Intentions, Trust and Frames:

A note on Sociality and the Theory of Games

Vittorio Pelligra

University of Cagliari and CRENoS

---

1 I thank Pierpaolo Battigalli, Luigino Bruni, John Davis, Shaun Hargreaves-Heap, Maurizio Pugno and Alessandra Smerilli who have read and commented on a previous version of the paper as well as the Editor and two anonymous referees for their useful suggestions. The usual disclaimer applies.
Abstract:
Psychological Game Theory (PGT) expands classical game theory allowing for the formal analysis of belief-dependent sentiments and emotions such as resentment, pride, shame, gratefulness, and the like. PGT incorporates these factors by relating agents' subjective expected utility to players' strategies, to their beliefs about others' strategies, but also to their beliefs about others' beliefs about their strategies, and so on. This paper argues that, thanks to the epistemic implications of this hierarchy of beliefs, PGT is well-endowed to address, and to some extent solve three of the most challenging problems recently emerged in classical game theory related respectively to the role of players' intentionality, to the self-fulfilling nature of trust and finally to the so-called social framing effects.

*Keywords*: Psychological games, intentions, trust, decision frames.

*JEL Classification*: C72, C79, C9.
“[In economics] we are only concerned with the ‘what’ and the ‘how’, and not at all with the ‘why’”

(Wicksteed, 1933[1910]:165).

1. Introduction

Philosophy of action and modern neurosciences agree with the fact that as social actors, but most importantly as human beings, we invariably perceive other agents’ behaviour as directed at certain targets. We operate in our social environments as “intentionality-detectors” (Metzinger & Gallese, 2003). Economic theory, however, tends to sterilize the impact of others’ behaviour and their intentions by focusing on market behaviour where the direct impact of behavioral interdependence is minimised. Two important exceptions to this neoclassical approach are the so-called new social economics\(^2\) and game theory.

In the early ‘40s the theory of games originally developed out of the dissatisfaction for the limitations of neoclassical theorizing about social interdependency and agents’ intertwinedness. Von Neumann and Morgenstern (1944) pointed out that most of the social situations in the economic domain are characterised, literally and not only metaphorically, by the same structure of a game. The most important ability to play a game is that of

\(^2\) In this paper, I will not consider the former approach. See Glaeser, Sacerdote & Scheinkman (1996, 2003); Glaeser & Scheinkman (2001, 2003), Gui & Sugden, (2005), and Scheinkman (2008) for a review.
accurately predicting other players’ behaviour. Consequently this ability should be at the core of any theory of strategic interactions, such as game theory aims to be. However, it is surprising and even disappointing to realise retrospectively that classical game theory after 50 years of developments, still basically assumes instead of explaining that agents are capable to predict others’ behaviour.

A promising step forward to fill this gap is constituted by Psychological Game Theory (PGT) (Geneakoplos, Pearce & Stacchetti, 1989; Battigalli & Dufwemberg, 2009), which develops analytical tools to formalise this kind of social abilities. This is done, on the one hand, by letting players’ payoffs depend not only on players’ actions, as it is in the classical theory, but also explicitly on players’ intentions, beliefs and emotions, and on the other, by providing solution concepts that take into account this complex causality.

The main thesis of the paper is that, thanks to the epistemic features of these analytical tools, PGT is well suited to address and, to some extent solve, three of the most challenging problems emerged in the domain of the classical game theory in the latest years: the first issue, perhaps the least explored of the three, refers to the problem of players’ intentionality, the second, to the problem of trust (and to that related of payoff’s endogenisation), and the third, to the so-called framing effect.

The remainder of the paper is organised as follows: section 2 discusses the concepts of “other” and “otherness” as conceived by classical game theory, in a historical-methodological perspective. Section 3 introduces the problem of intentionality by presenting a few experimental results that challenge the assumption of consequentialism. Section 4 outlines the very basic features of PGT linking it to the problem of intentionality.
in philosophy and neurosciences, and discussing the way PGT formalises the mind-reading process. Sections 5 focuses on the problem of trust and payoff endogenisation. Section 6 analyses the effects of decision frames for strategic decision-making and how PGT can take these effects into account. Section 7 ends the paper with some concluding remarks.

2. Game Theory, the others and their “otherness”.

Before addressing the three issues mentioned above, intentions, trust and frames, let me briefly depict the background where they are situated. The background here refers to the way classical game theory tackles the social dimension of interpersonal relationships; how, in particular, it describes and models the role of others and of their “otherness” in strategic interactions. At the end, a picture will emerge of a theory that oversimplify, if not elude, the “otherness” of social agents, and its implications. This is a consequence, I suppose, of a more general attitude of economists towards individuals. As John Davis convincingly argues, in fact: “in economics very little attention is given to the theory of the individual (...) Economics takes the individual as given” (2003:1).

The starting point of this story is, what else, Von Neumann-Morgenstern’s game theory: although the concept of interdependence lies at the core of their theory, the idea of interactive rational behaviour, as embodied in the minimax criterion, derives directly from their idea of individual rational behaviour. In other words, it seems as if it were defined independently of other players’ behaviour. A given course of action, in fact, is viewed as rational, if it maximises the minimum payoff one can get whatever the other
players do. This independence from others’ decisions can be read as a consequence of Von Neumann-Morgenstern’s desire to expunge any reference to the psychological dimension of the agents from their theory. Von Neumann’s 1928 original project is coherently embedded in this historical line, and its heritage strongly oriented his joint work with Morgenstern towards a particular characterisation of rational behaviour “capable of setting the players free from the necessity to form an expectation about the rivals’ actions and thoughts” (Giocoli, 2003:282), and of social interactions where all the details are expunged except for the mathematical properties of the payoffs.

Describing Von Neumann-Morgenstern’s formal world, Thomas Schelling points out two main weaknesses:

i) interdependence reduced to independence: a player “does not need to communicate with his opponent, he does not even need to know who the opponent is or whether there is one. (...) with a minimax criterion, a zero-sum game is reduced to a completely unilateral affair” (1960:105, emphasis added);

ii) de-psychologization of the prediction process of others’ behavior: “A randomised strategy is a deliberate means of destroying any possibility of communication, especially communication of intentions” (Ivi).

In the next and somewhat alternative step in the elaboration of a complete theory of games, John Nash (1950, 1951) develops a solution concept that embodies a rather different consideration of others’ role in determining one’s own optimal strategies. A set of strategies leads to a Nash equilibrium when they represent optimal replies to what
the player assumes other players would do. Therefore, when choosing a certain action, each player must have in mind which action the other players are going to choose. Each player forms conjectures about other players’ behaviour and they also know that the other players form similar conjectures as well. Notice, however, that these conjectures are always inspired by the assumption that all the players will act to maximize their payoff. The convergence to the equilibrium point, which must satisfy the requirement of conjectures’ (or beliefs’) mutual consistency, is in the end assured by the two assumptions of optimising behaviour and common knowledge of optimising conduct. Thus, in Nash’s theory agents’ intentions are assumed to be limited to the goal of payoffs maximisations. To appreciate the restrictiveness of these assumptions, one may note, for example, that the theory implies the coordination of expectations, that is, that two players’ beliefs about how a third player would play the game must necessarily be the same.

Players form unmodifiable strategies before the game begins and cannot engage in counterfactual reasoning, which would be needed to answer questions like: “what would I do if you did not conform to my assumption of optimising behaviour?” In this case the only thing to do is to assume that the other is behaving irrationally and thus suspending the game. That conclusion has being interpreted as an implication of the distinctive solipsistic nature of Nash’s theory: “Our theory (...) is based on the absence of coalitions in that it is assumed each participants act independently, without collaboration or communication with any of the others” (Nash, 1996:22). Commenting on this point Philip Mirowski emphasises that: “what would mean to play a game without any open
Bayesian game theory offers only a slightly different attitude. It provides a framework to analyse games of incomplete information, in which there is uncertainty, for instance, towards players' incentives or towards the history of the play. In a Bayesian game each agent’s private information is conceptualized assuming the existence of different types of players, that, as well as the structure of the game, are assumed to be common knowledge (Harsanyi, 1967-68). Players form and revise beliefs about which type is matched with. These beliefs can be considered as conjectures about players' types or players' incentives and not about their motives. Let’s consider a game in which A and B do not know each other’s payoff functions – suppose, for simplicity, that they do know their own payoff functions. This game can be modelled at least in two different ways. First: player A knows that player B’s strategy will depend on player A’s own payoff function. Thus, before choosing her move, A will form some expectation about B’s payoff. At the same time, B will form some expectation about A’s payoff function. Once these first-order expectations are established, A will form some second-order expectations about B’s first-order expectations, whereas B will form some second-order expectations about A’s first-order expectations, and so on. Notice that in the Bayesian approach these expectations are nothing but subjective probability distributions over the relevant mathematical objects. Of course, any model based on higher and higher order expectations would be much more complicated in the case of n-person games (with n> 2). In the “Harsany framework”, according to the words of its own champion, this way of
model players’ behaviour would be - “very natural - yet (...) rather impractical” (Harsany, 1994:137). Besides, the use of subjective probability distributions of various orders poses many technical difficulties. Although these difficulties can be somewhat overcome (Aumann, 1963, 1964), the above procedure to model higher and higher-order subjective probability distributions remains, in Harsany’s view – “a hopelessly cumbersome model for analysis of [incomplete information] games” (1994:150).

The second way to model incomplete information games is to transform them into imperfect information games by letting an external random device (Nature) determine the type of players (the active players) that are actually interacting in a given game. Each player tries to estimate the probability associated to each Nature’s move which, in turn, depends on what Harsany calls “the relevant social forces”. They will, in fact, try to estimate these probabilities as an outside observer would do, one restricted to information common to both players (cf. Harsanyi, 1967–68:176). Besides, each player knows that the other players will estimate these probabilities in the same way as they do (see Fudenberg & Tirole, 1991:210). An alternative interpretation of this common prior assumption considers players behaving as if both of them know the true numerical values of these probabilities. Thus, in Bayesian games each outcome is associated with a unique plan, a unique strategy, a unique set of intentions and, most importantly, a unique and common prior distribution of beliefs over types, beliefs that are common and imported from outside the game.

Thus, in Bayesian game theory, a game of incomplete information that can be used to describe players’ heterogeneity with regard to the set of conceivable intentions
underlying each moves, can be modelled either in a natural but, at best, impractical way, or in a practical yet very unintuitive and unrealistic way. In particular, such unrealism constitutes a serious limitation whenever we want to describe and analyse all those many social situations in which people’s motivations may be affected by belief-dependent emotions such as, inter alia, “anger, hatred, guilt, shame, pride, admiration, regret, rejoicing, disappointment, elation, fear, hope, joy, envy, malice, indignation, jealousy, surprise, boredom, sexual desire, enjoyment, worry and frustration” (Battigalli & Dufwemberg, 2005:41), that may be best modeled by considering internal or endogenous beliefs.

Summing up, the account of “the other” that classical game theory provides is based on four simple, though not unproblematic, assumptions: individual rationality, common knowledge of rationality, common priors and consequentialism. My claim is that the implications of these assumptions give us a far too simplistic picture of players’ “otherness”. This is true, in particular with regard to the important aspect of intentionality, especially because the prediction about the other’s intentions can be only grounded on his/her preferences about consequences. This, in turn, determines many of game theory’s problems in the explanation and prediction of real people’s behavior. In this area of research, game theorists may benefit from contributions coming from different fields such as philosophy of action and neurosciences\(^3\).

\(^3\) Efforts in this direction are attested by the work of Bacharach (1999), Singer & Fehr (2005), Ross (2005), Pugno (2005) and Camerer et al. (2005)
3. Does intention matter?

According to the account of intentionality that we have described in the previous section, in classical game theory players are exclusively motivated by what is in the payoff. This is basically what the assumption of consequentialism prescribes: players order their preferences over actions according to their preferences over the consequences these actions lead to. If action $a$ produces outcome $\alpha$, action $b$ produces outcome $\beta$ and action $c$ yields to outcome $\gamma$; action $a$ is preferred to $b$ and to $c$, as long as outcome $\alpha$ is preferred to $\beta$ and the latter to $\gamma$. Robust and growing experimental evidence, however, shows that, contrary to the consequentialist view, the same outcome may be variously assessed and acted upon, depending on the history of the moves that lead to it. That means that when deciding on how to behave in a strategic situation, real people take into account not only the prospective outcomes of their joint actions, but also other backward-looking factors, such as bygones and counterfactual arguments. Some examples may help to illustrate this point.

The first is drawn from McCabe, Smith and Rigdon (2003). They consider players' behavior in two variants of the trust game. In the “Voluntary” trust game (VTG, figure 1) player A can move “right” getting $20 both for him/her and player B. Otherwise he/she can play “down”. In this case player B is called upon to move. He/she can opt for playing “right”, leading to a $25-$25 distribution, or “down”, grabbing $30 for him/herself and giving $15 to player A. The “Involuntary” trust game (ITG, figure 2) is similar to the “voluntary” version, apart from the fact that now there is not outside option. The only
active player is player B. Classical game theory predicts that players B should prefer moving “down” in both games. On the other hand, outcome-based theories of social preferences that extend the utility function in order to account for altruism (Margolis, 2008) or inequality aversion (Fehr and Schmidt, 1999), lead to different predictions: some players may prefer “right” in both games and the number of these moves should be approximatively equal both in the VTG and in the IVG.

[Figure 1 and 2 about here]

However, McCabe and colleagues found a more complex behavioral pattern: not only many B players opt for the cooperative move (“right”), but the proportion is much higher in the VTG (94.11%) than in the ITG (40.74%).

They interpret this result as showing both that concerns about the distributive consequences of actions are only one of the components in the motivational structure of players and that yet another aspect, namely intention-detection, matter for strategic behavior. That is why nor standard subgame Nash equilibrium, neither purely outcome-based social preferences are consistent with the observed behavior. Despite the fact that in both games the consequences (the payoff distributions at terminal nodes) are the same, players B tend to ascribe a kind or trustful intention to A's playing “down” in the VTG, and respond to that kindness with a higher frequency of cooperative moves (B plays “Right”). In the ITG such an intention cannot be communicated because of the absence of the outside option. “By eliminating Player A’s opportunity cost associated
with playing down, we have restricted Player B’s ability to read her counterpart’s intentions unambiguously. In the voluntary trust game, an intentional move down the tree by Player A can be explained by Player B as an act of trust. In the involuntary game, however, a down move carries no such information because Player A had no choice but to move down in the game” (2003: 273).

The second example refers to an experiment carried-out by Falk, Fehr and Fischbacher (2003). Their design consists of four different variants of the so-called mini-ultimatum game. In the mini-ultimatum game there are two players, a proposer (P) and a respondent (R). P is endowed with 10 experimental points, and can choose to offer R either 2 or X points (keeping either 8 or 10 – X for himself). R, in turn, can either accept or reject the offer. If she accepts, the payoff distribution resulting from P’s offer is implemented. If she rejects, both players get nothing. In the four variants considered by Falk et al., X takes the values 5, 8, 2, and 0. Deriving predictions for these games is relatively straightforward. Classical game theory predicts that the [8, 2] offer (i.e., 8 to P and 2 to R), should be accepted regardless of the alternative, because any positive amount is better than nothing. Models based on altruism (Margolis, 1982) or inequality aversion, such as Fehr and Schmidt’s (1999), predict a positive rejection rate of [8, 2] in all games and, crucially, the same rejection rate in all variants, as the inequality associated with the [8, 2] outcome is constant.

The results of the experiment show that the percentage of subjects who reject the [8, 2] offer is 44.4%, 26.7%, 18% and 8.9% when X takes the values 5, 8, 2 and 0 respectively. They interpret this evidence as suggesting that individuals care not only about the final
outcome resulting from a particular combination of strategies (as indicated by the positive rejection rate of the [8, 2] offer when the alternative is identical, i.e. X = 2), but also about intentions, as signalled by the alternatives that have not been chosen. The highest rejection level of the [8, 2] split when the alternative is [5, 5], can be explained as a punishment for an unkind intention (an equitable alternative is available but not chosen). A slightly lower rejection rate can be expected when the alternative is [2, 8], because although this is better than [8, 2] from R’s point of view, he/she cannot infer a fully unfair intention from P’s refusal to choose a distribution that is disadvantageous for him/herself. These arguments do not hold when the alternative is [8, 2], because P has in fact no choice in this case. Finally, when the alternative is [10, 0], by choosing [8, 2], P may be signalling kind intentions, leading to expect an even lower rejection rate.

In our third example, Fehr and List (2004) observe the behavior of a sample of Chief Executive Officers (CEOs) in various forms of investment games. In particular they considered two variants of the game, one where the first player is endowed with a certain amount of money and has to decide which part of the endowment, if any, to send to player 2. If she sends a positive amount, that is tripled and given to player 2, who can, in turn, decide how much, if any, to send back. The second variant of the game is similar apart from the fact that the proposer has to decide whether to implement a sanction, if what she receives back from player 2 is less that what she had expected. The game theoretical solution to both games is for player 1 to send nothing to player 2, and for player 2 to send back nothing to player 1. However, what has been observed is not only that most of the experimental subjects decided to invest and to send back substantial
amounts of money but, most strikingly, that such amount increases in the treatment when the sanctions are available even if, actually, not used. This result is interpreted by Fehr and List as a sign that: “the availability of the sanctioning threat can be quite productive (...) If principals voluntarily refrain from using the punishment threat when it is available, agents exhibit significantly more trustworthiness than if the punishment threat is not available. Thus, if agents face no punishment threat, the mere fact the principal could have used the punishment option affects the agent’s trustworthiness in a positive manner” (2004, p.746).

What emerges from the results of these experiments is that for real people bygones are as relevant as the choices actually made. Subjects seem to asses and react in different ways to the same material payoff distribution when this is brought about by different courses of actions, by different combinations of moves. This evidence sheds light on the differences that emerge when individuals act on the basis of information predicting others’, and when they act on the basis of information about mental causes of other individuals’ behaviour. The rationale behind this “anomalous” behaviour may be the fact that different combinations of actions, apart from the outcome they produce, signal different messages about the players’ intentions. When they try to read each other’s minds subjects infer others’ intentions by what they did as well as by what they could have done and did not. Bygones help us to read other players’ minds, to infer their

---

intentions. To offer $2 in the first mini-ultimatum game appears to be different from offering the same amount in fourth variant, because in the former case the proposer willingly avoids the $5-$5 proposal which most of respondents consider fair. In the fourth variant, on the contrary, the alternative proposal ($0-$10) implies a sacrifice for the proposer, and therefore the $2-offer is accepted by respondent relatively more frequently. Since intentions can be inferred only by considering both what a player does and what he/she could have done and did not, it becomes clear why traditional game theory but also outcome-based theories of social preferences (e.g. altruism, inequality-aversion) cannot explain the observations collected in the experiments discussed above. This impossibility is a product of the assumption of consequentialism shared by those classes of models.

To be able to fully appreciate the novelty introduced by the rejection of consequentialism endorsed by psychological game theory, we should reflect on the very notion of intentionality and on the contributions that neurosciences have recently produced in the understanding of its nature.

4. Intentionality and PGT.

Intentionality is a central concern in the philosophical theory of action and it has attracted growing attention in the fields of neurosciences, as well as experimental and developmental psychology. In philosophy, the word intentionality means something very specific, notably, the nature of the mental states, desires, beliefs, goals, and other “propositional stances”. Intentionality is the way our mind puts us in relation to the
external world. According to Michael Bratman (1989), an intention is a plan of action the subject chooses and commits itself to in pursuit of a goal. An intention, thus, includes both a means and a goal. The latter fact is important to explain why the same action performed with different goals may be assessed in different ways, as the experimental evidence from the previous section shows. Consider, for example, the different way people (and the law) react to a murder or a manslaughter. Different reactions to the same action because, although the same causal link between action and consequences persists, the same action has different ends. In the same way: “If considered intentional, a critical remark can be seen as a hurtful insult; a collision in the hallway as a dangerous provocation; and a charming smile as a hint of seduction. But if considered unintentional, that same remark may be excused; the same collision may lead to a new friendship; and the same smile might simply indicate a good mood” (Malle & Knobe, 1997:101).

In the domain of strategic social interactions the role of intentions is crucial because, if we may come to know others’ intentions we will be able to express judgements and to predict their behaviour.

McCabe et al. (2000) investigated the extent to which cooperation is influenced by the representation form of the game and found that in an extensive form game, cooperation is achieved more often than in an equivalent game when represented by a normal form. Their explanation is based on the idea that the extensive form makes it easier to engage in the mind-reading process that allows players to infer others’ intentions from their moves: to read another person’s thoughts or intentions by placing themselves in the
position and information state of the other person” (2000:4404). This mind-reading process, in turn, via reciprocity, tends to favor players’ coordination on cooperative outcomes. That is consistent with the finding according to which achieving cooperation in games with private information is harder that in games of complete information. The lack of information about others’ incentive structure makes it difficult to infer their intentions from what they chose to do.

Neurosciences have recently shown that the capacity to understand others’ action attaching to them specific goals is typical of superior primates and in particular of human beings (Rizzolatti, et al., 2001), possibly because of the existence of sophisticated linguistic abilities (Tommasello, 2000). By the age of 4, children can infer what people know, think or believe on the basis on what they say or do: this ability is known as “mind-reading” (McCabe et al., 2000:4404). Neurosciences provide different theories to explain how this process works, commonly referred to as theories of mind or ToM. These theories can be distinguished into two large families: the “theories of theory of mind”, also known as “theory-theory” (TT) and the so-called “simulation theories” (ST).

The former theory postulates that agents tend to explain and predict others’ behavior by means of a set of causal laws that form a sort of folk psychology (Carruthers & Smith, 1996). According to TT, agents use simple explanatory laws to link the (unobservable) determinants of behavior (desires, beliefs and intentions) to external (observable) stimuli in order to predict people’s actions. This attributions process works on the basis of theoretical reasoning that involves (tacitly) shared causal laws.
The second class of theories, simulation theories (Davis & Stone, 1995), on the other hand, posit that the attribution of mental states to other people works through mental representation: agents are able to simulate the reasoning process, pretending to be, literally, in the “mental shoes” of the other agents whose behavior they are observing or want to predict. “First you create in yourself pretend desires and beliefs of the sort you take [the other] to have (...) these pretend preferences and beliefs are fed into your own decision-making mechanism, which outputs a (pretend) decision” (Gallese & Goldman, 1998:496).

The basic difference between the two classes of theories is that while TT describes mind-reading as a neutral, objective and detached theoretical process, ST considers it as a matter of actual replication of the same neural activities implied in the action the subject is trying to interpret or predict. According to ST, thus, social cognition is not only reasoning about others’ mental states, it is more like experiential insight of other minds.

Gallese & Goldman (1998) and following studies report evidence describing exactly this matching of mental activity between two subjects, an agent and an observer, they define as “mental mimicry”. This mechanism involves a class of neurons, the so-called “mirror neurons” that have the intriguing property of firing both when the subject performs a certain action as well as when she observes someone else doing the same action. Mental mimicry thus can be qualified not only as a theoretical activity, since the mirror neurons re-create in the observer the same mental activity that is ongoing in the brain of the observed. In interpreting the role of the mirror-neurons system, Gallese &
Goldman suggest that it may provide the neurophysiologic basis underlying human beings' ability to represent others' intentions by the observation of their actions.

Having briefly illustrated the role of intentions in the process of social cognition and how the ability to attach intentions to others’ actions may facilitate interpersonal coordination, it is necessary to explore how such an ability can be formalized in the context of abstract games. What I shall try to show in this part of the paper is that PGT provides useful formal tools to do exactly this in a way which is consistent with the conceptual framework of the ST.

Traditional game theory assumes that payoffs depend only on the actions chosen by each player. Outcome-based models of social preferences endorse the same logic by considering some combinations of ego's and alter's payoffs. Whatever motivates players' actions can be summarised by payoffs (or by some combinations of them) attached to the outcomes of the game. The players’ preference ordering over those outcomes is represented by their (extended) payoff vector. While this framework has gained considerable success in many areas within and outside economics, at the same time it has demonstrated as being too narrow to account for many social phenomena, especially those characterised by the presence of belief-dependent motivations. A variety of social emotions, as well as social norms such as trust and reciprocity, cannot be properly formalised in the classical theory in which payoffs are taken as exogenous. PGT develops tools that allow the formal analysis of many of those phenomena, making more manageable the natural but impractical way of modelling hierarchies of beliefs, referred to by Harsany.
In psychological games (Battigalli & Dufwenberg, 2005) players’ utility depends on their strategies and their (higher order) beliefs about their strategies. Geneakoplos, Pearce & Stacchetti (1989) originally developed various solution concepts for psychological games. In a psychological game each player considers each single action within the entire set of actions available at each node. In this way it is possible to infer other players’ intentions by counterfactually thinking about what they could have done and did not. A (psychological) equilibrium is obtained when both players maximise their payoffs and the hierarchies of beliefs are confirmed. Dufwenberg & Battigalli (2005), generalising Geneakoplos, Pearce & Stacchetti’s model, extend this framework even further:

i) allowing beliefs being revised as the game unfolds;

ii) defining a psychological sequential equilibrium, in the line of Kreps & Wilson (1982);

iii) considering non-equilibrium analysis in the line of rationalizability (Bernheim, 1984; Pearce, 1984).

In recent years, few theoretical models applying PGT (Rabin, 1993; Dufwenberg & Kirchsteiger, 2004) have been proposed to explain deviations from selfish behavior on the basis of players’ different assessments and their reaction about the negative or positive intentions of other players, as signalled both by their choices and by their bygones. Other hybrid models (Levine, 1998; Charness & Rabin, 1999; Falk & Fischbacher, 2006) combine the role of intentions with that of distributional concerns.
Summarizing: while standard game theory and outcome-based models of social preferences provide a over-simplified account of intentionality, mainly because of the assumption of consequentialism, PGT (and subsequent intention-based models of social preferences) allows to describe agents able to attribute intentions to others’ actions by observing others’ chosen actions and their results together with their un-chosen options. This process appears to be surprisingly similar to that described by ST, and it can formally be described and analyzed by PGT. It is important to notice, however, that while the latter considers the process as a deliberate and volitional act, ST posits that simulation occurs automatically and unconsciously. This, however, does not reduce the usefulness of the formal description of the mental process enabling us to interpret others’ intentions and goals.

5. The trust problem

Consider now the second problematic issue: the problem of trust and that related of payoff endogenization. It is worth noticing that in the most advanced formalization of PGT, Battigalli & Dufwenberg’s theory of dynamic psychological games (2005), 6 out of 9 examples are presented using games of trust. Although it is true that the problem of trust is related to that of intentionality, because a genuine trusting action can only be intentional and not incidental, this is only one side of the coin. In fact, as I shall explain in more detail later, a full account of the experimental evidence on trusting interactions, needs a class of formal models where payoffs are not exogenously given, as in classical game theory, but become endogenous. That is, payoffs depends not only upon actions and
beliefs about actions, but also upon beliefs about beliefs about beliefs, and so on (see Huang and Wu, 1994; Battigalli and Dufwenberg, 2007). All these models can only be developed within the theoretical framework provided by PGT.

A fiduciary interaction is characterized by three basic elements: i) potential negative consequences; ii) risk of opportunism; iii) lack of control. All these elements are summarised in the so-called “basic trust game” (Figure 3): point i) is described by $b<a$; point ii) depends on being $e>f$; point iii) is obtained modelling the game as a non-cooperative, two-stage sequential game.

FIGURE 3 ABOUT HERE

In the trust game, $A$ chooses first either L or R; in choosing L, players get a payoffs pair equal to $(a,d)$. But if $A$ chooses R, the choice passes to $B$, who, in turn, can choose either L or R. In the first case, she gets $e$ and $A$ gets $b$; in the second case $B$ gets $f$ and $A$ gets $c$. Given such a payoff matrix and the relations between its elements, the game theoretical advice for a rational course of action will be for $A$ to choose L and stop the game there. Although in this game $A$’s preferred outcome would be that described by $(c,f)$, such a situation is not an equilibrium outcome and, in fact, it is achievable only when both player $A$ deviates from her individually rational course of action choosing R, and player $B$ renounces her rational strategy by opting for R. However, as robust
experimental results show⁵, a significant number of As prefers to play R and a significant number of Bs resists the temptation of the opportunistic choice by playing R. Several different strategies have been followed to account for these data, revealing a continuous vacillation between a substantive definition of self-interest and a tautological one: heterodox theoretical models based on behavioral principles such as altruism, inequity aversion, on the one hand, and other, more radical moves, such as Ken Binmore’s (1998) “tautologism”, that suggests to rationalize the anomalous choices by considering payoffs a kind of utility measure as in the revealed preference theory. In Binmore’s view payoffs are nothing but ex-post indexes of preferences, and anomalous experimental results can be reconciled with standard theory, simply by re-describing the game to make it consistent with people’s choice. He considers a particularly strong form of revealed preference theory, in which, as a matter of definition, a player’s choice always reveal her preferences. Thus, once a game has been numerically specified, as noticed by Robert Sugden: “certain propositions about what a player will do (for example, that she will not choose a dominated strategy) are necessary truths, and not merely the implications of particular solution concepts which game theorists are free to dispute (Sugden 2001:F222)

Consider, for instance, the game of figure 4, the payoff vectors, are derived from the objective ones by means of the above-mentioned principles, altruism, inequality aversion and tautologism. However, all these explanations present important drawbacks that reduce their descriptive power (Pelligra, 2005a). Two of the emerging alternative

⁵ See Camerer (2003) for a review.
theories are based on reciprocity (Rabin, 1993) and on trust responsiveness (Pelligra, 2005b). All these theories consider an extended utility function that incorporates objective payoffs and transforms them according to some behavioral principle, which may render the trustful and trustworthy combination of choices an equilibrium of the trust game. For altruists $B$’s utility depends on $B$’s payoff plus $A$’s payoff weighted by some parameter $\alpha_B$, ranging between 0 and 1 (Margolis, 1982). Inequity averse $B$s aim both at maximizing their own objective payoff and minimizing the difference between $B$’s and $A$’s payoffs (Fehr & Schmidt, 1999). The negative impact of inequality is weighted by different parameters, $\alpha_B$ and $\beta_B$ (with $\alpha_B > \beta_B$) when $B$ occupies, respectively, the disadvantageous or the advantageous side of the distribution.

According to Rabin, (positive) reciprocity is the act of conferring benefits on people who have previously *materially* benefited you, while trust responsiveness is the act of conferring benefits on people who have shown that they expect you to do so, and have willingly exposed themselves to harm in the event you act on material self-interest. So trust responsiveness assumes that an explicit act of trust may “induce” or “elicit”, to some degree, a trustworthy response. In this respect, trust is said to be *responsive* or *self-fulfilling*. While reciprocity is based on the *joint action* of both material and
psychological incentives, trust responsiveness is exclusively based on a psychological-moral motivation.

Several experiments (Dufwenberg & Gneezy, 1998; Bacharach et al., 2005; Pelligra, 2005b) show that when you rule out the effect of altruism, inequity aversion, but also reciprocity, some instances of trustful and trustworthy behavior are still to be explained. From these tests, trust responsiveness seems to emerge as the most basic explanation for trust. However, apart from the differences in the descriptive power, altruism, inequity aversion and trust responsiveness differ also on more substantial grounds. While the former two, in fact, are theories of forward looking behavior, that consider intentions that players attribute to one another as irrelevant, the latter, on the contrary, is a theory of backward looking behavior. That means that in altruism and inequity-aversion based theories the payoff transformation is exogenous, that is external to the game, trust responsiveness implies an endogenous payoff transformation. You need endogenous payoffs when you want to model situations where the relational aspect is important, where the very fact of entering a relation, an interaction, affect players’ prior preferences. Such relational factors can be formalized using psychological game theory. Consider now, as an illustration, the following variant of the Simple Trust Game (G5).

6 The possibility of payoffs’ endogenisation may represent a step in the construction of a general theory of the individual in economics. As Davis (2003) argues, in fact: “the lack of attention given to the theory of individual in economics reflects economists’ widely held and generally unexamined belief that individuals are exogenous the the economic process” (p. 11). Endogenous payoffs are a way to describe how the “economic process”, the ongoing strategic interaction, somewhat modifies players’ identity.
Denote with $p \in [0,1]$ the probability that $B$ plays $R$; $1-p$ is the probability with which $B$ plays $L$. In the same way $q \in [0,1]$ represents $A$’s belief about $p$. Analogously, $r$ denotes $B$’s belief about $q$, that is, $B$’s belief about $A$’s beliefs about $B$’s choice. In this way we describe $B$’s hierarchy of beliefs, in particular, his first and second order beliefs. These beliefs are crucial to transform the standard game into a psychological one. I will restrict my formal discussion to the usual equilibrium analysis leaving aside considerations of out-of-equilibrium behavior.

Suppose $B$ observes $A$’s trustful choice ($A$ plays $R$); we are now in the second node of the game, where $B$ has to move. In this version of the trust game, $B$’s payoff from being opportunist is made up by a material part and a psychological one, which in turn depends on a “guilt factor”. The negative impact of guilt on $B$’s overall utility is a multiple $G \ (G>0)$ of $A$’s expectation $r$. The intuition underlying such a formalization is that $B$ suffers a psychological loss when he deliberately lets $A$ down knowing that $A$ has trusted him, and such loss is proportional to $B$’s belief about $A$’s expectation of $B$’s trustworthy behavior.
We can solve the game by isolating its psychological equilibria. In a psychological equilibrium players maximize their utility, and their first and second order believes are confirmed \((p=q=r)\). This particular game shows three of such equilibria:

1) in the first, \(A\) expects \(B\) to play trustworthily; given this, \(B\)'s psychological cost from frustrating \(A\)'s expectation becomes strong enough to lead \(B\) to the expected, trustworthy choice. \(A\) knows that and sets \(q\) accordingly \((q=1)\): she plays \(R\); \(B\) knows that as well, and sets \(r=q=1\): he plays \(R\). In the first equilibrium \(A\) plays \(R\) and \(p=q=r=1\), that is, \(B\) plays \(R\). Trust is self-fulfilling;

2) in the second equilibrium, \(A\) expects \(B\) playing opportunistically, that choice would not produce any psychological cost for \(B\). \(B\) knows that and sets \(q=0\); consequently \(B\) sets \(r=q=0\). In the second equilibrium \(A\) plays \(L\) and \(p=q=r=0\), that is, \(B\) plays \(L\);

3) the third (mixed-strategy) equilibrium is obtained by setting \(B\)'s payoff from opportunism and from trustworthiness equal, and imposing \(p=r\). In this third equilibrium, which only exists if \(pc+(1-p)b>a\), it follows that \(A\) plays \(R\) provided that \(p=q=r=(e-f)/G\) and \(0<(e-f)/G<1\). The associated payoffs are \(pc+(1-p)b\) for \(A\) and \((1-p)(e-rG)+pf\) for \(B\).
In this third case, both trustworthiness and opportunism may follow A’s trustful move, depending on the players’ beliefs. The denominator $G$ in (2) represents the impact of social sentiments, or internal reasons, on $B$’s utility. This factor, as well as the difference $(e-f)$, directly affects the probability of $B$’s trustworthy behavior.

Notice how, in this simple psychological trust game, the first two equilibria are strongly determined by the different intentions of A. If A intends to trust B and B correctly ascribes this intention to A's trustful move, we have an increase in the likelihood of a trustworthy response; on the other hand, when A's action is motivated by diffident intentions and these are correctly inferred by B, he/she will be pushed towards an opportunistic choice.

One important feature of this analysis is that the Simple Trust Game, once re-described as a psychological game, becomes a coordination game\(^7\). Which equilibrium will be selected depends, in fact, on the way players coordinate their first and second order expectations. In the pure strategy equilibria, if A expects B to behave trustworthily and such expectation is known also to B, the latter will be then induced to trustworthiness; both players coordinate on $r=q=1$. On the contrary, if A believes B to be an opportunist, this reduces B’s psychological cost of opportunism leading her to behave exactly as A expected. Coordination takes place on $r=q=0$

Once recognized that behavioral coordination depends on the players’ first and second order beliefs, we are induced to investigate the existence of equilibrium selection

\(^7\) Camerer and Thaler (2003) provide a similar interpretation.
devices. In other words, what favors or hinders beliefs coordination. This question leads us directly to the third problem of our original list, the problem of framing.

6. Social Framing Effects

Most of the work done on the role of framing in decision-making focuses on individual choices, neglecting strategic environments almost completely. However, if we consider Kahneman and Tversky’s originally definition of the framing process as “controlled by the manner in which the choice problem is presented as well as by norms, habits and expectancies of the decision maker” (1987:257), its importance to the understanding of interactive decision and thus, for traditional and psychological game theory, clearly emerges.

Coordination games as well as all the situations where the role of players' expectations is explicitly taken into account determine the problem of multiplicity of equilibria. While on the one hand empirical evidence shows that real people are remarkably capable to coordinate over the optimal outcomes using theoretically non-relevant information (see Metha, Starmer & Sugden, 1994), on the other, since the pioneering contribution of Thomas Schelling (1969), and apart from a few exceptions (Metha et al., 1995; Bacharach & Sthal, 2000; Janssen, 2001; Bacharach, 2006), theorists have been unable to provide a satisfactory explanation of such an ability. Focal points theories aims at understanding the process through which players connect their strategies with the context of the game using the informational richness of the latter to find in the environment clues useful to coordinate their choices. Here is the connection between the coordination problem and
the framing problem. Different frames may, in fact, favor or hinder the reading of such information. In psychological games, coordination is even more problematic because an equilibrium requires not only behavioral consistency, but also higher order beliefs' compatibility. This fact has convinced Battigalli and Dufwenberg that “assuming equilibrium may be assuming too much especially in psychological games” (2009:9). If the latter may be true in theory, on the empirical side, the kind of multiplicity of equilibria implied in psychological games may constitute an element of realism, precisely because it leaves room for the working of frames as beliefs and strategies correlating device. If frames, in fact, influence beliefs, in PGT, in turn beliefs influence motivations for action.

In the context of the psychological trust game, we have already observed how the perception we have about the idea the others have of us may influence our decision to be trustworthy or opportunist. Such perception develops and strengthens in relation to others’ actions, and particularly in relation to our interpretation of such actions, namely our beliefs about others’ beliefs. Such interpretation, in turn, is strongly affected by the description or the frame within which actions take place. This point is subtle and needs clarification. The idea underlying the concept of focal point posits that, in a strategic situation, a particular frame may signal cues that favour choices’ coordination. This is so, because a frame influence beliefs about others’ choices. My argument goes a step further. If agents care for others' intentions, then the description of a situation influences behavior independently from beliefs about others’ choices. In this way there arises a second channel through which frames signal cues, affecting now, beliefs about
others’ beliefs. If a certain strategic situation may give rise to belief-dependent emotions (pride, shame, resentment, guilt, embarrassment, ...), as it is often the case, then motivations may depend on beliefs directly. In PGT players’ choices depends on beliefs about other’s choices and about other’s beliefs in a direct way, so if beliefs change, actions could change too. This mechanism is not surprisingly consistent with a series of empirically well-established phenomena: first, it is much easier to coordinate towards cooperative outcomes in experimental games, when they are represented in strategic-form instead of in normal-form since the former facilitates mind-reading, and therefore coordination (McCabe et al., 1998).

Second, according to the so-called “false consensus effect”, people who expect others to cooperate tend to be more willing to cooperate themselves (Ross et al., 1977; Butler et al., 2009). This phenomenon, which is hard to reconcile with existing models, is a consequence of the coordinating process at work in psychological games. Third, a series of experimental studies show how subjects’ behavior in the same public good- or prisoners’ dilemma-game, is significantly different when the game is modified by non-theoretically relevant variables, as, for instance, the mere semantic description of the situation ("Community Game" vs. "Wall Street Game", Ross & Ward, 1996; Liberman, Samuels & Ross, 2004), the label of co-players ("opponent" vs. "partner", Burnham, McCabe & Smith (2000), or the actions set ("give" treatment vs “take” treatment, Dufwenberg, Gächter & Hennig-Schmidt, 2008). All these variants induce a change in players’ beliefs about others’ beliefs that affect their motivations and, thus, their actual choices.
The fact that in all these games PGT predicts multiple equilibria constitutes, in my interpretation, not a drawback of the theory, as Battigalli and Dufwenberg seem to think, but instead an element of realism, leaving room to the working of the framing-effects just mentioned and attenuating the violation of “description invariance” that the framing-effect implies, with no recourse to cognitive bias.

7. Conclusions

The main thesis I wanted to put forward is that, thanks to the epistemic consequences of the hierarchy of beliefs formalised in PGT, we became capable to address, and to some extent, solve three of the most theoretically and empirically relevant problems that affect classical game theory.

First: intentions are important to players to ascribe meaning to actions, as the evidence reviewed in section 3 shows; PGT enriches the standard account of intentionality and helps explaining why different reactions may originate from the same action. Simulation theories postulated that this intention attribution process works through mental mimicry, that is, by activating the same neuronal circuits that are active in the brain of the subject who actually performs the action. PGT allows to formalise this process by enabling theorists to manage an entire hierarchy of beliefs, as well as the counterfactual reasoning implied in the “mind-reading” activity.

Second: trust is a relational construct. My claim is that it can be better understood if we assume that trustfulness may elicit trustworthiness, a mechanism that has been dubbed “trust responsiveness” (Pelligra 2005a, 2005b). Such responsiveness implies that
payoffs cannot be taken from outside the game. PGT provides a method to endogenize such payoffs and to analyse, in this way, belief-dependent emotions, such as pride and guilt, that are supposed to play a major role in explaining trustful and trustworthy behaviour.

Third: psychological equilibria are affected by a problem of (higher order) belief coordination. This fact makes the presence of a coordinating device necessary. While in general, multiplicity of equilibria is viewed as problematic, I maintain that within the conceptual framework of PGT, it introduces an element of realism, as it leaves room for the working of decision (social) frames that facilitate the coordination of players' beliefs towards a unique equilibrium.
REFERENCES


[59] - PUGNO, M. 2005. The happiness paradox: a formal explanation from psychoeconomics. Mimeo, Economics Department, University of Trento,


FIGURES

FIGURE 1: The Voluntary Trust Game

A

B

15,30

20,20

25,25

FIGURE 2: The Involuntary Trust Game

A

B

15,30

25,25

FIGURE 3

G3: The Simple Trust Game

Player A

L

R

Player B

L

R

Player A

a

b

Player B

d

e

f

b<a<c; f<e;
FIGURE 4
G4: “Trust Game re-described”

<table>
<thead>
<tr>
<th>Objective Payoffs</th>
<th>Player A</th>
<th>Player B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>B altruist</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>2.5</td>
</tr>
<tr>
<td>B inequity averse</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tautologism</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

FIGURE 5
G5: “The Simple Trust Game with endogenous Guilt”

<table>
<thead>
<tr>
<th>Player A</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Player B</th>
<th>1-p</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Player B</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-p</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td></td>
</tr>
</tbody>
</table>

| 1-p      |   |
| a        |   |
| c        |   |

\[ c > a > b ; e > f ; G > 0 ; \]