DISCUSSION PAPER SERIES

IZA DP No. 1641

Neuroeconomic Foundations of Trust and Social Preferences

Ernst Fehr
Urs Fischbacher
Michael Kosfeld

June 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Neuroeconomic Foundations of Trust and Social Preferences

Ernst Fehr
University of Zurich
and IZA Bonn

Urs Fischbacher
University of Zurich

Michael Kosfeld
University of Zurich

Discussion Paper No. 1641
June 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1641
June 2005

ABSTRACT

Neuroeconomic Foundations of Trust and Social Preferences

This paper discusses recent neuroeconomic evidence related to other-regarding behaviors and the decision to trust in other people's other-regarding behavior. This evidence supports the view that people derive nonpecuniary utility (i) from mutual cooperation in social dilemma (SD) games and (ii) from punishing unfair behavior. Thus, mutual cooperation and the punishment of free riders in SD games is not irrational, but better understood as rational behavior of people with corresponding social preferences. We also report the results of a recent study that examines the impact of the neuropeptide Oxytocin (OT) on trusting and trustworthy behavior in a sequential SD. Animal studies have identified Oxytocin as a hormone that induces prosocial approach behavior, suggesting that it may also affect prosocial behavior in humans. Indeed, the study shows that subjects given Oxytocin exhibit much more trusting behavior, suggesting that OT has a direct impact on certain aspects of subjects' social preferences. Interestingly, however, although Oxytocin affects trusting behavior, it has no effect on subjects' trustworthiness.

JEL Classification: A13, C90

Keywords: social preferences, foundations of trust, neuroeconomic

Corresponding author:

Ernst Fehr
Institute for Empirical Research in Economics
University of Zürich
Blümlisalpstrasse 10
8006 Zürich
Switzerland
Email: efehr@iew.unizh.ch

Neuroeconomics merges methods from neuroscience and economics to better understand how the human brain generates decisions in economic and social contexts. Neuroeconomics is part of the general quest for microfoundations – in this case, the microfoundation of individual decision-making in social contexts. The economic model of individual decision-making is based on three concepts – the action set, preferences, and beliefs. Economists assume that an individual will choose his preferred action for a given set of available actions and a given belief about the states of the world and the other players' actions. Neuroeconomics provides a microfoundation for individual beliefs, preferences, and behavior; it does so by examining the brain processes associated with the formation of beliefs, the perception of the action set, and the actual choice. Moreover, since the set of available actions can be framed in different ways and different frames of the same action set sometimes elicit different behaviors, neuroeconomics may also contribute to a deeper understanding of framing effects.

This paper discusses recent neuroeconomic evidence related to other-regarding (nonselfish) behaviors and the decision to trust in other people's nonselfish behavior. As we will show, this evidence supports the view that people derive nonpecuniary utility (i) from mutual cooperation in social dilemma (SD) games and (ii) from punishing unfair behavior in these games. Thus, mutual cooperation that takes place despite strong free riding incentives, and the punishment of free riders in SD games is not irrational, but better understood as rational behavior of people with corresponding social preferences. Finally, we report the results of a recent study that examines the impact of the neuropeptide Oxytocin (OT) on trusting and trustworthy behavior in a sequential SD. Animal studies have identified Oxytocin as a hormone that induces prosocial approach behavior, suggesting that it may also affect prosocial behavior in humans. Indeed, the study shows that subjects given Oxytocin exhibit much more trusting behavior, despite the fact that OT does *not* change their explicit beliefs about others' behavior. Thus, it seems that OT has a direct impact on certain aspects of subjects' social preferences. Interestingly, however, although Oxytocin affects trusting behavior, it has no effect on subjects' trustworthiness.

At the general level, economic theory has been reluctant to assume anything specific about human preferences, except for the fact that they satisfy the axioms of revealed preference theory. In practice, however, economists often make the strong assumption that individual preferences are exclusively self-regarding. However, a large body of evidence (Colin F. Camerer, 2003, Ernst Fehr and Urs Fischbacher, 2003) now suggests that a substantial share of the people exhibits social preferences and that an even larger share typically shows trust in the existence of these preferences. Sequential SD games that are played only once are a neat vehicle for demonstrating the behavioral relevance of social preferences. This game can be described as follows: there are two players, A and B, each of whom has an initial endowment of $10. First, player A decides whether to keep his endowment or to send it to player B. Then player B observes A's action and decides whether to keep her endowment or to send it to A. The experimenter doubles each transfer payment, i.e., both players are better off if they transfer their endowments than if they both keep them. This situation mimics a sequential economic exchange in the absence of contract enforcement institutions. B has a strong incentive to keep her endowment regardless of whether A transferred or not; if A anticipates this behavior, however, he has little reason to transfer his endowment. A mutually beneficial exchange can only take place if A trusts B and if B behaves nonselfishly by transferring her endowment.

Literally hundreds of experiments, with stake levels up to several months' income, have confirmed that a large share of subjects in the role of player B reciprocates player A's trust and that an even larger share of subjects in the role of A trusts B (Camerer, 2003, Fehr and Fischbacher, 2003). Moreover, if we add a third decision stage to this game by giving player A the option of rewarding or punishing B at a cost to himself, many A players reward those B players who reciprocated their trust and punish those B players who did not do so (Fehr et al., 1997). Why do we observe these strong deviations from the predictions of the standard model? What are the driving forces behind the decision to trust, to reciprocate trust, and to punish non-reciprocation? To what extent do emotional factors play a role here, and how do they interact with the human ability for rational deliberation? Can reciprocation and

punishment best be modeled by assuming that they are preferred behaviors, or are these behaviors just a reflection of subjects' bounded rationality, as some authors have claimed (Larry Samuelson, forthcoming). In the following we will show that neuroeconomic studies can help answer these questions.

I. Neural evidence for a taste for the punishment of unfair behavior

In a recent paper, (Dominique DeQuervain et al., 2004) combined a two-player sequential SD game with Positron Emission Tomography (PET) imaging of subjects' brains; PET is one method for measuring the activation of different brain areas. In this game, player A had the opportunity of punishing player B after observing whether B reciprocated A's trust by assigning up to 20 punishment points to B. The monetary consequences of the punishment depended on the treatment conditions and will be explained below. Player A's brain was scanned with PET when A received information about B's decision and during his decision about whether to punish B. The main purpose of this study was to examine what happens in A's brain when B abuses his trust. The study was led by the hypothesis that player A has a taste for punishing B if B intentionally abuses his trust. Models of social preferences and reciprocity developed in the past 5-8 years suggest this hypothesis. If it is correct, we should observe the activation of reward-related brain areas during and after A's decision to punish. This activation of reward-related areas could be due to the satisfaction a player anticipates if he decides to punish player B for unfair behavior.

An important prerequisite for this study was the existence of neuroscientific knowledge about the key components of the brain's reward circuits. Fortunately, many recent studies have shown that an area in the midbrain, the striatum, is a key part of reward-related neural circuits. Single neuron recording in non-human primates (Wolfram Schultz, 2000) and neuroimaging studies with humans using money as a reward medium (John P. O'Doherty, 2004) indicate clearly that the striatum is a key part of reward-related neural circuits. Moreover, if A punishes B because he anticipates deriving satisfaction from punishing, one

should observe activation predominantly in those reward-related brain areas that are associated with goal-directed behavior. There is strong evidence from single neuron recording in non-human primates (Schultz, 2000) that the dorsal striatum is crucial for the integration of reward information and behavioral information in the sense of a goal-directed mechanism. Several recent neuroimaging studies support the view that the dorsal striatum is implicated in processing rewards resulting from a decision (O'Doherty, 2004). The fact that the dorsal striatum also responds to expected monetary gains in a parametric way is of particular interest from an economic viewpoint: if subjects successfully complete a task that generates monetary rewards, the activation in the dorsal striatum increases as the expected monetary gain grows. Thus, if A's dorsal striatum is activated when punishing B, we have a strong piece of evidence indicating that punishment is rewarding.

To examine the activation of striatal areas during the decision to punish, subjects' brains were mainly scanned in those SD trials in which B abused A's trust.[1] In the condition termed "costly" (C), the punishment was costly for both A and B. Every punishment point assigned to B cost experimental $1 for A and reduced B's payoff by experimental $2. In the condition termed "free" (F), punishment was not costly for A. Every punishment point assigned to B cost nothing for A while B's payoff was reduced by $2. In a third condition, which we call "symbolic" (S), punishment had only a symbolic (and no pecuniary) value. Every punishment point assigned to B cost neither A nor B anything. Thus, A could not reduce B's payoff in this condition.

The hypothesis that punishment is rewarding predicts that the contrast F – S will show the activation of reward related brain areas after A's trust has been abused[2]. The rationale behind this prediction is that A is likely to have a desire to punish B both in the F and the S

[1] Player A played the game seven times with seven different subjects in the role of player B.

[2] Brain activations in neuroimaging are always measured in one condition relative to another condition. Thus, the F – S contrast provides information about those brain areas that are more highly activated in the F relative to the S condition.

condition because B intentionally abused A's trust, but A cannot really hurt B in the S condition. Thus, the purely symbolic punishment in the S condition is unlikely to be satisfactory because the desire to punish the defector cannot be fulfilled effectively, and in the unlikely case that symbolic punishment is satisfactory, it is predicted to be less so than punishment in the F condition.

The F – S contrast is ideal for examining the satisfying aspects of effective punishment because – except for the difference in the opportunity to punish effectively – everything else remains constant across conditions. However, punishment should also generate satisfaction from an economic viewpoint if it is costly. If there is indeed a taste for punishing defectors and if subjects actually do punish because the cost of punishing is not too high, the act of punishment is analogous to buying a good. Rational subjects buy the good as long as the marginal costs are below the marginal benefits. Thus, an economic model based on a taste for punishment predicts that punishment in the C condition should also be experienced as satisfactory, implying that reward related areas will also be activated in the C – S condition.

Questionnaire and behavioral evidence indicates that player A indeed had a strong desire to punish the defectors. In fact, almost all subjects punished maximally in the F condition, while most subjects still punished in the C condition, albeit at a lower level. This reduction in the level of punishment makes sense because punishment was costly in the C condition. Most importantly, however, the dorsal striatum was strongly activated in both the F – S contrast and the C – S contrast, indicating that punishment is experienced as satisfactory. Moreover, the data show that those subjects in the C condition who exhibit higher activations in the dorsal striatum also punish more. This positive correlation can be interpreted in two ways: first, the higher level of punishment could cause the increased activation of the dorsal striatum, i.e., the higher satisfaction. Second, the greater anticipated satisfaction from punishing could cause the higher level of punishment, i.e., the activation in the striatum reflects – in this view – the anticipated satisfaction from punishing. It would be reassuring from an economic viewpoint if

the second interpretation were the correct one because it relies on the idea that the anticipated rewards from punishing drive the punishment decision.

Both the popular press and neuroscience often claim that emotions are an overpowering force that inhibit rational behavior. Emotions like anger are known to play an important role in punishing defectors (Ernst Fehr and Simon Gächter, 2002). Thus, it is theoretically possible that anger overrides rationality and induces subjects to punish the defector "blindly". However, if it could be shown that, while anger is important in these situations, subjects decide rationally about how much they want to punish a defector, one could argue in favor of an economic approach. According to this approach, emotions like anger have a motivational impact because they change the hedonic consequences of different actions; yet, given the hedonic consequences of different actions, subjects decide rationally by weighing the costs and benefits of the actions.

DeQuervain et al. (2004) provide two pieces of evidence in favor of an economic approach. The first piece of evidence is related to the C – F contrast. Subjects face a nontrivial trade off in the C condition between the benefits and costs of punishing, whereas the decision is much simpler in the F condition because no costs exist. Thus, certain parts of the prefrontal cortex (Brodmann areas 10 and 11), which are known to be involved in integrating the benefits and costs for the purpose of decision-making, should be more strongly activated in the C condition than in the F condition. This is in fact the case. The second piece of evidence is based on the observation that most subjects punished maximally in the F condition. Thus, the differences in striatum activation across these subjects cannot be due to different levels of punishment. However, if different striatum activations reflect differences in the anticipated satisfaction from punishment, those subjects who exhibit higher striatum activations in the F condition (although they punish at the same maximal level) should be willing to spend more money on punishment in the C condition. The data again supports this prediction.

II. The Rewards of Mutual Cooperation

Models of social preferences and reciprocity are based on the idea that a substantial share of people prefers mutual cooperation over unilateral defection in a SD. These models are based on behavioral evidence indicating that many second movers in a sequential SD reciprocate player A's trust. However, skeptics (Samuelson, forthcoming) have argued that self-interest might also explain behavior that is seemingly consistent with social preferences, if subjects treat one-shot games as if they were repeated games involving the possibility for future punishment.

Neuroeconomic evidence may be able to resolve this debate. One possibility is to show that mutual cooperation yields higher utility than unilateral defection. However, computing the brain contrast between mutual cooperation and unilateral defection is not ideal because any difference in brain activation could be due the fact that the scanned player cooperates in one situation and defects in the other. The measured activations might have nothing to do with the special hedonic consequences of the mutual cooperation outcome but might be caused by the behavioral difference. There is, however, another way to solve this problem. One of the first neuroeconomic studies (James K Rilling et al., 2002) reports activations in the striatum when subjects experience mutual cooperation with a human partner compared to mutual cooperation with a computer partner. Thus, despite the fact that the subject's monetary gain is identical in both situations, mutual cooperation with a human partner seems to be experienced as a more rewarding outcome, indicating that extra benefits from mutual cooperation extend beyond the mere monetary gain. Unfortunately, however, the Rilling et al. study is based on a repeated SD. A repeated dilemma game involves a host of other confounding influences which might shed doubt on the interpretation of brain activations in terms of social preferences. A recent paper based on a one-shot sequential SD has solved this problem (Rilling et al., 2004). The authors show again that the mutual cooperation outcome with a human partner generates higher striatum activation than the mutual cooperation outcome with

a computer partner.[3] Moreover, the mutual cooperation outcome with a human partner also generates higher activations than does earning the same amount of money in a trivial individual decision-making task. A further study showing that the mere viewing of faces of people who previously cooperated in a SD activates reward related areas (Tania Singer et al., 2004) indicates the special hedonic qualities of mutual cooperation. This result suggests that people derive more utility from interactions with cooperative people not just because they can earn more money in these interactions but because these interactions are rewarding per se.

III. The Neurobiology of Trust

Neuroeconomics is not restricted to the use of imaging techniques. A recent study (Michael Kosfeld et al., 2005) examined the neurobiological basis of trusting and trustworthy behavior in a sequential SD. Animal studies on the neurobiology of certain forms of prosocial behavior (Thomas R. Insel and Larry J. Young, 2001) suggest the hypothesis that the neuropeptide Oxytocin (OT) might provide a biological basis for trusting behavior in humans. OT facilitates maternal behavior and pair bonding in different species. Specifically, OT seems both to permit animals to overcome their natural avoidance of proximity and to inhibit defensive behavior, thereby facilitating approach and biparental care.

Kosfeld et al. examined the hypothesis that OT facilitates trust and trustworthiness by comparing behavior in an SD in a group of subjects that received OT with that of subjects in a control group that received placebo. Their results indeed show that subjects with OT exhibit significantly more trusting behavior; however, OT does not affect player B's trustworthiness. More specifically, the percentage of players A who trusts maximally in a SD increases from

[3] In the Rilling et al studies the ventral and not the dorsal striatum is activated. This makes sense because the brain contrasts were measured after subjects who cooperated were informed whether their (computer or human) opponent also cooperated. Thus, the contrast measures the experienced and not the anticipated extra benefits of mutual cooperation with a human partner.

21% to 45% whereas the transfers of player B remain constant between the OT and the placebo group. Kosfeld et al. also measure how OT affects subjects' calmness, wakefulness, and mood, to control for the possibility that such side effects are responsible for the effect of OT on trusting behavior. However, a sizeable and significant effect of OT on trust remains, even after controlling for these indirect effects. The direct effect of OT increases the probability of trusting maximally by 20 percentage points.

An interesting question is whether OT operates at the level of subjects' beliefs about others' trustworthiness or whether it operates at the level of subjects' preferences. Recent research (Iris Bohnet and Richard Zeckhauser, 2004) shows that the decision to trust is not shaped by risk aversion, but by exploitation aversion, i.e. by the fear of being fooled by player B. Thus, in the same way OT overcomes the animals' natural tendency to avoid others, OT might also overcome the "natural" fear of being exploited by others in a SD. The results of the Kosfeld et al study show that OT does not affect subjects' beliefs about player B's trustworthiness. Although subjects with OT and the placebo hold the same beliefs, subjects with OT make themselves more vulnerable to exploitation by sending more money to B. Thus, it seems that subjects with OT are more willing to take the risk of being exploited, suggesting that OT affects subjects' exploitation aversion. This effect is insofar interesting as economists usually assume that preferences are stable. However, if preferences are based on actual or anticipated emotions, they may be much less stable than typically assumed because emotions are often transient. Moreover, as the OT study suggests, preferences and the underlying affective states can be deliberately shaped over short periods of time by administering the "right" substance.

IV. Conclusions

We have discussed recent neuroeconomic evidence on social preferences and trust in this paper. However, the implications of neuroeconomic studies go far beyond these areas of research (Camerer et al., 2005). Neuroeconomic studies are likely to provide insights into how

the human reward system is linked to decision making in intertemporal choice (Samuel M. McClure et al., 2004) and risk (Hans C. Breiter et al., 2001) and how affect and cognition interact to generate decisions (Alan G. Sanfey et al., 2003). Such studies enable us to go beyond the prevailing "as if" approach in economics by uncovering the neural mechanisms behind individual decisions. In the long term, it may well be that neuroeconomic insights fundamentally change the current preferences and beliefs approach that prevails in economics. For example, economics assumes that an individual's beliefs about the other player's actions do not depend on the individual's preferences. This assumption precludes motivated belief formation, making it difficult to understand questions of religious beliefs, ideology, aggression towards outgroup members, the structure and the content of political and economic advertising campaigns that appeal to people's emotions, etc. Perhaps, however, there are neural and affective mechanisms which allow preferences to influence beliefs and vice versa. Reputation formation may provide an example: if we are cheated in a social exchange, we have a strong affective reaction that shapes our preferences towards the opponent (Singer et al., 2004 study). This affective reaction may also shape our beliefs about the opponent's future behaviors. We would be surprised if such affect guided belief formation obeyed the rules of Bayesian updating.

References

Bohnet, Iris and Zeckhauser, Richard. "Trust, Risk and Betrayal." Journal of Economic Behavior & Organization, 2004, 55(4) pp. 467-484.

Breiter, Hans C.; Aharon, Itzhak; Kahneman, Daniel; Dale, Anders and Shizgal, Peter "Functional Imaging of Neural Responses to Expectancy and Experience of Monetary Gains and Losses." Neuron, 2001, 30(2), pp. 619-39.

Camerer, Colin F. Behavioral Game Theory - Experiments in Strategic Interaction. Princeton, New Jersey: Princeton University Press, 2003.

Camerer, Colin; Loewenstein, George and Prelec, Drazen. "Neuroeconomics: How Neuroscience Can Inform Economics." Journal of Economic Literature, 2005, forthcoming.

DeQuervain, Dominique; Fischbacher, Urs; Treyer, Valerie; Schellhammer, Melanie; Schnyder, Ulrich; Buck, Alfred and Fehr, Ernst. "The Neural Basis of Altruistic Punishment." Science, 2004, 305, pp. 1254-58.

Fehr, Ernst and Fischbacher, Urs "The Nature of Human Altruism." Nature, 2003, 425, pp. 785-91.

Fehr, Ernst and Gächter, Simon "Altruistic Punishment in Humans." Nature, 2002, 415, pp. 137-40.

Fehr, Ernst; Gächter, Simon and Kirchsteiger, Georg "Reciprocity as a Contract Enforcement Device: Experimental Evidence." Econometrica, 1997, 65(4), pp. 833-60.

Insel, Thomas R. and Young, Larry J. "The Neurobiology of Attachment." Nature Reviews Neuroscience, 2001, 2(2), pp. 129-36.

Kosfeld, Michael; Heinrichs, Markus; Zak, Paul, Urs Fischbacher and Fehr, Ernst. "Oxytocin Increases Trust in Humans." Nature, June 2, 2005, pp. 673-676.

McClure, Samuel M.; Laibson, David I.; Loewenstein, George and Cohen, Jonathan D. "Separate Neural Systems Value Immediate and Delayed Monetary Rewards." Science, 2004, 306, pp. 503-7.

O'Doherty, John P. "Reward Representations and Reward-Related Learning in the Human Brain: Insights from Neuroimaging." Current Opinion in Neurobiology, 2004, 14(6), pp. 769-76.

Rilling, James K.; Sanfey, Alan G.; Aronson, Jessica A.; Nystrom, Leigh E. and Cohen, Jonathan D. "Opposing Bold Responses to Reciprocated and Unreciprocated Altruism in Putative Reward Pathways." Neuroreport, 2004, 15(16), pp. 2539-243.

Rilling, James K.; Gutman, David A.; Zeh, Thorsten R.; Pagnoni, Giuseppe; Berns, Gregory S. and Kilts, Clinton D. "A Neural Basis for Social Cooperation." Neuron, 2002, 35, pp. 395-405.

Samuelson, Larry. "Foundations of Human Sociality: A Review Essay." Journal of Economic Literature, forthcoming.

Sanfey, Alan G.; Rilling, James K.; Aronson, Jessica A.; Nystrom, Leigh E. and Cohen, Jonathan D. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." Science, 2003, 300, pp. 1755-58.

Schultz, Wolfram "Multiple Reward Signals in the Brain." Nature Reviews Neuroscience, 2000, 1(3), pp. 199-207.

Singer, T.; Kiebel, S. J.; Winston, J. S.; Kaube, H.; Dolan, R. J. and Frith, C. D. "Brain Responses to the Acquired Moral Status of Faces." Neuron, 2004, 41(4), pp. 653-62.