UNIVERSITÀ
CATTOLICA
del Sacro Cuore

DIPARTIMENTO DI ECONOMIA INTERNAZIONALE
DELLE ISTITUZIONI E DELLO SVILUPPO

Guido Merzoni - Federico Trombetta

Foundations of trust, interpersonal relationships and communities

N. 1201

V&P VITA E PENSIERO

UNIVERSITÀ CATTOLICA DEL SACRO CUORE

DIPARTIMENTO DI ECONOMIA INTERNAZIONALE
DELLE ISTITUZIONI E DELLO SVILUPPO

Guido Merzoni[*] - Federico Trombetta[**]

Foundations of trust, interpersonal relationships and communities[***]

N. 1201

V&P VITA E PENSIERO

[*] Corresponding author: DISEIS and Faculty of Political and Social Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 20123 Milano – ITALY. guido.merzoni@unicatt.it.

[**] Faculty of Political and Social Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1 20123 Milano - ITALY.

[***] We would like to thank Carlo Beretta, Simona Beretta and Mario Maggioni for useful discussions. Federico Trombetta thanks Michela Boccola for her support. Financial support from the Catholic University of Milan under the D.1 research project on "Identità comunitaria, motivazioni individuali e cooperazione" is gratefully acknowledged.
A previous and different version of a few paragraphs of this paper are contained in Trombetta (2012).

Prima di essere pubblicati nella Collana Quaderni del Dipartimento di Economia internazionale, delle istituzioni e dello sviluppo edita da Vita e Pensiero, tutti i saggi sono sottoposti a valutazione di due studiosi scelti prioritariamente tra i membri del Comitato Scientifico composto dagli afferenti al Dipartimento.
I Quaderni del Dipartimento di Economia internazionale, delle istituzioni e dello sviluppo possono essere richiesti alla Segreteria (Tel. 02/7234.3788 - Fax 02/7234.3789 - E-mail: segreteria.diseis@unicatt.it).
www.unicatt.it/dipartimenti/diseis

Università Cattolica del Sacro Cuore, Via Necchi 5 - 20123 Milano

www.vitaepensiero.it

ISBN 978-88-343-2527-8

Abstract

We claim that the emergence of trust is best explained by relation-based arguments. After briefly surveying alternative explanations which concentrate on material payoffs both with self-centered and with other-regarding preferences, we examine theoretical discussions of cooperative and trust behavior framed in terms of attitudes, esteem and, most of all, intentions. An important implication of all these approaches is that the relational element makes human interactions different, as it is also documented by a lot of evidence produced by neuroeconomic experiments.

When trust is based on relations and on the recognition of the others' intentions, efficient outcomes are brought about by the agents' (at least) partial disregard for the maximization of their material payoff and by heavily personalized interactions. Both these features are distinctive of the functioning of communities and the particular way how they work and solve coordination problems.

Keywords: trust, community, intentions
J.E.L. classification: C70, C72, D63

INDEX

1. Introduction

As many economists, as well as Pope Benedict XVI in the Encyclical Caritas in Veritate[1], remind us, trust is the engine of the well-functioning of any economic and social system.

From an economic point of view, trusting someone entails a risk: it is a bet over the relationship with another person, which may appear irrational from a myopic perspective. As explained by James[2], «*in the language of economics, trust can be viewed as an expectation, and it pertains to circumstances in which agents take risky actions in environments characterized by uncertainty or informational incompleteness. To say "A trusts B" means that A expects B will not exploit a vulnerability A has created for himself by taking the action*».

In other words, A trusts if, when and to the extent he bets on the relationship with B, accepting the risk that his behavior could not be trustworthy. If Bis selfishand self-interested, the risk A takes is hopeless and he certainlywill bebetrayed. Knowing this,A should not trust B.

However, this is not a good description of what happens in reality: the empirical world of economic interactions (and also of a lot of other types of interaction) shows that, in some way, trust emerges.And when this doesn't happen, economic and social consequences are extremely negative. According to Chami and Fullenkamp's[3] description of the Russian transformation

[1] "In fact, if the market is governed solely by the principle of the equivalence in value of exchanged goods, it cannot produce the social cohesion that it requires in order to function well. *Without internal forms of solidarity and mutual trust, the market cannot completely fulfil its proper economic function*. And today it is this trust which has ceased to exist, and the loss of trust is a grave loss", Benedict XVI (2009), n.35.

[2] James (2002), p.291.

[3]Chami-Fullenkamp (2002).

from a centrally-planned to a free market economy, «*while some part of this decline in living standards can be attributed to the failure of the communist era economic institutions, much of it is due to the fact that lack of trust has prevented the growth of a new, effective economic infrastructure to replace the old and failed one*». But trust is central also in our well established market economies, characterized by a general proliferation of agency relationships: trust plays a pivotal role in a huge number of human relationships and, as a consequence, also in a lot of economic, political and social interactions, as, for instance, in intertemporal trade without complete contracts[4] or agency relationships in labour interactions, political representation or international relationships[5].

In all these examples, the decisions both to trust someone and to behave as a trustworthy individual are not consistent with the pursuit of short-term individual self-interest, and as a consequence this decision needs to be explained in a different way.

Such explanations abound in the mainstream,as well as in behavioral and neuro-economic literatures.Most are based on the consideration of people attitudes towards the outcome of their interactions, be they only self-centered but explicitly taking into account the possible repetition of the game played, or other-regarding in terms of equity and fairness, so that betraying other people trust implies an unfair division of the output, which should be dismissed.

Relatively few papers abandon a pure consequentialist view of trust and explicitly consider the real relational dimension of agents interactions. In those papers, trusting and trustworthy behavior is explained as motivated by the acknowledgement of the other party attitudes ("I take pride of the esteem of the other

[4] Greif (2006).

[5] Colombo-Merzoni (2006), Bull (1987), Tabarrok (1994), Kydd (2000).

players and this motivates me to be trusting or trustworthy") and intentions ("I appreciate the other party's move of trusting me and decide to positively correspond to it, even though this decision is not maximizing my material payoff.") towards the decision-maker.

Considering the role of attitudes are Levine (1998), on the effect of altruism and spitefulnesson reciprocity, and Ellingsen and Johannesson (2008), on how social esteem promotes prosocial behavior and may be crowded out by control systems and monetary incentives.

A growing stream of literature on the so-called *psychological games*addresses instead the role of intentions, by making payoff belief-dependent (Rabin, 1993; Dufwenberg – Kirchsteiger, 2004).

Support for the importance of attitudes and intentions emerge both from behavioral experiments, in particular in McCabe et al. (2003), and fromneuroconomic experiments (Rilling et al., 2002, Rilling et al. 2004, Chang et al., 2011), where players are registered having different behaviors and neurological activations when playing with human counterparts as opposed to automata, despite facing identical material payoffs.

Indeed, there seems to be something absolutely distinctive in human interactions that goes beyond purely material outcomes.

We claim that this is not only interesting *per se*, but also because the way we use to justify and explain trust is a mirror of our vision of human beings and, as a consequence, of the functioning ofcommunities we belong to. Indeed, there is a huge difference between thinking at communities as being madeof-self-centered material welfare maximisersand as being madeof-subjects that, in their decisions, take into account also *the others* and their welfare or, indeed and even more, their attitudes and intentions. As underlined, again, by James[6], «*on the one hand, I may trust because it is prudent for me to do so, if I believe (rationally) that my partner has an incentive to be trustworthy. On the other hand, if Ibelieve that my trading partner retains an incentive to exploit my trust, then I may still choose to trust. While prudence may suggest otherwise, if I choose to trust I do so out of the "hope" that my partner will not exploit my trust*». This distinction, as pointed out by the same author, recalls the one by AmartyaSen[7]between «*sympathy*» and «*commitment*»: «*a person is sympathetic when his concern for another's welfare directly affects his own utility. By contrast, a person is committed if she is willing to undertake an activity that clearly conflicts with her self-interest and sympathetic preferences (i.e. if the activity does not benefit her)*»[8]. It is clear that there is a huge difference between a society in which each economic action is justified only by egoism or indeed by sympathy[9] and one in which commitment based on real relationships between individuals could be, at least, a credible possibility.

This paper is organized as follows. In the next section we introduce the trust game, the analytical tool we use as a reference for our study of trust, and briefly survey explanations of cooperative behavior based on material payoffs both with self-centered and other-regarding preferences. In section 3,we examine explanation based on reciprocal behavior, starting from an experiment by McCabe et al. (2003) and then considering theoretical discussions of cooperative behavior framed in terms of attitudes and esteem on one hand, and intentions and psychological games on the other; finally, we briefly discuss the implications of the psychological games approach to explain-

[6] James (2002), p. 303.

[7]Sen (1977).

[8] James (2002), p. 303.

[9] Obviously, in the sense of Sen.

cooperative behavior in the trust game. Section 4 contains a brief survey of neuroeconomic papers supporting the vision that the relational element makes human interaction different.

Finally, in the concluding section we try to highlight some of the possible implications for the conception of communities as coordination-promoting environments brought about by the proposed relational explanations of trust.

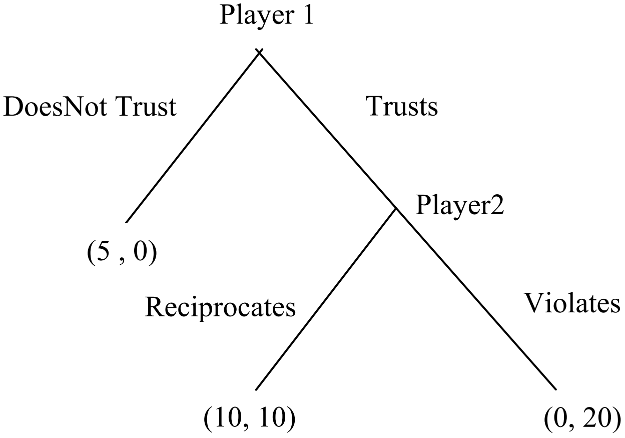
2. Trust games and cooperative behavior based on material payoffs: repetition, reputation and other-regarding preferences.

2.1. The trust game

The most important analytical tool used to study the problem of trust is the so-called trust game, first introduced as such by Kreps (1990), which is a one-side prisoner's dilemma with sequential choices. In the version shown in Fig. 1, player 1 must choose between a costly action that implies an investment of his endowment on the second player (and a risk, betting on the trustworthiness of the second player) and the free decision not to act. If the investment is made, the sum is multiplied by a factor (4 in figure 1's example) and player 2 must decide between a fair division of the multiplied endowment and the egoistic choice to keep the whole sum for himself.

It is immediately clear that, if the game is played one-shot under complete information, in the subgame perfect equilibrium the choice of player 1 is not to trust.In fact, using backward induction he knows that player 2 maximizes his individual welfare keeping the whole sum, and as a consequence the best choice for him is not to trust.

Figure1 - *The extensive-form of a trust game*



This equilibrium is Pareto-inefficient: if the players chose to behave “cooperatively”, in fact, they could reach a situation in which both are better off. As explained by Kreps (1990), however, the inefficient equilibrium is the unique equilibrium of the trust game, obviously with the imposition that the game is played only once and «*absent other considerations*».

It is immediately obvious that the analysis is, at this point, partial and – most importantly – contradictedby facts: interactions with a structure of incentiveslike the one described above not infrequently lead to a different outcome. The cooperative behavior prevails in quite a large number of cases.

2.2. Repetition and reputation

A traditional explanation of cooperative play in similar situations by mainstream game theory is based on repeated interactions (Kreps-Wilson, 1982; Milgrom-Roberts, 1982) and leads to the so called *Folks' Theorems*. If the same two players interact "many times" they will find the way to reach an efficient outcome. This result is proved for repeated games with infinite horizon, with uncertain duration and with finite horizon under incomplete information.[10]Even the most classical *homo oeconomicus*, under some conditions, seems to be able to trust his counterpart.

A feature of most of these settings, which many perceive as a weakness,is that they are characterized by multiplicity of *equilibria*; this certainly limit the usefulness of those analyses to make predictions, since the same game could be played in many different ways.

Cooperation in repeated trust (and other prisoners' dilemma-like) games may emerge also when repetitions could be interrupted by a change of partner, as shown by Colombo and Merzoni (2006), provided that the relationship is perceived as sufficiently stable.[11]Kreps(1990) shows that the creation of a reputation over an array of choices can sustain trust even when the interacting players are not always the same, obviously under the assumptions that past behavior is observable to everybody interested and that reputation within the group is a sufficiently valuable asset to discourage the deviation motivated by an immediate gain.

[10] The classical reference here is Fudenberg-Maskin (1986).

[11] See also Ghosh-Ray (1996), Kranton (1996), Colombo-Merzoni (2004).

2.3. Other-regarding preferences: pure and impure altruism, fairness and inequity aversion

When trust and cooperation are based on repetition the cooperative strategy is played because each player has the reasonable expectation that, in the long run, this is the choice maximizing individual welfare. Clearly, it is a purely egocentric vision of trust as “others” do not enterin the players utility functions. However, trusting and trustworthy behaviors may follow from the players regard for other players, in terms of pure (Chami and Fullenkamp, 2002) or impure altruism (Andreoni, 1990), as well as of concern for fairnessand aversion to inequity (Fehr-Schmidt, 1999).

Chami and Fullenkamp (2002) show how trust, modeled as symmetric pure altruism between a principal and an agent within a firm, allows to limit agency costs more effectively than standard remedies such as incentive contracts or monitoring devices, or indeed a paternalistic attitude of the principal. They assume that the principal’s utility is given by

$$U_p = u_p(r + x - w) + \beta_p[u_a(w) - v(e)] \qquad (1)$$

and the agent’s utility is given by

$$U_a = u_a(w) + \beta_a\big[u_p(r + x - w)\big] - v(e) \qquad (2)$$

where r is a rental fee earned by the principal, x are the revenues of the firm, w is the agent’s wage and $v(e)$ expresses the agent’s disutility of effort e. Parameters β_p and β_a express the degree of altruism of the principal and the agent respectively, i.e. the weight they place of the other utility.

When both the principal and the agent are egotistic, and so $\beta_a = \beta_p = 0$, the agency problemis contrasted by using standard incentive contracts, and the equilibrium has the usual second-best, inefficient properties.An altruistic principal, with

$\beta_p > 0$, facing a non-altruistic agent, with $\beta_a = 0$, would behave paternalistically and lead the firm to an even worse situation, where the agent exerts a very inefficient level of effort, causing the firm to succumb to competition of firms using incentive contracts. However, when altruism is symmetric, and so $\beta_a = \beta_p$ is at least approximately true, the agent cares of the positive effect of his effort on the principal's utility and the principal is willing to reward the agent by insuring her against business risks.This results in the agent working harder and the firm becoming overall more efficient.

Andreoni (1990) proposes a theory of impure altruisms in the private provision of public goods. He assumes that in their contributions to financing public goods agents take into account not only the effect on the overall amount of public goods that will be provided, as they would if they were purely altruistic, but also the "warm glow" produced by the mere act of giving. Hence, in a simple setting with only one private and one public good and n agents, agent i's utility function is assumed to be

$$U_i = U_i(x_i, G, g_i) \qquad (3)$$

where x_i is the amount of the private good consumed by agent i, $G = \sum_{i=1}^{n} g_i$ is the total amount of the public good and g_i is agent i's contribution to the provision of such a public good. Agent i's contribution g_i enters the utility function both indirectly, by adding to the total amount of the public good provided and directly, as the act of giving is utility-enhancing per se. The optimal choice is then found as a function also of the individual budget constraint and of other agents contributions. Pure altruism and pure egoism are special cases of (3), with the utility functions being $U_i = U_i(x_i, G)$and $U_i = U_i(x_i, g_i)$ respectively. Impure altruism is probably the most common case and the degree of altruism would typically varies across a given population.

The main focus of the paper is to provide a model able to explain some empirical puzzle related to public goods contribution. In particular, it is shown that with impurely altruistic agents, at difference to some previous theoretical results, but consistently with empirical evidence, redistributions of income would not be neutral to public goods provision: a transfer from less to more altruistic individuals increases the total amount of public goods provided.

Other-regarding preferences characterize also Fehr and Schmidt (1999). They focus on the role of fairness and inequity aversion in determining the outcomes of strategic interactions in a wide variety of settings. In particular they build a theory of fairness, which is able to reconcile apparently contradictory results of experimental studies. They “model fairness as self-centered inequity aversion”[12], so that agents do not care about the overall equity of the distribution of payoff in the entire population they belong to, but they evaluate their own payoffs comparing them with a reference outcome, being concerned with their relative, and not only with their absolute, values.In a setting with n agents and a vector $x = x_1, \dots, x_n$ of monetary payoffs, one for each agent, the utility function of agent i is assumed to be

$$U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} max|x_j - x_i, 0| - \beta_i \frac{1}{n-1} \sum_{j \neq i} max|x_i - x_j, 0| \qquad (4)$$

where$\alpha_i \geq \beta_i$ and $0 \leq \beta_i \leq 1$. Whileα_i represents agent i’s degree of aversion to disadvantageous unequal outcomes,β_i captures agent i’s degree of aversion to advantageous unequal outcomes. So, agents do not like any unequal outcomes, but

[12] Fehr-Schmidt (1999) p.819.

dislike the most situations where inequity is at their own disadvantage.

Fehr and Schmidt (1999) show thata population where a share of agents have such a concern for fairness, while all the others are purely self-interested, will be characterized by the prevalence of fair behavior in setting like the ultimatum game, the gift exchange game and the public good game with punishment and by the prevalence of unfair behavior in others, like the competitive market game or the public good game without punishment. Hence, to explain behavior apparently in contrast with self-interest, one needs not to resort to extreme assumptions on agents preferences, which would be impossible to reconcile with observed self-interested behavior in other settings. Assuming some heterogeneity of preferences in the population does seem both reasonable and useful to produce more encompassing explanation of observed behavior.As one of the main conclusions of their analysis, the authors note that the observed outcome in many experimental and real world settings cannot be explained by focusing only on the preference structure of the agents involved or on the strategic environment where they operate, but on the interaction between the two.

2.4. A more decisive move toward really human interactions

In the settings presented in this sectioncooperation has an outcome-based explanation: players are modeled as interested only in their own or other players material payoffs and no feeling or psychological motive plays a role. The fact that a subject is interacting with another person, with a computer or with a monkey does not apparently make any difference, as long as he knows that the person, the computer or the monkey are rational subjects, whose aim is the maximization of their utility.

In the next section we concentrate on alternative explanations, which add more structure to the representation of the interaction among players, who, when deciding how to behave, also look at their opponents intentions and suitably reciprocate kind or unkind behavior.

3. The relational dimension of trust building

3.1. An example on reciprocity

McCabe, Rigdon and Smith (2003) highlight that “the others” may not really matter, or not so much, in terms of material welfare, but rather for the interpretation of their actions. In their “trust–reciprocity” model the first player thinks that the second will interpret his cooperative action as a “gift” of trust motivated by his perception of the second player’s trustworthiness. As a consequence, player 2 will indeed behave as trustworthy.

It seems to be decisive the fact that player 1 accepts a risk, chooses to trust player 2 and bets on the relationship. But are real interactions described by such model?

Figure 2 - *McCabe et al.'s voluntary trust game*

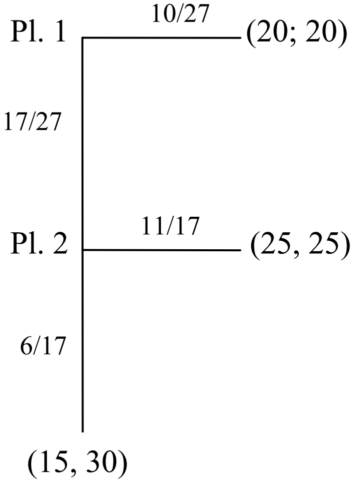
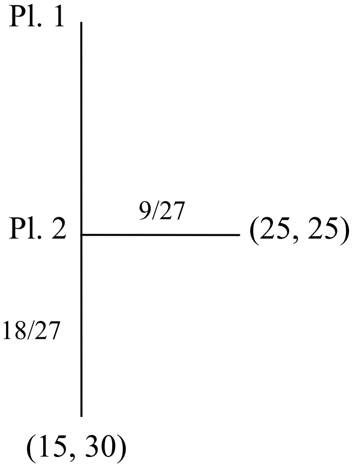


Figure 3 - *McCabe et al.’s involuntary trust game*



McCabe et al.(2003) propose the two versions of a trust game shown on figure 2 and 3. The first is “voluntary”: player 1 can choose between a bet on the relationship with player 2 and a honorable exit option (20, 20); the second is “involuntary”: player 1 is forced to play cooperatively, he has no alternatives.

If concerns for fairness or inequity aversion were at work, as in the model by Fehr and Schmidt (1999) this variation would not cause any difference in the players’ behavior: if what matters is only the material pay-off (that determines the inequity aversion), player 2’s choices should be the same in both situations. But, once again, the outcome of experiments are different. As indicated by the frequency numbers in figure 2 and 3, in the voluntary trust game 17 players 1 out of 27 move “down”, betting on the relationship with player 2. And, in 11 cases out of 17, more than 64% of the times, player 2 chooses to be trustworthy.

The experiments on the involuntary trust game show an opposite behavior. Here obviously the entire set of 27 player 1 moves down. But players 2 knows that, behind players 1 actions, there is no bet on the relationship. There is no trust, so there is no reciprocity: in fact, in 18 cases out of 27, more than 66% of the times, player 2 chooses the untrustworthy alternative.

This result should alert about a fact: pay-offs by themselves are not sufficient for the explanation of human behavior. Relations between human beings, the way they affect agents’ preferences and allow agents to express and interpret their intentions toward each other, play a role that is often decisive, and that stresses the importance – at least in principle – of the models that, in some ways, try to take them into account.

3.2. Explanations based on relation-affected agents' preferences

A first class of models considers the role of relations in trust-building by representing their influence on agents' preferences over the possible outcomes of their interactions: what is optimum for me depends on my reading of how the other players see me, i.e. on their attitudes toward me, or on the way they interpret, and eventually assess, what I do, i.e. on their esteem.

3.2.1 Attitudes

When a strategic interaction is setinside a human relation each agent behavior will be driven not only by material outcomes, but also by other people attitudes toward him. The strategy chosen will then have a reciprocal motive: we are more inclined to positively take into account of our opponent welfare when she shows an altruistic attitude toward us, while we tend to reciprocate spitefulness with spite.

This observation is at the core of Levine (1998), which has the declared objective of explaining apparently altruistic behavior in experiments on Ultimatum bargaining and Public Goods contribution games by taking into account reciprocal altruistic and spiteful behavior.

Levine (1998) assumes that agent i's utility, adjusted by taking account the other agents' utilities and attitudes, is

$$v_i = u_i + \sum_{j \neq i} \frac{a_i + \lambda a_j}{1 + \lambda} u_j \qquad (5)$$

where u_i is agent i's direct utility, $-1 < a_i < 1$ is agent i's coefficient of altruism and $0 \leq \lambda \leq 1$ is a coefficient meant to reflect agents' regard of other agents attitudes toward them.

According to equation (1), each agent adjusted utility depends not only on his own direct utility, but also on other agents' direct utilities: this relation may positive, $a_i > 0$, representing altruism, or negative, $a_i < 0$, representing spite. Most importantly for the sake of ourdiscussion, the way other agents' utilities affect agent i's adjusted utility depends on the other agents attitudes: my altruism (spite) is reinforced by other agents' altruistic (spiteful) attitude toward me, and weakened by their spitefulness (altruism). This latter effect goes through parameter λ: while $\lambda = 0$ would mean that agents behavior is affected only by pure altruism or spite, the larger is λ, the more the agent cares about other agents attitudes toward him.

It is further assumed that agents differ in their degree of altruism or spite, the agents involved in any given interaction are drawn from a common distribution known to everyone, while each agent's type is private information. Hence, when agents interact, they play a signaling game: a more altruistic move, as for example a demand for a smaller share of the amount to be divided in an Ultimatum bargaining game, will convey the message that the agent is altruistic andencourage the opponent to reciprocate, e.g. accepting the offer.

Levine claims that a model of pure altruism (or spite) cannot explain the experimental results he considers, since, for instance,the degree of spitefulness consistent with such a model and the observed rejection rate of respondents in Ultimatum bargaining experiments would correspond to much larger demands by proposers than the ones actually registered. Hence, according to Levine (1998), although experimental results contradicting the assumption of selfishness cannot be explained without considering some degree of altruism or spite, a complete account needs to allow for agents caring of each other's attitudes.

As for trust building, the consideration of attitudes may help to explain McCabe et al.'s puzzle, where pure altruism cannot. Indeed, a pure altruistic responder should behave exactly in the same way in the voluntary and involuntary version of the trust game; however, a responder who takes account of the proposer's attitudes would consider a trusting move as a signal of the proposer's altruism, value such an attitude and decide to reward trust with a trustworthy beahviour.

3.2.2 Esteem

Agents are often keen to gain from other people an approval of their behavior, i.e. in having other people's esteem. Ellingsen-Johanesson(2008) explicitly model this motive and discuss how the desire for social esteem may provide incentives for pro-social behavior. They focus on the possible motivational crowding out of relational motives coming from the use of control systems and pecuniary incentive schemes to motivate agents: explicit incentive schemes may be read by agents as a signal of their principal lack of esteem and so trigger non cooperative behavior.[13]

Yet, for the sake of our argument the main point of interest of Ellingsen-Johanesson(2008) is their recognition of the role of esteem in trust-building. They assume that each agent cares about the esteem of his counterparts and that the value of esteem depends on the agent's assessment on who provides it. Ellingsen-Johanesson(2008) assume that agent i's utility can be represented as

$$u_i = m_i + \theta_i m_j + \hat{\theta}_{ji} \qquad (6)$$

[13] In slightly different frameworks, similar crowding out effect are explored by Benabou-Tirole (2006) and Kreps (1997).

where m_i is the material payoff of agent i, θ_i is agent i's degree of altruism and $\hat{\theta}_{ji}$ is a measure of agent i's feeling of being esteemed, which in turn depends on agent j's salience $\sigma(\theta_j)$andesteem for agent i, θ_{ji} , as follows

$$\hat{\theta}_{ji} = E_{\theta_i}[\sigma(\theta_j)\theta_{ji}|\theta_i]. \qquad (7)$$

Note that the opponent's salience is a function of her type, which is represented by her degree of altruism. Hence, each agent's consideration for the opinions of his counterparts is lager as she is more pro-social in terms of altruism.

As in Levine (1998) above, the identity of each agent's counterpart affects his utility, the way he sees the value of the relation with her and, ultimately, the possibility of trust building; however, while for Levine this effect is positively or negatively synergetic with the agents' degree of altruism, Ellingsen and Johanesson see it as independent from and additively separable to other arguments of agents utility.

Ellingsen and Johanesson directly apply their setting to the analysis of a trust game, and, in particular, use it to interpret McCabe et al. (2003) observations.Certain values of the parameters, a trusting behavior of the proposer engaged in the voluntary game, and so actually choosing to trust, will be interpreted as a signal of a rather altruistic proposerby the responder, who will chose to be trustworthy to benefit from the (for him) highly relevant esteem of the proposer. This will not happen in the involuntary version of the game, since no choice is made and so no signal of his altruism is conveyed by the proposer.

3.3 Intentions and psychological games

The final step of the inclusion of a fully-fledged relational dimension in the representation of strategic interaction is due to the analysis of the so-called psychological games, first introduced by Geanakoplos et al.(1989). As noted by Dufwenberg and Kirchsteiger (2004), in psychological games "a player's payoff depends not only on what strategy profile is played, but possibly also onwhat are the player's beliefs about other players' strategic choices or beliefs"[14], since other players' intentions directly affect his utility. The same action is interpreted differently if I believe my opponent's intentions were to favor me and if I instead think she wanted to hurt me; and in turns this would depend on her beliefs on how I will play as well as on the set of alternatives she has.Hence, belief and intentions turn out to be deeply intertwined.

This add a psychological element to the material payoff agents obtain as the outcome of playing games. Rabin (1993) shows that these aspects become particularly relevant when reciprocal motives, a typically relational dimension, affect individual choice; for instance, as it is often natural, agents may prefer to be kind with people who have shown to be kind with them, and unkind with unkind people. Rabin assumes that agent i's expected utility is

$$U_i(a_i, b_j, c_i) = \pi_i(a_i, b_j) + \tilde{f}_j(b_j, c_i)[1 + f_i(a_i, b_j)] \qquad (8)$$

where $\pi_i(a_i, b_j)$ is his material payoff, depending on the strategy he chooses, a_i, and on his belief on the strategy chosen by player j, b_j; $\tilde{f}_j(b_j, c_i)$is player i's belief about how kind player j is to him, depending also on player i's belief about what player j believes player I's strategy is; $f_i(a_i, b_j)$ is player i's kindness to player j. As Rabin (1993) notes "If player i believes that

[14] Dufwenberg – Kirchsteiger (2004) p. 273.

player j is treatinghim badly-$\tilde{f}_j(b_j, c_i) < 0$ -then player i wishesto treat player j badly, by choosing an actiona_i such that$f_i(a_i, b_j)$ is low or negative. Ifplayer j is treating player i kindly, then$\tilde{f}_j(b_j, c_i)$ will be positive, and player i will wishto treat player j kindly."[15]

Dufwenberg – Kirchsteiger (2004) extends the analysis of Rabin (1993), which is developed for normal form games, to consider the dynamic structure of many strategic interaction. They derive a specific solution concept for dynamic games where players have a concern for relational dimensions: sequential reciprocity equilibrium. This solution concept allows to address the issue, classical in whole game theory, of optimality of equilibrium strategies out of the equilibrium path; with psychological games the effects of the sequential structure is particularly complex, since, as noted by the authors "As play unravels in a sequentialgame, a player who revises his beliefs may have to also revise beliefs about how kindother players are, since kindness depends on beliefs. Therefore, the way that the playeris affected by reciprocity concerns may differ dramatically between different parts of thegame tree."[16]

In such a more complex setting, however, the main intuition reported above for the simpler normal form setting survives.

As for our trust-building main concern, a trusting behavior may be interpreted as signaling a kind intention of the proposer towards the responder, particularly if the responder believe that the proposer's belief is that the responder will not play in a trustworthy manner. This interpretation does only make sense if the proposer has indeed a choice between trusting or not trusting the responder, as in the voluntary trust game in

[15] Rabin (1993) p. 1287.

[16]Dufwenberg-Kirchsteiger (2004) p. 271.

McCabe et al. (2003); if the trust game is involuntary, though, the lack of alternatives for the proposer deprives is observed action of any meaning as a signal of kind intentions, $\tilde{f}_j(b_j, c_i) = 0$, and so the responder is left with the only concern for material utility, behaving accordingly.

4. Evidence from the neuroeconomic research

Is there a deeper way to study the reasons for trust? According to supporters of the neuroeconomic approach, there is. And the answer is simple: we must search inside the brain.

This approach, taken at its face value, implies a vision of the world that is doubtful for a huge number of philosophical and methodological reasons, too long to be analyzed here. Are we just our brains? However, with some precautions, we can ask the brain about the reasons for trust, and a lot of scholars, by monitoring subjects playing a trust game with neuro-imaging techniques, actually did so.

In general, these studies seem to contain insights supporting the idea of social preferences, but we must stress a big caveat: with the present level of knowledge of our brain, which is really tentative, we should consider neuroeconomics results as insights, without trying to find there the final evidence supporting a theory against another.

A first paper we consider in this stream of literature is McCabe et al.'s[17]: these scholars monitor with the functional magnetic resonance 12 subjects playing a trust game: some with a human being as counterpart, other with a computer.

Comparing brain activations of the two groups McCabe et al. register that, in cooperative subjects belonging to the group playing with humans there is a stronger activation in the pre-

[17] McCabe et al. (2001).

frontal cortex and, interestingly, also in the anterior paracingulate areas, known to be connected with the theory of mind, i.e. the ability to understand and interpret other people. These activations do not occur in non cooperative subjects, suggesting that, when we choose to renounce an immediate monetary reward in order to reach a cooperative outcome, the comprehension of the counterpart is very important.

This can be seen as an insight of the pivotal role of beliefs about others' behavior; but they are important in any strategic interaction for the simple fact that this interaction exists. More specifically, those activations could be a sign of the importance of the understanding of others' intentions when a subject decides (consistently with Rabin, 1993), or of the influence of the well being of the counterparts on the decision making (as theorized by the hypothesis of altruism or inequity aversion). However, in general, McCabe et al.(2001) seems to support explanations of trust going beyond the vision of self-centered and welfare maximizing subjects, with an obvious connection with the behavioralist approach in McCabe et al.(2003).

Rilling et al. (2002) monitored with the functional magnetic resonance a couple of female subjects playing a repeated prisoner's dilemma against a computer or, alternatively, another human being. The result is interesting because, in their post-experiment interviews, subjects indicated the outcome where both cooperated as the most satisfactory from their individual point of view, even more satisfactory than the more rewarding (in material terms) outcome where the subject defects and the counterpart plays cooperatively[18].

This behavioral result is linked with a consistent brain activation: when the outcome is the double cooperation, researchers

[18] According to the interviewed subjects, this latter outcome provokes guilt or inhibits future cooperation.

observe a (comparatively) strong activation of the orbitofrontal cortex and of the anteroventral striatum, known to be part of the reward brain circuit.

The authors, as a consequence, support the idea that the achievement of a cooperative outcome is a fact rewarding *per se*, although in a different way if compared with monetary incentives.

For the sake of our argument a very interesting observation made by the authors is that cooperative playing is correlated with the striatal activation only if the two parties are humans, suggesting there is something special and not related to their material outcome in human interactions.

The reward circuit is important also in another paper written by Rilling et al. (2004). Here in fact the activation of the striatum, between other areas, are registered only when the partner who cooperates is a person, and not when it is a computer[19].

This fact seems to suggest that the cooperative response of a computer is not «*sufficiently rewarding to provoke a robust response in midbrain dopamine neurons; and that there is something particularly rewarding about positive social interactions with other people*»[20].

Cooperation (with a human being!) seems to be a rewarding activity.

[19] Subjects actually face every time a computer, but sometimes experimenters explain them that their partner behind the monitor is a person and not a computer.

[20] Rilling et al. (2004), p. 2543.

5. Implications for communities as coordination-promoting environments

In the previous sections we have considered examples where intentionally trusting behavior can play a crucial role to promote trustworthiness and to foster a well-functioning economic environment based on trust relationships. Then, we have surveyed a literature that tries to explain such a phenomenon in terms of signaling and recognition of attitudes and intentions that go beyond consequentialist assessments of the value of interactions in term of material outcomes. Finally we have noted that the peculiarity of the relational source of trust can be traced also in the neural activity going on inside the brains of the parties involved in some strategic interactions.

Having an explanation of trust that is not outcome-based, but relation-based has two implications for our consideration of communities as environment where human interactions take place and help us define their specific nature as institutional settings for solving coordination problems.

First, when trust is based on relations, it is the very recognition of the intrinsic value of such relations and the agents’ disregard for the maximization of their material payoff, which makes them choose virtuous behavior. This eventually leads to more efficient outcomes and indeed higher material payoffs. Hence, in environments where trust is important,disregarding material payoffs allows agents to enhance them.

Second, as trust is built on the recognition of the others’ intentions towards me, as my trustworthy attitudes are nurtured by the way the others behave with me, as, in some sense, my very personality is modeled by my relations with other people, then an environment favorable to trust is characterized by heavily personalized interactions.

At difference with institutional settings like markets and the State, which ignore the recognition of individual identities and base their operations on objective measure of the value of outcomes,[21]heavily personalized interactions where the parties involved concentrate on relations and disregard material outcomes characterize the functioning of communities and the particular way how they work and solve coordination problems. This deserves further attention, since settings where trust has a crucial role in enhancing efficiency are increasingly common and so communities, as coordination-promoting environments, far from being the relic of a distant past, may regain a central role in our near future.

[21] On this see also Merzoni (2010).

References

Andreoni, J. (1990), Impure altruism and donation to public goods: a theory of warm-glow giving. *The Economic Journal*, 100, pp. 464-477.

Benabou, R. and Tirole, J. (2006), Incentives and prosocial behavior.*The American Economic Review*, vol. 96, 5, pp. 1652-1678.

Benedict XVI (2009), *Encyclical letter "Caritas in Veritate"*, Città del Vaticano: Libreria Editrice Vaticana.

Bull, C. (1987), "The Existence of Self-Enforcing Implicit Contracts", *Quarterly Journal of Economics*, 102, 1 pp. 147-159.

Chami, R. and Fullenkamp, C. (2002), Trust and efficiency. *Journal of Banking and Finance*, 26. 1785-1809.

Chang, L.J., Smith, A., Dufwenberg, M. and Sanfey, A.G. (2011), Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70:560-572.

Colombo, F. and Merzoni, G. (2008), For how long to tie your hands? Stable relationships in an unstable environment. *Journal of Economics*, 95(2), pp. 93-120.

Colombo, F. and Merzoni, G. (2006), In praise of rigidity. The bright side of long term contract in repeated trust games. *Journal of Economic Behavior and Organization*, 59(3), pp.349-373.

Colombo, F. and Merzoni, G. (2004), "Reputazione, flessibilità e durata ottima dei contratti", *Economia Politica*, 21(2), pp. 233-268.

Dufwenberg, M. and Kirchsteiger, G. (2004), A theory of sequential reciprocity. *Games and Economic Behavior*, 47: 268-298.

Ellingsen, T. and Johannesson (2008), Pride and Prejudice: The Human Side of Incentive Theory, *American Economic Review*. 98, 3: 990-1008.

Fehr, E. and Schmidt, K.M. (1999), A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114: 817-868.

Fudenberg, D. and Maskin, E. (1986), The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, vol.54, pp. 532-554.

Geanakoplos, J., Pearce, D., Stacchetti, E. (1989), Psychological games and sequential rationality. *Games and Economic Behavior*, 1, 60-79.

Ghosh, P., Ray, D.(1996), Cooperation in community interaction without information flows. *Review of Economic Studies*, 63: 491-519.

Greif, A. (2006), The Birth of Impersonal Exchange: The Community Responsibility System and Impartial Justice, *Journal of Economic Perspectives*, 20(2), 221-236.

James, H. S. Jr. (2002), The trust paradox: a survey of economics inquiries into the nature of trust and trustworthiness. *Journal of Economic Behaviorand Organization*, 47, pp. 291-307.

Kranton, R.E. (1996), The formation of cooperative relationships, *Journal of Law, Economics. and Organization*, 12: 214-233.

Kreps, D. M. (1997), Intrinsic Motivation and Extrinsic Incentives, *The American Economic Review*, vol. 87, 2, Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association, pp. 359-364.

Kreps, D.M. (1990), Corporate culture and the economic theory. In: *Perspectives on positive political economy*. Alt J., Shepsle K. (eds), Cambridge University Press, Cambridge. pp. 90-143.

Kreps, D.M. and Wilson, R. (1982), Reputation and imperfect information. *Journal of Economic Theory,* 27(2): 253-279.

Kydd, A. (2000), Trust, reassurance, and cooperation. *International Organization*, 54(2): pp. 325-357.

Levine, D.K. (1998), Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics,* 1: 593-622.

McCabe, Kevin A., Houser, Daniel, Ryan, Lee, Smith, Vernon e Trouard, Theodore (2001), A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, USA, 98: 11832-11835.

McCabe, K., Rigdon M.L. and Smith, V.L. (2003), Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization,* 52:267-275.

Merzoni, G. (2010), Towards a positive vision of Global Governance, in Beretta, S. and Zoboli, R. (a cura di), *Global Governance in a Plural World*, pp. 137-142, Vita e Pensiero, Milano.

Milgrom P.R. and J. Roberts (1982), Predation, reputation, and entry deterrence. *Journal of Economic Theory*, 27(2): 280-312.

Rabin, M. (1993), Incorporating fairness into game theory and economics. *American Economic Review,* 83: 1281-1302.

Rilling, J.K., Gutman, D.A., Zeh, T.R. et al. (2002), A neural basis for social cooperation. *Neuron,* 35: 395-405.

Rilling, J.K., Sanfey, A.G., Aronson, J.A. et al. (2004), Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport*, 15: 2539-2543.

Sen, A. (1977), Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs,* 6(4), 317-344.

Tabarrok, A. (1994), A Survey, Critique, and New Defense of Term Limits. *Cato Journal,*14 (2): 333-50.

Trombetta, F. (2012), Behind Trust. How alternative explanations of trust shape our vision of human beings and communities. In Beretta S. and M. Maggioni (eds.) *"The whole breadth of reason. Rethinking Economics and Politics"*, Venezia, Marcianum Press.

Printed by
Gi&Gi srl - Triuggio (MB)
December 2012

ISBN 978-88-343-2527-8

9 788834 325278