (2004) formulation of perceived kindness, player A is perceived as kind by B when he chooses an action 'increasing' B's payoff, even in case A's own payoff increases too. If B is highly sensitive to reciprocity (θ_B^r high), she does not mind if the action chosen by A leads to a (relatively) higher material payoff also to him. She looks only at her own payoff and if it is higher than the (equitable) payoff that she believes it would be possible to 'receive' from A, then she reciprocates by playing the 'kind' action *Share*, independently of her perception of A's kindness.

[17] For the limit case $\theta_B^r = 2$, the unique *PSE in beliefs* is again $\alpha_B = 1$, $\alpha_A = \beta_B = 1$.

Keeping in mind that in every equilibrium we require (by definition) $\alpha_A = \beta_B$, we summarize in the following table all the possible *PSE* of the psychological game in Figure 3, for all values of the positive parameter θ_B^r.

Player B's *r*eciprocity sens.	Pure Equilibrium $(Dissolve, \mathbf{Take})$	Pure Equilibrium $(Continue, \mathbf{Share})$	Mixed Equilibrium
$\theta_B^r \in \left(0,\ \frac{2}{3}\right]$	Yes	No	No
$\theta_B^r \in \left(\frac{2}{3}, 1\right)$	No	No	$\alpha_B = 0,\ \alpha_A = \frac{3\theta_B^r - 2}{2\theta_B^r}$
$\theta_B^r = 1$	No	No	$\alpha_B \in [0,1]\,,\ \alpha_A = \frac{1}{2}$
$\theta_B^r \in (1,2)$	No	No	$\alpha_B = 1,\ \alpha_A = \frac{3\theta_B^r - 2}{2\theta_B^r}$
$\theta_B^r \in [2, +\infty)$	No	Yes	No

Table 2 *P*sychological *S*equential *E*quilibria, given a reciprocity concerned B

Finally, we remark that these results depend on the adopted definition of perceived kindness, which is inspired by Dufwenberg and Kirchsteiger (2004). Rabin (1993) proposes a different definition, which yields different psychological utilities and partially different equilibria. In particular, for no value of the (positive) parameter θ_B^r measuring B's sensitivity to reciprocity the equilibrium $(Continue, Share)$ exists. This is because according to Rabin's formulation A is perceived as kind by B only when he 'increases' B's payoff by decreasing his own payoff. When the action chosen by A increases both his payoff and the payoff of B, he is perceived as 'neutral' by B, who does not play the 'kind' action *Share*, even in case she is highly sensitive to reciprocity.

2.5. The 'incomplete information' case: some hints

It is probably not realistic to assume that players know one another's psychological propensities. Unless one models interaction within a family (or amongst friends) and there is not an ex-ante stage in which all players' feelings sensitivities are (elicited and) transmitted to their co-players, we are dealing with a psychological game with (some form of) incomplete information.

In order to extend the analysis of psychological games to include incomplete information, let $\theta = (\theta_A, \theta_B)$ denotes a vector of parameters that summarize all the payoff-relevant aspects of the game that are not common knowledge; θ_i $(i = A, B)$ is a component known to player i only (such as his sensitivity to certain psychological motivations). It is common knowledge that θ belongs to a parameter space $\Theta = \Theta_A \times \Theta_B$. Elements of

Θ are called *states of Nature*. As underlined by BD, in general, in dynamic psychological games, 'one can assume for simplicity that players do not get more refined information about the states of Nature as the play unfolds, they only observe the actions chosen in previous stages of the game' (p. 37). In the simple two-stage game we analyze we can maintain this assumption without loss of generality. It is relatively easy to generalize our construction of the belief space in order to include beliefs about the state of Nature. However, since states of Nature are exogenous, also players' hierarchies of *initial* beliefs about the state of Nature are exogenous, and the model may specify assumptions about such exogenous beliefs.

With this extended framework in place, one can regard the trust game with guilt aversion and reciprocity of B analyzed in the previous subsection as psychological games with incomplete information.

Attanasi, Battigalli and Nagel (2007) provide a specific framework in order to analyze and solve (one-stage) psychological games with *incomplete information*. In the simple trust game under consideration (assuming that $\theta_A = 0$ for each possible belief-dependent motivation and that this is common knowledge), none of the parameters θ_B^g and θ_B^r is known to A: he supposes that (θ_B^g, θ_B^r) is drawn from a joint distribution. Thus, A does not know the true values of θ_B^g and θ_B^r, but has a prior on each of the two. With this extended framework in place, they can regard trust games in which there is not an ex-ante communication of feeling sensitivities between truster and trustee (standard case) as psychological games with incomplete information.

Attanasi and Nagel (2006) and Attanasi, Battigalli and Nagel (2007) remark that another source of incomplete information has to be taken into account. It is possible that psychological utilities of players sensitive to guilt and/or to reciprocity are not linear, differently from what we assume throughout this chapter. Finally, Attanasi and Nagel (2006) shows that repeated psychological games with incomplete information have a non-standard signaling game structure.

3. Review of the Experimental Literature on Psychological Games

In this section, we review and interpret some experimental papers giving support to theories of belief-dependent motivations. We mainly concentrate on social dilemma games belonging to the same family of the one presented in section 2. We divide those contributions into two subgroups: in section 3.1 we discuss the two main behavioral

theories aiming to explain the causal relations between beliefs and actions; in section 3.2 we present in deeper detail the main experimental studies analyzing the relations between emotions (expressed as belief-dependent motivations) and actions.

3.1. Relationship between Actions and Beliefs

Strategic settings in which players' private interest conflicts with their collective interest (like trust games) have been studied both by experimental economists and psychologists. One regularity among all these studies involves the relationship between individual's beliefs and actions: in particular, expectations of others' cooperation and one's actual cooperation are robustly and positively correlated. However, two competing classes of (causal) theories have been proposed to explain this relationship.

The first class of theories, coming from economics, suggests that beliefs cause actions. These theories are defined by Croson and Miller (2004) *reaction theories*, since individuals choose actions to *react* to beliefs.

The second class of theories, proposed by psychologist, suggest that actions cause beliefs. There is a number of different specific theories in this class: Croson and Miller (2004) provide a critical analysis of many of them. They conclude that, whichever of these theories one adopts, they all have the similar property that expectations are projected from own (anticipated) behavior. They refer to these theories as *projection theories*.

There are several studies (e.g. Messick *et al.* (1983), Schroeder *et al.* (1983) and Weimann (1994)) showing that manipulated expectations can affect choices, thus providing evidence to reaction theories. However, other studies support projection theories. Yamagishi and Sato (1986), for example, argue that choices are justified ex-post by adjusting expectations. Dawes *et al.* (1977) state that the difference between the variance of the players' estimates of others' actions and the observers' estimates of others' actions demonstrates 'false' consensus.

Almost all the experimental works on the relation between actions and beliefs in social dilemmas have assumed one theory or another in the experimental design (for example, psychologists tend to elicit beliefs *after* individuals make decisions in strategic settings, economists *before*). However, none of these previous experiments were designed to explicitly distinguish between these competing classes of theories. Starting from the fact that both classes of theories are consistent with the existing evidence of a positive relationship between beliefs and actions in social dilemmas, Croson and Miller (2004) design a particular social dilemma experiment in order to explicitly provide a clean

separation of these two classes of explanations: their results are consistent with the reaction hypothesis as opposed to the projection hypothesis. This is not to conclude that people don't project in dilemma settings, but that in settings where reaction and projection predict different outcomes, reaction has the stronger effect.

That would lead us to conclude that, introducing belief elicitation in an experimental setting, the influence of belief elicitation on subsequently chosen actions should be greater in size with respect to the one that actions could have on subsequent beliefs. However, the experimental literature surveyed in the next subsection, among other things, shows that in trust games (a particular case of social dilemma games) the influence of belief elicitation (and transmission) on subsequent chosen actions is small: different kinds of 'belief manipulation' do not affect players' behavior.

Nonetheless, our prediction is that reaction theories and projection theories should apply even more strongly to psychological games, because of a structural 'casual relation system': actions cause beliefs (*projection*); since beliefs are part of players' psychological utilities, when they update as the play unfolds, players' (total) utilities update; therefore, the optimal action is chosen according to the 'new' updated utilities. Hence, beliefs cause actions (*reaction*) also through the psychological utilities.

3.2. Relationship between Actions and Beliefs-dependent Motivations

In this subsection, we mainly concentrate on the relations between actions and feelings in *trust games*.[18] In the review below, we also analyze the techniques of belief elicitation and transmission used in the previous experimental settings to investigate players' sensitivity to some feelings. Belief elicitation is essential to measure the influence of belief-dependent motivations on players' behavior in psychological games with incomplete information. Nonetheless, feeling elicitation and transmission among players is essential to represent a psychological game with complete information in a laboratory. We first analyze the main belief elicitation techniques in the experimental literature. Then, we discuss some feeling elicitation and transmission techniques.

Dufwenberg and Gneezy (2000) analyze the relevance of *trust responsiveness*[19] in a

[18] In order to ensure uniformity while summarizing the different experiments, in each trust game described in this section, we call *truster* the player who has to choose to place trust (or not) and *trustee* the player who has to choose to fulfill trust (or not).

[19] *Trust responsiveness* (or the *self-fulfilling property of trust*) is 'a tendency to fulfil trust because you believe that it has been placed on you' (Bacharach, Guerra and Zizzo (2001), p. 1). Actually, Dufwenberg and Gneezy (2000) talk about '*let-down aversion*', which could be thought of as synonymous

simple trust game, in which the truster may take $x \in [0, 20]$ Dutch guilders, or leave them and let the trustee split 20 guilders between them. *After* the truster and the trustee have simultaneously made their choices (strategy method), there is beliefs elicitation both from truster (first-order beliefs) and from trustee (second-order beliefs), but there is no belief transmission. They conclude that trust responsiveness, which predicts that trustees who have higher guesses will be more likely to fulfil trust, should hold even if belief transmission did not take place.[20] This result is not in line with Croson (2000), which shows how belief elicitation distorts behavior in the context of social public goods and prisoner dilemma experiments.

Morrison and Rutstrom (2002) explore the possibility that psychological games are associated with the relationship between prior beliefs about others' behavior and the actual revealed actions of those individuals. In other words, they investigate whether a surprise or a confirmation of prior beliefs have 'payoff' consequences. The authors develop and implement an experimental design that allows for belief elicitation in an investment game, where material self-interest would predict the absence of either trust or reciprocity. They elicit truster's first-order beliefs *before* he chooses and trustee's second-order beliefs *before* she learns about the decision made by her partner. Their results, supporting existing evidence that rejects the Nash equilibrium in monetary payoffs, are consistent with Rabin (1993), given that most trustees engage in reciprocity as a way of repaying kindness by the trusters. They also support somehow Fehr and Schmidt's (1999) inequality aversion model. However the strongly significant correlation between beliefs and amounts sent back (by the trustees) cannot be explained by the inequality aversion model.

Bacharach, Guerra and Zizzo (2001) test trust responsiveness in basic 2×2 trust games with different payoff structures. In their experimental setting, beliefs are elicited in an incentive-compatible way on the part both of trusters and of trustees, using the same technique of Dufwenberg and Gneezy (2000), but (a descriptive statistics about)

to the subsequently adopted term of '*guilt aversion*'. Comparing Bacharach, Guerra and Zizzo (2001) and Guerra and Zizzo (2004) with Dufwenberg and Gneezy (2000) and Charness and Dufwenberg (2005), it seems clear that *trust responsiveness* and *guilt aversion* are two sides of the same coin: they describe the same feeling, the former from a positive point of view and the latter from the negative one. Moreover, they both predict the same behavior for the trustee: positive correlation between her second-order beliefs (as statement of truster's confidence that she would fulfil trust) and trust fulfilment. However, the term '*trust responsiveness*' seems too much 'context-dependent', whereas the term '*guilt aversion*' looks more 'general', being a statement about a motivation that can be extended also to games outside the family of trust ones.

[20] Dufwenberg and Gneezy (2000) do not only look at trust games but also at dictator games. They report evidence in favor of *let-down aversion* for dictator games too.

trusters' first-order beliefs are transmitted to each trustee, *before* eliciting her second-order beliefs.[21] This is common knowledge among all participants. We think that because of beliefs (elicitation and) transmission from trusters to trustees, Bacharach, Guerra and Zizzo's (2001) results may have not been very general. Belief transmission is motivated in the paper by the desire to enable trustees to form definite enough beliefs about the trusters' confidence on them. While this is reasonable, this may have biased trust responsiveness upwards for at least two reasons: first of all, it may draw the attention of subjects to the truster's confidence in them, possibly making any psychological factor leading to trust responsiveness more salient than otherwise; secondly, it is not obvious that in real world people typically receive this kind of information. Guerra and Zizzo (2004) use two simple trust games to measure directly or indirectly the robustness of trust responsiveness in three experimental treatments: beliefs are not elicited; beliefs are *elicited* but not transmitted; beliefs are *elicited* and *transmitted* from trusters to trustees, as in Bacharach, Guerra and Zizzo (2001). Since trusting and fulfilling rates are quite similar in all three treatments, they conclude that trust responsiveness is robust with respect to any kind of 'belief manipulation', hence strengthening the case for the real-world significance of trust responsiveness. Differently from Bacharach, Guerra and Zizzo (2001) and Guerra and Zizzo (2004), we do not agree with the 'neutrality' of their beliefs (elicitation and) transmission procedure on players' behavior, while we completely agree with the 'neutrality' of the beliefs elicitation alone.

In this review of belief-elicitation (and transmission) experimental works, we mainly concentrated on 2×2 trust games. However, there are some other experimental papers on belief-elicitation in other kinds of (social dilemma) games, which states interesting rules aiming to provide the right incentive for subjects in the lab to reveal their true first-order beliefs: Nyarko and Schotter (2002), who test behavior in two-player 2×2 constant-sum games with a unique mixed equilibrium; Costa-Gomes and Weizsäcker (2006), who test behavior in two-player 3x3 games with a unique equilibrium in pure strategies; Rey Biel (2006), who tests behavior in two-player 3x3 constant-sum games with a unique equilibrium in pure strategies and with different number of rounds of iterated deletion of (strictly) dominated strategies required to reach the Nash equilibrium.

Let us now examine the role of *pre-play communication* as a technique of *feeling* (*elicitation* and) *transmission*, able to reproduce psychological games with *complete information* in the lab.

A number of experiments (e.g., Dawes *et al.*, 1977) provide evidence that face-to-

[21] Each trustee receives a report of the mean value of first-order beliefs of her non-coplayers.

face communication can greatly improve cooperation, even when the dominant strategy (with selfish preferences) is 'defection'. However, as Roth (1995) points out, there may be many confusing and uncontrolled effects in a face-to-face contact.

Clark *et al.* (2001) show that non-binding pre-play communication moves outcome in the direction of the Pareto-dominant equilibrium in games which converge to a Pareto-dominated equilibrium in the absence of communication. They first reproduce the experimental findings of Cooper *et al.* (1990) suggesting an important role for pre-play communication for the coordination on the Pareto-dominant equilibrium in a simple two-player coordination game. Then they go on and study other games in which coordination failures take place in the absence of communication, in order to evaluate the argument in Aumann (1990) according to which the effectiveness of communication is sensitive to the payoff structure. The authors find that when players have a strict preference over the opponent's choices, communication still influences outcomes, but its impact is considerably lessened. In line with Aumann's argument, they also find that agreements to play Nash equilibrium are fragile. Finally, they find that when communication is effective, it does not always work in the direction one might expect, i.e. it does not necessarily lead to the Pareto-dominant equilibrium.

Charness and Dufwenberg (2006) run one-shot interactions between participants in a simple experimental trust game (not so different from the one presented in section 2)[22] in order to test for *guilt aversion*. The authors examine the effect of non-binding pre-play communication (cheap talk from trustee to truster) on players' trust and cooperation. Their experimental results suggest that there are differences in the level of efficiency among the various forms of communication. Beliefs are higher when promises are made. It seems that statements of intent are instrumental in changing perceptions and facts about what people do. As the authors admit, measuring beliefs is crucial to the guilt-aversion story and also to the reciprocity model. *After* collecting the strategic decisions made by participants (they use the strategy method), the authors elicit truster's first-order and trustee's second-order beliefs. As the game is one-shot and beliefs are transmitted only after strategies are chosen, the belief elicitation should not have any impact on participants' prior choices. The authors do not allow face-to-face interaction, but instead allow the trustee to send a *free-form* message to the truster. This permits a clean, controlled test of the function of communication: they only permit a single message from one party to the other. Moreover, as they are interested in whether some

[22] The simple one-shot trust game they represent in the lab differs from the one in section 2, because, after the action profile $(Continue, Share)$ they allow for a random draw, introduced in order to capture the essence of 'hidden action'.

particular type of message has an impact or not, they give the sender a blank piece of paper to write any (anonymous) message, instead of restricting the message space.

Masclet *et al.* (2003) test *informal sanctions* contribution in public good games. Each agent can express disapproval on others' contribution decision thus showing how this non-monetary system is able (as much as monetary sanctions) to increase contribution levels. Moreover, they stress the fact that punishment may be a particular form of communication and, in a repeated game, it serves as a form of pre-play communication for future periods.

Starting from Charness and Dufwenberg's (2006) intuition on the importance of pre-play communication in psychological games and from Masclet *et al.* (2003) result of a non-monetary punishment scheme able to create 'psychological pressure', Attanasi, Battigalli and Nagel (2007) build a simple mechanism to elicit and transmit agents' sensitivity to some particular feelings in a psychological game of trust similar to the one analyzed in section 2. Differently from Charness and Dufwenberg (2006), who introduce a form of 'free' pre-play communication in which the truster can send any kind of signal (maybe having no relations with social preferences) to the trustee, Attanasi, Battigalli and Nagel (2007) introduce a well-structured feelings' elicitation scheme which enables to elicit trustee's sensitivity to both distribution-dependent preferences *à la* Fehr and Schmidt (1999) and to belief-dependent motivations *à la* GPS. In particular, they design a reliable mechanism of elicitation and transmission (from trustee to truster) of a good approximation of trustee's feelings sensitivities (to guilt, reciprocity or both). Hence, when this mechanism is in place it is like the two parameters θ_B^g and θ_B^r in Figures 2 and 3 are both *public information* between the two players. In that way, they are able to compare players' behavior in the 'almost complete information treatment' with the one in the 'incomplete information treatment', i.e. when they play the (one-stage) game of trust without any kind of ex-ante belief-dependent motivations transmission. Their experimental results show that eliciting B's psychological preferences without transmitting them to A does not modify players behavior. This could be interpreted as an indirect proof of the 'existence' of the psychological preferences: the authors do not induce them, they only elicit something that was already in place. Moreover, their experimental results show that eliciting and transmitting B's psychological preferences and thus letting the two players playing the (almost) *complete information* one-stage psychological game of trust leads them to behave in a visibly different manner (with respect to their behavior in the corresponding *incomplete information* psychological game setting). More precisely, the *public information* of the trustee's psychological preferences in the one-stage psy-

chological game results in truster's perception of trustee's belief-dependent motivations (like guilt aversion and/or reciprocity) which would be otherwise underestimated. That in turn ends in a more cooperative behavior for both players. The authors provide also theoretical findings on psychological games with complete and incomplete information that are well matched by their experimental results.

Using the same feeling elicitation and transmission procedure, Attanasi and Nagel (2006) test players' behavior in a *finitely repeated game*, in which they elicit beliefs at the beginning of each period. This is the first experimental work allowing for reputation or repeat-game effects in psychological games too. The authors are able to disentangle between the (standard) effect of reputation and the (non-standard) effect of public information of trustee's feeling sensitivity on players' behavior. They also provide specific *theoretical findings* for repeated psychological games. These games present several features that one cannot find in framing and solving other kinds of dynamic psychological games. They deeply investigate such features. Moreover, by concentrating on repeated psychological games, they both define specific intertemporal psychological utility functions and introduce two (behaviorally derived) *monotonicity conditions* on the dynamics of players' beliefs. These conditions allow them to select among the multiplicity of sequential psychological equilibria coming out by applying Battigalli and Dufwenberg (2005) and Dufwenberg and Kirchsteiger (2004) solution concepts for psychological games with *guilt aversion* and with *reciprocity*, respectively. The subset of equilibria satisfying their monotonicity conditions matches very well the behavior of participants in their experimental sessions.

4. Conclusions

In this chapter, we provided the main theoretical insights of the general psychological game framework, through an example of a simple trust game in which belief-dependent motivations are involved.

From a theoretical point of view, our conclusion is that there is little reason to assume that equilibrium coordination is easier in psychological games than in standard games. In fact, since psychological games often seem more complicated, and since problems of equilibrium multiplicity are likely to be enhanced in psychological games (as shown in section 2), assuming equilibrium may be assuming too much, especially in psychological games. Another reason to feel skeptical about a fully fledged equilibrium analysis in psychological games is the following: it is often argued that players learn to play Nash

equilibrium because, through recurrent strategic interaction, they come to hold correct beliefs about the actions of the opponents. This is not enough for a psychological equilibrium: since utilities depend on hierarchical beliefs, players would have to be able to learn the beliefs of others, but, unlike actions, beliefs are typically not observable ex post. GPS, the first paper on psychological games, restrict attention to equilibrium analysis, but in many strategic situations there is little compelling reason to expect players to coordinate on an equilibrium, hence one may wish to explore alternative assumptions. However, giving up the equilibrium assumption does not necessarily mean giving up on predictive power. Following this direction, BD develop a framework for analyzing interactive epistemology in psychological games, without postulating equilibrium play. In particular, building on an epistemic theme due to Battigalli and Siniscalchi (2002), they extend Pearce's (1984) classical notion of (extensive form) rationalizability to psychological games. The concept captures psychological forward induction in simple games like those in Figure 1 and Figure 2 (section 2), and in more complicated games for which long chains of beliefs about beliefs may be needed to get clear-cut predictions.

In the second part of the chapter, we discuss the main experimental works that directly or indirectly refer to psychological game-theoretic explanations as able to rationalize their experimental results. In the last two decades, many experimental economists have studied the relation between actions and feelings in social dilemma games, allowing for *ex-post explanations* claiming for belief-dependent motivations as source of the 'irregularities' in the experimental results. Attanasi, Battigalli and Nagel (2007) are the first to provide a direct test for psychological games in the lab, clearly distinguishing between psychological games with complete and incomplete information. Using a similar feelings elicitation and transmission method, Attanasi and Nagel (2006) extend the analysis of the relationships among actions, beliefs and feelings to *dynamic* (repeated) settings: they state the evolution of beliefs, their correlation with the played action profile and the way through which they incorporate players' belief-dependent motivations. These works are only a starting point, although they are indicative of the viability of that line of experimental research on psychological games.

References

[1] Attanasi, G., P. Battigalli and R. Nagel (2007) 'Public Information of Psychological Preferences in One-Stage Trust Games: an Experimental Study', mimeo.

[2] Attanasi, G. and R. Nagel (2006) 'Actions, Beliefs and Feelings: an Experimental Study on Dynamic Psychological Games', mimeo.

[3] Aumann, R. (1990) 'Nash Equilibria are not Self-Enforcing', in J. Gabsiewicz, J. F. Richard and G. Wolsey (eds), *Economic Decision Making Games*, *Econometrics and Optimisation*, Amsterdam: Elsevier, pp. 201-6.

[4] Bacharach, M., G. Guerra and D. J. Zizzo (2001) 'Is Trust Self-Fulfilling? An Experimental study', mimeo.

[5] Battigalli, P. (2007) 'Reciprocity and Psychological Games', presented at the international conference 'Reciprocity: Theory and Facts' - Verbania (Italy), publicly available at http://www.igier.uni-bocconi.it/battigalli

[6] Battigalli, P. and M. Dufwenberg (2007) 'Guilt in Games', *American Economic Review*, Papers and Proceedings, 97, 170-176.

[7] Battigalli, P. and M. Dufwenberg (2007) 'Dynamic Psychological Games', mimeo, August 2005 (previous version: IGIER working paper 287).

[8] Battigalli, P. and M. Siniscalchi (1999) 'Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games', *Journal of Economic Theory*, 88, 188-230.

[9] Battigalli, P. and M. Siniscalchi (2002) 'Strong Belief and Forward Induction Reasoning', *Journal of Economic Theory*, **106**, 356-391.

[10] Charness, G. and M. Dufwenberg (2006) 'Promises and Partnership', *Econometrica*, 74, 1579-1601.

[11] Clark, K., S. Kay and M. Sefton (2001) 'When are Nash equilibria self-enforcing? An experimental analysis', *International Journal of Game Theory*, 29, 495-515.

[12] Costa-Gomes, M. and G. Weizsäcker (2006) 'Stated Beliefs and Play in Normal-Form Games', mimeo.

[13] Croson, R. T. A. (2000) 'Thinking like a game theorist: factors affecting the frequency of equilibrium play', *Journal of Economic Behavior and Organization* 41, 299–314.

[14] Croson, R. T. A. and Miller, M. (2004) 'Explaining the Relationship between Actions and Beliefs: Projection vs. Reaction', mimeo.

[15] Dawes, R., Mctavish, J., and Shacklee, H. (1977) 'Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation', *Journal of Personality and Social Psychology*, 35, 1-11.

[16] Dawes, R. (1989) 'Statistical Criteria for Establishing a Truly False Consensus Effect', *Journal of Experimental Social Psychology*, 25, 1-17.

[17] Dawes, R. and Mulford, M. (1996) 'The False Consensus Effect and Overconfidence: Flaws in Judgment or Flaws in How We study Judgment?', *Organizational Behavior and Human Decision Processes*, 65, 201-211.

[18] Dufwenberg, M. (1995) 'Time-Consistent Wedlock with Endogenous Trust', in On Rationality and Belief Formation in Games, Doctoral Dissertation, Department of Economics, Uppsala University.

[19] Dufwenberg, M. (2006) 'Psychological Games', entry for *The New Palgrave Dictionary of Economics* (2nd edition).

[20] Dufwenberg, M. (2002) 'Marital Investment, Time Consistency and Emotions', *Journal of Economic Behavior and Organization*, 48, 57-69.

[21] Dufwenberg, M. and U. Gneezy (2000) 'Measuring beliefs in an experimental lost wallet game', *Games and Economic Behavior*, 30, 163-182.

[22] Dufwenberg, M. and G. Kirchsteiger (2004) 'A Theory of Sequential Reciprocity', *Games and Economic Behavior,* 47, 268-298.

[23] Elster, J. (1998) 'Emotions and Economic Theory', *Journal of Economic Literature*, 36, 4774.

[24] Falk, A. and U. Fischbacher (2006) 'A Theory of Reciprocity', *Games and Economic Behavior*, 54, 293-315.

[25] Fehr, E. and S. Gächter (2000) 'Fairness and Retaliation: The Economics of Reciprocity', *Journal of Economic Perspectives*, 14, 159-181.

[26] Fehr, E. and K. M. Schmidt (1999) 'A Theory of Fairness, Competition, and Cooperation', *Quarterly Journal of Economics*, 114, 817-868.

[27] Geanakoplos, J., D. Pearce and E. Stacchetti (1989) 'Psychological Games and Sequential Rationality', *Games and Economic Behavior*, 1, 60-79.

[28] Geanakoplos, J. (1996) 'The Hangman Paradox and the Newcomb's Paradox as Psychological Games', Cowles Foundation Discussion Paper No. 1128.

[29] Gilboa, I. and D. Schmeidler (1988) 'Information Dependent Games: Can Common Sense Be Common Knowledge?', *Economics Letters,* 27, 215-221.

[30] Guerra, G. and D. J. Zizzo (2004) 'Trust Responsiveness and Beliefs', *Journal of Economic Behavior and Organization*, 55, 25-30.

[31] Kolpin, V. (1992) 'Equilibrium Refinements in Psychological Games', *Games and Economic Behavior*, 4, 218-231.

[32] Kreps D. M. and R. Wilson (1982) 'Sequential Equilibria', *Econometrica*, 50, 863-894.

[33] Krohne, H. W. (2003) 'Individual differences in emotional reactions and coping', in R. J. Davidson, K. R. Scherer and H. H. Goldsmith eds., *Handbook of Affective Sciences*, 698-725. New York: Oxford University Press.

[34] Masclet, D., C. Noussair, S. Tucker and M. C. Villeval (2003) 'Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism', *American Economic Review*, 93, 366-80.

[35] Messe, L., and Sivacek, J. (1979) 'Predictions of Others' Responses in a Mixed-Motive Game: Self-Justification or False Consensus?', *Journal of Personality and Social Psychology*, 37, 602-607.

[36] Messick, D., Wilke, H., Brewer, M., Kramer, R., Zemke, P., and Lui, L. (1983) 'Individual adaptations and structural change as solutions to social dilemmas', *Journal of Personality and Social Psychology*, 44, 294-309.

[37] Morrison, W. and E. E. Rutström (2002) 'The role of beliefs in an Investment Game Experiment', mimeo.

[38] Nyarko, Y. and A. Schotter (2002) 'An Experimental Study of Belief Learning Using Elicited Beliefs', *Econometrica*, 70, 971-1005.

[39] Penta, A. (2004) 'Psychological Games, Interactive Epistemology, and Reciprocity', undergraduate dissertation, Bocconi University.

[40] Rabin, M. (1993) 'Incorporating Fairness into Game Theory and Economics', *American Economic Review*, 83, 1281-1302.

[41] Rey Biel, P. (2006) 'Equilibrium Play and Best Response to (Stated) Beliefs in Constant Sum Games,' mimeo.

[42] Roth, A. (1995) 'Bargaining Experiments,' *Handbook of Experimental Economics*, John Kagel and Alvin Roth eds., 253-348, Princeton University Press.

[43] Schroeder, D., Jensen, T., Reed, A., Sullivan, D., and Schwab, M. (1983) 'The Actions of Others as Determinants of Behavior in Social Trap Situations,' *Journal of Experimental Social Psychology*, 19, 522-539.

[44] Tangney, J.P. (1995) 'Recent Advances in the Empirical study of Shame and Guilt', *American Behavioral Scientist*, 38, 1132-1145.

[45] Weimann, J. (1994) 'Individual Behavior in a Free Riding Experiment', *Journal of Public Economics*, 54, 185-200.