Year: 2007

Social neuroeconomics: the neural circuitry of social preferences

Fehr, Ernst; Camerer, Colin F

Abstract: Combining the methods of neuroscience and economics generates powerful tools for studying the brain processes behind human social interaction. We argue that hedonic interpretations of theories of social preferences provide a useful framework that generates interesting predictions and helps interpret brain activations involved in altruistic, fair and trusting behaviors. These behaviours are consistently associated with activation in reward-related brain areas such as the striatum and with prefrontal activity implicated in cognitive control, the processing of emotions, and integration of benefits and costs, consistent with resolution of a conflict between self-interest and other-regarding motives.

Social neuroeconomics:

The neural circuitry of social preferences

Ernst Fehr**
University of Zurich, Institute for Empirical Research in Economics
efehr@iew.uzh.ch

Colin F. Camerer
California Institute of Technology
camerer@hss.caltech.edu

Abstract:
Combining the methods of neuroscience and economics generates powerful tools for studying the brain processes behind human social interaction. We argue that hedonic interpretations of theories of social preferences provide a useful framework that generates interesting predictions and helps interpret brain activations involved in altruistic, fair and trusting behaviors. These behaviours are consistently associated with activation in reward-related brain areas such as the striatum and with prefrontal activity implicated in cognitive control, the processing of emotions, and integration of benefits and costs, consistent with resolution of a conflict between self-interest and other-regarding motives.

(93 words)

As in behaviorist psychology, the long-standing tradition in economic theory has been to treat preferences and beliefs as impossible or difficult to observe directly; instead, their effects were thought to be only revealed by direct choices. The emerging neuroeconomic approach [1-4] rejects the premise of unobservability, and seeks a microfoundation of social and economic activity in neural circuitry, using fMRI, TMS, pharmacological interventions, and other techniques. The neuroeconomic approach hopes to unify mechanistic, mathematical and behavioral (choice-based) measures and constructs. Byproducts of such an ambitious program might include better understanding of individual differences and development over the human lifecycle (including disorders and expertise), insights into the effects of direct and social learning, empirical discipline of evolutionary modeling, and advice for how economic rules and institutions can be designed so that people react to rules in a socially-efficient way.

In this review we discuss the neural circuitry involved in altruistic, fair and trusting behaviors. Traditional economic analyses generally makes the simplifying assumption that people are exclusively self-regarding, but there is now a large body of experimental evidence [5, 6] indicating that many people exhibit social preferences, i.e, their preferred choices are based on a positive or negative concern for the welfare of others, and on what other players believe about them. Social neuroeconomics tries to understand the brain processes that govern these regular deviations from purely self-interested behavior. Social neuroeconomics combines the tools of social cognitive neuroscience [7-9] with well-structured tasks taken from economic theory (Box 1 and Table 1). These tasks come equipped with benchmark theoretical predictions about rational play and social efficiency of outcomes, which are useful for interpreting the results and cumulating regularity across studies.

Roughly speaking, there are two viewpoints in economic and biological sciences about why pro-social behaviors occur. One view is that behavior in one-shot anonymous games indicates a reflexive behavior that is highly adapted for repeated interactions in which immediate pro-social behavior earns *future* benefits. In this view, pro-sociality in one-shot games results from bounds on rationality in fully responding to changes in economic structure

[7]. The other view is that pro-social behavior reflects robust social preferences for treating others generously or reciprocally, and those preferences are similar to preferences for other kinds of primary and secondary rewards (the "reward interpretation", see Table 1).

This paper focuses on recent studies that provide some tentative evidence of neural activity which might eventually distinguish these two broad viewpoints. Some of the questions asked in social neuroeconomics include: What are the neural networks and the motivational forces behind charitable donations, rejections in ultimatum bargaining games, punishment of greedy behavior in third party punishment games or decisions to trust and to reciprocate trust altruistically? To what extent do emotional and rewarding factors play a role here, and how do they interact with the human ability for rational deliberation?

Theories of social preferences and the brain

Economic theories of social preferences [10-17] model the motivational forces driving the deviations from economic self-interest in a precise way. In theories of reciprocal fairness [11, 16, 17], for example, players are assumed to positively value kind intentions, and to negatively value hostile intentions, of other players. Thus, if player A reduces B's payoff to his own benefit, a reciprocal player B will punish A, whereas if bad luck led to a redistribution of income from B to A, a reciprocal player B will not punish [18]. In contrast, if a player is motivated by inequity aversion [13], i.e. a dislike of unequal outcomes per se, then bad luck will induce player B to take action to redistribute income [19]. Likewise, some theories postulate an individual's desire to increase the economic welfare of the group they belong to [14, 15], to experience a warm glow from altruistic giving to worthy causes [10], or to maintain a positive social image [20].

Theories of social preferences are based on the concept of decision utility [21]. A decision utility is a numerical measure which is thought to underly observed behaviour (e.g., the action chosen from a set of choices is inferred to have the highest numerical decision utility). Decision utility can, in principle, be distinguished from (a) experienced utility, which

is the hedonic experience associated with the consumption of a good or an event, and from (b) anticipated utility, which is the anticipation of experienced utility at the time of decision-making. One of the central questions in social neuroeconomics, which recent studies address, is how the brain constructs decision utilities when a person's behaviour reflects their own rewards but is also governed by competing motives such as warm glow altruism, reciprocity or inequity aversion. This general question implies a host of other important questions such as: Is self-interest a primary motive that needs to be constrained by appropriate inhibitory machinery? If so, which brain circuitry is involved in these inhibitory processes? To what extent are these processes related to emotion regulation? Are deviations from economic self-interest partly governed by positive hedonic consequences associated with non-selfish behaviours and, if so, are these complex social rewards represented in the striatum and the OFC like primary or monetary rewards [22, 23] or do they rely on different neural circuitries?

Social preferences and reward circuitry

Theories of reciprocity and inequity aversion imply that subjects prefer the mutual cooperation outcome over the unilateral defection outcome in the canonical prisoners' dilemma game although unilateral defection leads to a higher economic payoff (Table 1abc). Although these theories do not make assumptions about the hedonic processes associated with fairness related behaviors (because they rely on inferred decision utilities), a plausible interpretation of these theories is that subjects in fact derive higher hedonic value from the mutual cooperation outcome [24]. Indeed, there is questionnaire evidence (M. Kosfeld, E. Fehr and J. Weibull, unpublished) supporting the view that mutual cooperation in social exchanges has special subjective value, beyond the value that is associated with monetary earnings (Table 1d). Therefore, a natural question is whether we can find neural traces of the special reward value of the mutual cooperation outcome. Two neuroimaging studies [25, 26] report activation in the ventral striatum, when subjects experience mutual cooperation with a human partner compared to mutual cooperation with a computer partner. Despite the fact, that

the monetary gain is identical in both situations, mutual cooperation with a human partner is associated with higher striatal activity, consistent with the reward hypothesis, given substantial evidence from other studies with primary and secondary rewards that the striatum is activated by anticipated reward.

Social preference theories also predict that subjects prefer to punish unfair behavior such as defection in public good and PD games (Table 1e) because leaving an unfair act unpunished is associated with higher disutility than bearing the cost of punishing an unfair act. In this view, it is natural to hypothesize that the act of punishing defection involves higher activation of reward circuitry. A study using PET [27] examined this hypothesis in the context of a PD game with a punishment opportunity similar to the one described in Table 1e. This study showed that the dorsal striatum (caudate nucleus) is strongly activated in the contrast between a real punishment condition (in which the assignment of punishment points hurt the defector in economic terms) and a symbolic punishment condition (in which the assignment of punishment points did not reduce the defector's economic payoff). In another study [28] subjects first played a sequential PD with (confederate) fair and unfair opponents. The focal subjects were then scanned (using fMRI) when a slight pain – an electrical shock – was administered either to themselves or to confederate partners who behaved fairly or unfairly. Both men and women exhibited empathic responses in anterior cingulate and anterior insula when the fair partner received pain. However, only men report a higher desire for revenge against unfair partners, and also exhibit activation in the nucleus accumbens (NAcc) and orbitofrontal cortex (OFC) when unfair partners are shocked. Male revenge-desire ratings across subjects are also correlated with the estimate of NAcc activity, consistent with the view that there is reward value in observing the punishment of unfair partners.

Further evidence that decisions involving social preferences are associated with activity in reward circuitry comes from fMRI studies of charitable donations [29, 30] and reaction to offers in a take-it-or-leave-it ultimatum bargaining game [31]. Ventral tegmental (VTA) and striatal areas are both activated by receiving money and non-costly donations, indicating that 'giving has its own reward' [29] Across subjects, those who made more costly donations also

had more activity in the striatum. Decisions to donate, whether costly or not, activate the subgenual area, which is densely connected with mesolimbic dopaminergic and serotonergic pathways and is implicated in social attachment mechanisms and in regulating the release of the neuromodulator oxytocin via the anterior hypothalamus. In [30] subjects are in two conditions – a forced donation and a voluntary donation condition. In the former, subjects passively observed that money is transferred to their account or to the charities account. In the voluntary condition the subjects could decide whether to accept such monetary transfers. Both in the forced and the voluntary condition subjects reported higher satisfaction if they themselves receive more money or if the charity receives more money (controlling for the subject's cost of this transfer). Moreover, in both conditions activations in dorsal and ventral striatum are positively correlated with the money that goes to the charity and to the subjects themselves. Finally, a recent ultimatum game study [31] provides evidence suggesting that the fairness of a bargaining offer – controlling for the absolute size of the monetary gain – is associated with activations in the ventral striatum. The same dollar bargaining offer of, say $5, elicits higher striatal activation if the offer represents a fair share (say 50%) of the amount which is being bargained over, compared to when that dollar offer represents a small share (say, only 15%).

The activations observed in these studies and several others indicate that social rewards commonly activate the dorsal or ventral striatum (see Figure 1A). There is substantial overlap between these areas of activation and activation observed in studies of reinforcement learning or anticipated money reward (Figure 1B). This overlap is consistent with the hypothesis that social preferences are similar to preferences for one's own rewards in terms of neural activation.

Do activations in reward circuitry predict choices?

The above evidence is consistent with the view that costly pro-social acts of charitable donation and punishment of unfair behaviors are both rewarding. However, the hedonic

interpretation of social preference theories also implies that such acts occur *because* they are rewarding. Evidence for causality is also important for moving from correlation to causality, and because some studies suggest monetary gains *and* losses might both be fully processed by a unitary system, centered on the striatum [32, 33]. If this unitary activity holds more generally for positively and negatively valenced goods, the mere fact that studies show higher fMRI BOLD responses for costly altruistic acts may indicate a costly experience rather than a rewarding one. But if it could be shown that higher activations in the striatum *imply* a higher willingness to act altruistically, the case for the reward interpretation would be strengthened considerably (since it is implausible to observe this relation between striatum activation and altruistic acts if striatum activation represents the cost of the act rather than its reward value).

Neuroimaging data do not allow causal inferences, but it is possible to move towards causality by predicting from neural activity in one treatment to choice behaviour in another treatment ("out of treatment" forecasting). For example, in [27] individual differences in caudate nucleus activation when punishment is costless for the punisher predicts how much individuals actually pay for punishment when it is not costless. Likewise, in [30] individual differences in striatal activity in the condition in which donations are forced predicts subjects' willingness to donate money to charities in the condition in which donations are voluntary (Figure 2). These results further support the reward interpretation of social preferences, which in turns provides support for the hypothesis of a common neural currency of socially-preferred and other primary and secondary rewards [34].

The role of the prefrontal cortex (PFC) in decisions involving social preferences

If people have social preferences the brain must compare social motives and economic self-interest and resolve conflict between them. Several studies indicate that the prefrontal cortex plays a decisive role in such conflict resolution. For example, in the contrast between costly punishment condition and costless punishment of players who behaved unfairly, the ventromedial PFC (BA 10, 11) has been implicated [27], consistent with the hypothesis that

this area is involved in the integration of separate benefits and costs in the pursuit of behavioral goals [35]. The crucial role of VMPFC in decisions involving social preferences is also supported by evidence [36] that subjects with brain lesions in VMPFC reject ultimatum game offers more frequently, suggesting that the cost of rejecting positive offers has less weight in the decision process if VMPFC is impaired. Finally, in charitable donations [29] the contrast between altruistic decisions involving costs and no costs also activated the VMPFC (BA 10, 11, 32) and the dorsal anterior cingulate cortex (ACC). Since the ACC is thought to play a key role in conflict monitoring [37], activity in this region is consistent with the existence of a trade off between self-interest and pro-social motives.

The role of the VMPFC in decisions involving costly altruism is also interesting because of related activation in this region in other studies. The VMPFC is involved in emotional processing and moral judgment [38, 39] and in integrating the value of consumer products and their prices [40]. Lesions to VMPFC are also associated with poor choices in various situations [41, 42] which require integrating costs and benefits. These studies and those on the VMPFC's role in expression of social preference suggest a general role in integrating emotional feelings about costs and benefits, regardless of whether these choices involve economic consumption goods or "non-economic" goods such as the subjective value of acting altruistically.

Two neuroimaging studies [31, 43] suggest that the dorsolateral (DLPFC) and ventrolateral (VLPFC) prefrontal cortex are also likely to play an important role in the processing of decisions involving social preferences. These studies examined the neural circuitry involved in the recipient's behavior in an ultimatum game where the rejection of low positive offers involves a motivational conflict between fairness and economic self-interest. The first study [43] reports activation of bilateral DLPFC, bilateral anterior insula (AI) as well ACC in the contrast between "unfair>fair" offers. In addition, the higher the activation of right AI the more likely a subject rejects an unfair offer suggesting that AI activation may be related to the degree of emotional resentment of unfair offers. Due to the role of ACC in conflict monitoring [37], the activation of ACC in this task may reflect the motivational

conflict between fairness and self-interest when facing unfair offers. Finally, the DLPFC activation may represent the cognitive control of the emotional impulse to reject unfair offers. The second study [31] also finds that AI is more active during rejected trials. In addition, the right VLPFC is more activated (relative to a resting baseline) when unfair offers are accepted, which may indicate that this region down-regulates the resentment associated with unfair offers.

The interpretation that DLPFC activity represents the cognitive control of the impulse to reject implies that interfering or disrupting DLPFC activity reduces the control of the impulse and should, thus, increase the rejection rate. Knoch et al. [44] examined this hypothesis by reducing the activation in right and left DLPFC with low-frequency transcranial magnetic stimulation (TMS). Surprisingly, the study found that TMS of right DLPFC increases the *acceptance* rate of unfair offers relative to a placebo stimulation (from 9% to 44%) while TMS of left DLPFC did not affect behavior significantly. This finding suggests that right DLPFC is *causally* involved in controlling the impulse that pushes subjects towards accepting unfair offers, i.e., with controlling economic self-interest. Interestingly, the disruption of right DLPFC only affects subjects' fairness related behaviors but not their fairness judgments, i.e., they still judge low offers as very unfair, but they nevertheless accept them more frequently and more quickly.

Trust, reputation and social preferences

Social preferences models predict that trusting other individuals, by making investments that may not be repaid, is not just a decision involving monetary risk. Reciprocal and inequity averse subjects derive a special disutility from betrayal of trust, along with the associated economic loss, which is consistent with behavioral studies [45] indicating a pure aversion to social betrayal. The first evidence that the brain distinguishes between social trust and monetary risk-taking comes from [46] who infused the synthetic neuropeptide oxytocin intranasally to players in a trust game. OT-infused players were more trusting than a placebo

control group, although OT-players' beliefs about the chances of being repaid were not higher and OT did not affect risk-taking in a pure risk condition. Thus, OT seems to limit the fear of betrayal in social interactions, consistent with animal evidence that it inhibits defensive behavior and facilitates maternal behavior and pair bonding [47]. The hypothesis that the fear-of-betrayal-reducing effect of oxytocin might be due to a reduced activation of the amygdala is consistent with a study [48] showing that OT dampens amygdala activity and its connections to the brainstem if subjects view emotionally arousing pictures. Amygdala involvement has been shown to occur in assessing the trustworthiness of faces [49, 50] and the processing of ambiguous events [51], which both have social implications.

Since trust decisions are also likely to involve perspective taking, they should also activate areas implicated in theory of mind tasks such as the paracingulate cortex and the posterior superior temporal sulcus (pSTS) [52]. One of the earliest neuroeconomic studies [53] reports activation of the paracingulate in a trust game when subjects play against another person compared to a computerized opponent. Another study found that pSTS is activated simply by showing the faces of intentional cooperators compared to nonintentional agents [28].

In repeated trust games a player learns about his opponent's choices so that the opponent acquires a reputation. In game theory this reputation is defined as the subjective probability that the opponent is the type of player who prefers to reciprocate trust. In this approach players' preferences and their subjective beliefs are distinct concepts and a rational player's beliefs are not colored by his preferences or his emotions towards the opponent. It is interesting to examine whether the brain also makes this distinction, i.e., whether the neural networks involved in hedonic preferences and emotional processing are distinct from the networks involved in assessing the opponent's reputation or whether there is substantial overlap in these neural networks.

Preliminary evidence suggests that the latter is likely to be true. In one study [54] players faced a series of cooperative and noncooperative opponents in a sequential PD game.

The authors found that simply displaying the faces of cooperative partners (relative to neutral faces) in a subsequent gender assessment task activated striatal and emotion related areas such as the amygdala, the insula and the putamen. This suggests that a trustworthy person's face automatically trigger emotions and reward expectations, as if simply seeing another person's face activates its representation as a future exchange value.

The importance of the striatum in learning the opponent's trustworthiness has been demonstrated by two other studies. In [55] the activity in caudate nucleus signals whether the other player reciprocates an earlier move. A further study using the same trust-game paradigm showed specializations in the cingulate for encoding decisions of others and oneself [56]. In [57] trustors repeatedly face three partners whose (fictional) profiles make them seem morally good, bad or neutral (instilling a prior belief about trustworthiness). By design, all three fictional partners repay in the trust game with the same frequency. During the outcome phase the caudate nucleus activates more strongly for repayment outcomes from the neutral partner, but not from the other partners, presumably because the neutral partner represents unpredictable outcomes and there is more to learn.

Conclusions and Research Directions

In this review we showed how theories of social preferences guided the conduct of neuroeconomic experiments and the interpretation of the resulting brain data. One emerging theme of the studies reviewed above is that social reward activates circuitry that overlaps, to a surprising degree, to circuitry which anticipates and represents other types of rewards. These studies reinforce the idea that social preferences for donating money, rejecting unfair offers, trusting others, and punishing those who violate norms, are genuine expressions of preference. The social rewards are traded off with subjects' economic self-interest and the dorsolateral and the ventromedial prefrontal cortex are likely to be crucially involved in the balancing of competing rewards. These processes can also be altered by treatments like oxytocin infusion

and TMS disruption, actually changing behavior in ways that are consistent with hypotheses derived from fMRI.

Economics and other social sciences may benefit from social neuroeconomics because of the potentially unifying force of neural data for choice based approaches. Most of economics assumes for example, that beliefs about other people's behavior are based on a rational assessment of the available information; they are neither directly affected by preferences nor are they disturbed by emotions. The results in [54] suggest however, that beliefs about other people's trustworthiness may be strongly affected by reward and emotion circuitry. This study also creates a paradigm for exploring how rapidly reputations increase and decrease, and whether good reputations built up by particular players in racial, gender or class-based social groups might generalize to new players in those same groups. This kind of group-based neural generalization of expected reward could be important in understanding the powerful role of social networks and physical cues in labor market discrimination, for example.

Another example illustrating the potential of neuroeconomics comes from recalling the economist's concepts of risk preferences, time preferences (the willingness to postpone consumption) and social preferences. Most economic analyses treat these as separate types of preference. However, suppose all three types of preference share some common neural circuitry for controlling automatic emotional impulses (based on fear, temptation, and selfishness, respectively), by integrating all the costs and benefits of choices. If there is such a shared basis of preference, it is important for economists and other social scientists to understand, and this can be best established by data from imaging, lesion patient studies, and other neuroscientific measures. For example, two studies by Knoch et al. [44, 58] show that disruption of right, but not left, DLPFC with TMS increases both risk-taking in choice tasks and self-interested choices in ultimatum games. A natural next step is to investigate whether disruption of right DLPFC also makes people more impatient.

Future studies should exploit the wide range of tools available to neuroscientists, using multiple measures at the same time (e.g., hormone measurement or TMS and fMRI), combined with the parametric value and predictions about more complex games. Focusing on neural bases also opens up research directions which are, perhaps surprisingly, unexplored in economics. For example, the standard analysis in game theory assumes that players are "in equilibrium" (i.e., they correctly anticipate, from introspection or learning, what others are planning to do). In equilibrium analysis there is no room for differences in strategic skill. Yet one study [59] found that skill varied systematically across players (as measured by belief accuracy and differences in earnings). When making choices, more skilled players showed more activity in the ventral striatum and precuneus (as if they anticipated higher money rewards) and less skilled players showed more activity in the insula (as if they were feeling discomfort from strategic uncertainty).

These directions might eventually provide a biological basis for a mathematical characterization of social exchange that is rooted in neural details but can also make predictions about activity in strategic interaction and market trading, and about how behavior can change when causally manipulated by pharmacology, TMS, and other tools.

References

1. Glimcher, P.W., and Rustichini, A. (2004) Neuroeconomics: the consilience of brain and decision. *Science* 306, 447-452
2. Camerer, C., Loewenstein, G., and Prelec, D. (2005) Neuroeconomics: How Neuroscience Can Inform Economics. *Journal of Economic Literature* 43 1, 9-64
3. Fehr, E., Fischbacher, U., and Kosfeld, M. (2005) Neuroeconomic foundations of trust and social preferences: Initial evidence. *American Economic Review* 95, 346-351
4. Sanfey, A.G., Loewenstein, G., McClure, S.M., and Cohen, J.D. (2006) Neuroeconomics: cross-currents in research on decision-making. *Trends in Cognitive Sciences*
5. Fehr, E., and Fischbacher, U. (2003) The nature of human altruism. *Nature* 425, 785-791
6. Camerer, C.F. (2003) *Behavioral Game Theory - Experiments in Strategic Interaction.* Princeton University Press
7. Adolphs, R. (2003) Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience* 4, 165-178
8. Blakemore, S.J., Winston, J., and Frith, U. (2004) Social cognitive neuroscience: where are we heading? *Trends Cogn Sci* 8, 216-222

9.   Lieberman, M.D. (2007) Social cognitive neuroscience: a review of core processes. *Annu Rev Psychol* 58, 259-289

10.   Andreoni, J. (1990) Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving. *Economic Journal* 100, 464-477

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

12.   Levine, D.K. (1998) Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1, 593-622

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

14.   van Lange, P.A.M. (1999) The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation. *Journal of Personality and Social Psychology* 77, 337-349

15.   Charness, G., and Rabin, M. (2002) Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117, 817-869

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

17.   Falk, A., and Fischbacher, U. (2006) A theory of reciprocity. *Games and Economic Behavior* 54, 293-315

18.   Blount, S. (1995) When social outcomes aren't fair - the effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* 63, 131-144

19.   Dawes, C.T.*, et al.* (2007) Egalitarian motives in humans. *Nature* 446, 794-796

20.   Rotemberg, J. (2007) Minimally acceptable altruism and the ultimatum game. *Journal of Economic Behavior & Organization* in press

21.   Kahneman, D. (1994) New Challenges to the Rationality Assumption. *Journal of Institutional and Theoretical Economics-Zeitschrift Fur Die Gesamte Staatswissenschaft* 150, 18-36

22.   O'Doherty J, P. (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr Opin Neurobiol* 14, 769-776

23.   Knutson, B., and Cooper, J.C. (2005) Functional magnetic resonance imaging of reward prediction. *Current Opinion in Neurology* 18, 411-417

24.   Thibaut, J.W., and Kelley, H.H. (1959) *The Social Psychology of Groups*. Wiley

25.   Rilling, J.K.*, et al.* (2002) A neural basis for social cooperation. *Neuron* 35, 395-405

26.   Rilling, J.K.*, et al.* (2004) Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* 15, 2539-2243

27.   DeQuervain, D.*, et al.* (2004) The neural basis of altruistic punishment. *Science* 305, 1254-1258

28.   Singer, T.*, et al.* (2006) Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439, 466-469

29.   Moll, J.*, et al.* (2006) Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America* 103, 15623-15628

30.   Harbaugh, W.T., Mayr, U., and Burghart, D.R. (2007) Neural Responses to Taxation and Voluntry Giving Reveal Motives for Charitable Donations. *Science* 316, 1622-1625

31.   Golnaz, T., Satpute, A.B., and Lieberman, M.D. (2007) The Sunny Side of Fairness_ Preference for Fairness Activates Reward Circuitry. *Psychological Science* in press

32.   Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007) The neural basis of loss aversion in decision-making under risk. *Science* 315, 515-518

33.   Seymour, B.*, et al.* (2007) Differential encoding of losses and gains in the human striatum. *J Neurosci* 27, 4826-4831

34.   Montague, P.R., and Berns, G.S. (2002) Neural economics and the biological substrates of valuation. *Neuron* 36, 265-284

35.   Ramnani, N., and Owen, A.M. (2004) Anterior prefrontal cortex: insights into funtion from anatomy and neuroimaging. *Nature Reviews Neuroscience* 5, 184-194

36.   Koenigs, M., and Tranel, D. (2007) Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *J Neurosci* 27, 951-956

37.   Botvinick, M.M.*, et al.* (2001) Conflict monitoring and cognitive control. *Psychological Review* 108, 624-652

38.   Moll, J.*, et al.* (2005) Opinion: the neural basis of human moral cognition. *Nat Rev Neurosci* 6, 799-809

39.   Koenigs, M.*, et al.* (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 908-911

40.   Knutson, B.*, et al.* (2007) Neural predictors of purchases. *Neuron* 53, 147-156

41.   Damasio, A.R. (1995) *Descartes' Error: Emotion, Reason and the Human Brain.* Hayrer Collins

42.   Bechara, A., Damasio, H., Tranel, D., and Damasio, A.R. (1997) Deciding advantageously before knowing the advantageous strategy. *Science* 275, 1293-1295

43.   Sanfey, A.G.*, et al.* (2003) The neural basis of economic decision-making in the ultimatum game. *Science* 300, 1755-1758

44.   Knoch, D.*, et al.* (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314, 829-832

45.   Bohnet, I., and Zeckhauser, R. (2004) Trust, risk and betrayal. *Journal of Economic Behavior & Organization* 55, 467-484

46.   Kosfeld, M.*, et al.* (2005) Oxytocin increases trust in humans. *Nature* 435, 673-676

47.   Insel, T.R., and Young, L.J. (2001) The neurobiology of attachment. *Nat Rev Neurosci* 2, 129-136

48.   Kirsch, P.*, et al.* (2005) Oxytocin modulates neural circuitry for social cognition and fear in humans. *J Neurosci* 25, 11489-11493

49.   Winston, J.S., Strange, B.A., O'Doherty, J., and Dolan, R.J. (2002) Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience* 5, 277-283

50.   Adolphs, R.*, et al.* (2005) A mechanism for impaired fear recognition after amygdala damage. *Nature* 433, 68-72

51.   Hsu, M.*, et al.* (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680-1683

52.   Frith, U., and Frith, C.D. (2003) Development and neurophysiology of mentalizing. *Philos Trans R Soc Lond B Biol Sci* 358, 459-473

53.   McCabe, K.*, et al.* (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America* 98, 11832-11835

54.   Singer, T.*, et al.* (2004) Brain responses to the acquired moral status of faces. *Neuron* 41, 653-662

55.   King-Casas, B.*, et al.* (2005) Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78-83

56.   Tomlin, D.*, et al.* (2006) Agent-Specific Responses in the Cingulate Cortex During Economic Exchanges. *Science* 312, 1047-1050

57.   Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8, 1611-1618

58.   Knoch, D.*, et al.* (2006) Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. *Journal of Neuroscience* 26, 6469-6472

59.   Bhatt, M., and Camerer, C.F. (2005) Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior* 52, 424-459

60.   Mikula, G. (1972) Reward Allocation in Dyads Regarding Varied Performance Ratio. *Zeitschrift Fur Sozialpsychologie* 3, 126-133

61.   Kahneman, D., Knetsch, J.L., and Thaler, R. (1986) Fairness as a constraint on profit seeking: entitlements in the market. *American Economic Review* 76, 728-741

62.   Eckel, C., and Grossman, P. (1996) Altruism in Anonymous Dictator Games. *Games and Economic Behavior* 16, 181-191

63.   Dana, J., Cain, D.M., and Dawes, R.M. (2006) What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes* 100, 193-201

64.   Güth, W., Schmittberger, R., and Schwarze, B. (1982) An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization* 3, 367-388

65.   Henrich, J.*, et al.* (2001) In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review* 91, 73-78

66.   Fehr, E., and Fischbacher, U. (2004) Third-party punishment and social norms. *Evolution and Human Behavior* 25, 63-87

67.   Camerer, C., and Weigelt, K. (1988) Experimental Tests of a Sequential Equilibrium Reputation Model. *Econometrica* 56, 1-36

68.   Berg, J., Dickhaut, J., and McCabe, K. (1995) Trust, Reciprocity and Social History. *Games and Economic Behavior* 10, 122-142

69.   Ledyard, J. (1995) Public Goods: A Survey of Experimental Research. In *Handbook of Experimental Economics* (Kagel, J., and Roth, A., eds), 111-194, Princeton University Press

70.   Fehr, E., and Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415, 137-140

74.   Preuschoff P. K., Bossaerts P., Quartz, S. R. (2006) Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron* 51, 381-90.

Box 1: Measuring social preferences with games

Experimental games enable measurement of how much of their own economic payoff players are willing to sacrifice to increase or decrease the payoffs of others [5, 6]. These games are typically played one-shot, with anonymous partners and with real monetary stakes. They provide a solid collection of empirical regularities from which the study of neural activity can proceed.

In a "dictator" game [60, 61] one player – the dictator – is given a sum of money which he can allocate between herself and another player – the recipient. Dictator allocations are a mixture of 50% offers and 0% offers (i.e., the dictator keeps everything), and a few offers in between 50 and 0%, but the allocations are sensitive to details of how the game is described [6], the dictator's knowledge of who the recipient is [62] and whether the recipient knows that she is part of a dictator game [63].

In an ultimatum game, the recipient can reject the proposed allocation [64]. If she rejects it both players receive nothing. Rejections are evidence of negative reciprocity [11], the motive to punish players who have treated you unfairly, or inequity aversion[13], a distaste for unfair outcomes. The strength of these motives can be measured by how much a recipient loses by rejecting a proposed allocation. Offers of less than 20% are rejected about half the time; Proposers seem to anticipate these rejections and consequently offer around 40% on average. Cross-cultural studies, however, show that across small-scale societies ultimatum offers are more generous when cooperative activity and market trade are more common [65].

In a third party punishment game two players, the dictator A and the recipient B, participate in a dictator game [66]. A third player, the potential punisher C, observes how much A gives to B; then C can spend a proportion of his endowment on punishing A. This game measures to what extent "impartial" and "unaffected" third parties are willing to stick up for other players at their own expense, enforcing a sharing norm by punishing greedy dictators. Between 50 and 60% of the third parties punish selfish deviations from the equal split suggesting that giving less than 50% in the dictator game violates a fairness norm. In principle, the third party punishment option can be used to measure economic willingness to

punish violation of *any* social norm (e.g., a violation of etiquette, breaking a taboo, or making a linguistic slur).

In a trust game [67, 68], two players, A and B each have an initial endowment. First, A decides whether to keep his endowment or to send it to B. Then B observes A's action and decides whether to keep the amount she received or share some if it with A. The experimenter doubles or triples A's transfer, so that both players are better off collectively if A transfers money and B sends back a sufficient amount. This situation mimics a sequential economic exchange in the absence of contract enforcement institutions. B has a strong incentive to keep all the money and repay none to A; if A anticipates this behavior, however, there is little reason to transfer so a chance for mutual gain is lost. Empirically, A's invest about half of their endowment and B's repay about as much as player A invested [6]. Player A's invest less than they do in risky choices with chance outcomes however, which indicates a pure aversion to social betrayal and inequality [45].

In a public goods game [5, 6, 69], which represents a generalization of the Prisoners' Dilemma game (see Table 1), players have a token endowment they can simultaneously invest in any proportion to a private project or a public project. Investment into the public project maximizes the aggregate earnings of the group but each individual can gain more from investing into the private rather than the public project. Typically, players begin by investing half their tokens on average (many invest either all or none). When the game is repeated over time, with feedback at the end of each decision period, investments decline until only a small fraction (about 10%) of the players invest anything. When players are allowed to also punish other players at a cost to themselves, many players who invested punish the players who did not invest, which encourages investment and leads players close to the efficient solution in which everyone invests [70].

_____

Table 1a. Representation of prisoners' dilemma in terms of material payoffs

	Cooperate (C)	Defect (D)
Cooperate (C)	4, 4	0, 5
Defect (D)	5, 0	1, 1

Table 1b. Utility representation of prisoners' dilemma if players are inequity averse

	Cooperate (C)	Defect (D)
Cooperate (C)	4, 4	$0 - 5\alpha$, $5 - 5\beta$
Defect (D)	$5 - 5\beta$, $0 - 5\alpha$	1, 1

Table 1c. Utility representation of prisoners' dilemma if players are inequity averse with parameters $\alpha = 1$ and $\beta = 0.5$

	Cooperate (C)	Defect (D)
Cooperate (C)	4, 4	–5, 2.5
Defect (D)	2.5, –5	1, 1

Table 1d. Player A's actual average ranking of outcomes in the prisoners' dilemma

|	Cooperate (C)	Defect (D)
Cooperate (C)	3.2	1.6
Defect (D)	2.7	2.5

Table 1e. Utility representation if an inequity averse player A ($\alpha = 1$, $\beta = 0.5$) punishes a selfish player B for defection

|	Cooperate (C)	Defect (D)
Cooperate (C)	4, 4	–2, 0
Defect (D)	2.5, 0	1, 1

Subjective payoffs after player A punished a selfish player B for defection

Table 1. In the prisoners' dilemma (PD) each of 2 players makes one of 2 choices: cooperate or defect. The PD can either be played sequentially, where one player moves before the other, or simultaneously, where both players choose without knowing what the other one does. **a**. The PD can be thought of as a paradigm for any kind of exchange that is not enforced by third parties. Assume for instance that both player A and B possess a good that they value with 1; however they both value the other player's good with 4 so that exchanging the goods is beneficial for both. In this case cooperation means to send the good to the other player while defection means keeping one's good. In case of mutual cooperation (exchange) both players receive an economic payoff of 4, in case of mutual defection both receive 1. If player A (the row player) defects and player B (the column player) cooperates, A receives an economic payoff of 5 (the value of the own good plus the value of the other's good) while B receives 0. See lower left corner of Table 1a; the first (red) number in each cell is A's payoff, the second (blue) number is B's payoff. Regardless of what B does, it is always in the self-interest of A to defect. The same holds for B. Thus, the unique equilibrium outcome in the game is (defect, defect). However, if both players defect they are worse off than if both cooperate, hence the dilemma. **b**. If both players have a strong enough preference for reciprocity [11] or if they are inequity averse [13] their subjective preferences transform the game. In case of inequity aversion a player suffers from receiving less than the other with parameter α_i (envy), and also from receiving more than the other with parameter β_i (compassion). An inequity averse player i's subjective payoff U_i is a function of her own economic payoff x_i and of the payoff differences $(x_j - x_i)$ between the two players: $U_i(x) = x_i - \alpha_i(x_j - x_i)$ if player i is worse off than player j $(x_j - x_i \geq 0)$, and $U_i(x) = x_i - \beta_i(x_i - x_j)$ if player i is better off than player j $(x_i - x_j \geq 0)$. Inequity aversion makes unilateral defection less attractive by reducing the subjective payoff from 5 to $(5 - 5\beta)$ while being the victim of the other player's unilateral defection is particularly painful because it reduces the subjective payoff from 0 to -5α. **c**. Representation of subjective payoffs for the special case of $\alpha = 1$ and $\beta = 0.5$. Here we can see that an inequity averse player A values the mutual cooperation outcome with 4 whereas the unilateral defection outcome is only valued with 2.5. Thus, if a player believes that the other player cooperates the player subjectively prefers to also cooperate, rending mutual cooperation an equilibrium. However, mutual defection also remains an equilibrium: if an inequity averse player believes that the other player defects he or she prefers to defect, too. **d**. Subjects' ordinal ranking of the four outcomes in a PD game similar to the one in Table 1a (M. Kosfeld, E. Fehr and J. Weibull, unpublished). Subjects had to assign a number between 1 and 4 to each cell of a prisoners' dilemma. 4 represented the most highly valued outcome and 1 the

least valued outcome. The numbers in the table represent the outcome from the perspective of player A only. On average, subjects valued the mutual cooperation outcome highest (3.2) while the other player's unilateral defection receives the lowest average valuation (1.6). The hypothesized ordinal ranking of outcomes by an inequity averse player A in Table 1c is identical to the actual ordinal ranking of outcomes in Table 1d. In particular, the average player A (Table 1d) prefers cooperation if she believes the opponent cooperates, and prefers defection if she believes the opponent defects. **e**. Suppose now that the prisoners' dilemma is played sequentially – A first chooses, then B chooses, knowing what A did. Then, A observes whether B cooperated and defected. After this A has the chance to punish B at a cost to himself. An inequity averse player A will never punish B in case of mutual cooperation or mutual defection because the players' payoffs are equal in these cases. However, he may be willing to punish if B defected unilaterally. Suppose, for example, that A can spend 1 money unit on punishment such that B's income is reduced by 5 money units, causing a material payoff distribution of $(-1, 0)$ and a utility of $U_A = -1 - \alpha(1 - 0) = -1 - \alpha$. If A does not punish B the material payoff distribution is $(0, 5)$ and A's utility is $U_A = 0 - \alpha(5 - 0) = 0 - 5\alpha$. In Table 1e we assumed $\alpha = 1$, implying that A's subjective payoff is -5 if he does not punish while if he punishes it is -2. But for any $\alpha \geq \frac{1}{4}$ player A is subjectively better of if he punishes B. Note also that in this case only mutual cooperation is part of an equilibrium because a rational player B anticipates that A punishes and hence, it is not in the self-interest of B to defect.
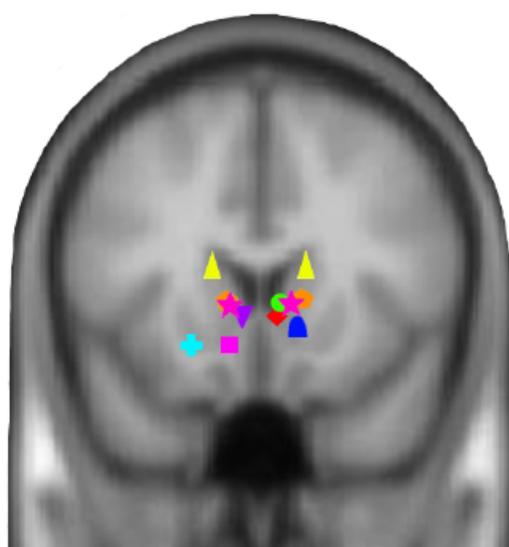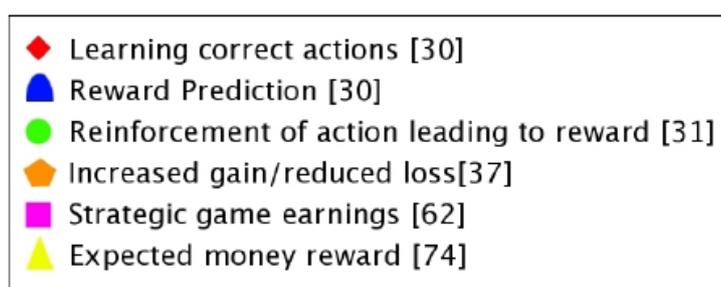
(A)
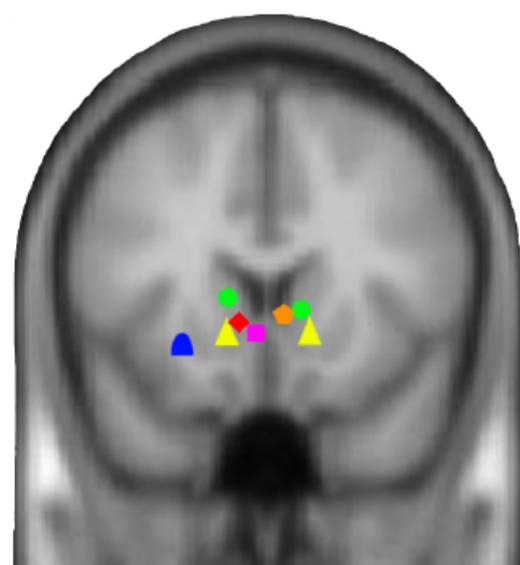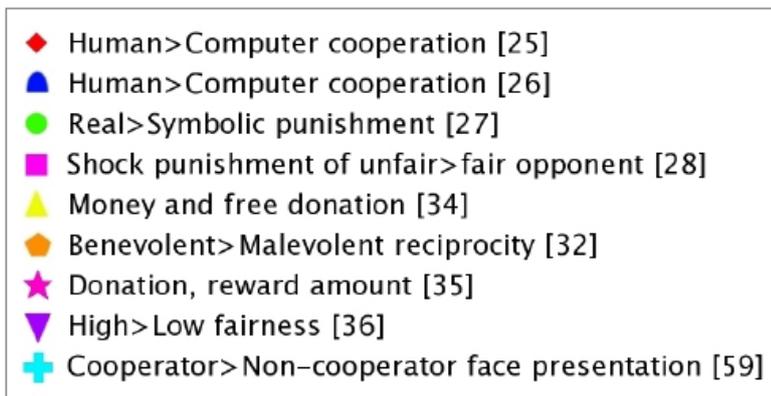
(B)

Figure 1. Parallelism of rewards for oneself and for others: Brain areas commonly activated in (A) nine studies of social reward and (B) a sample of six studies of learning and anticipated monetary reward. (All are projected onto Tailarach coordinate y=15; voxels of peak activation in original studies range from y=4 to y=24).
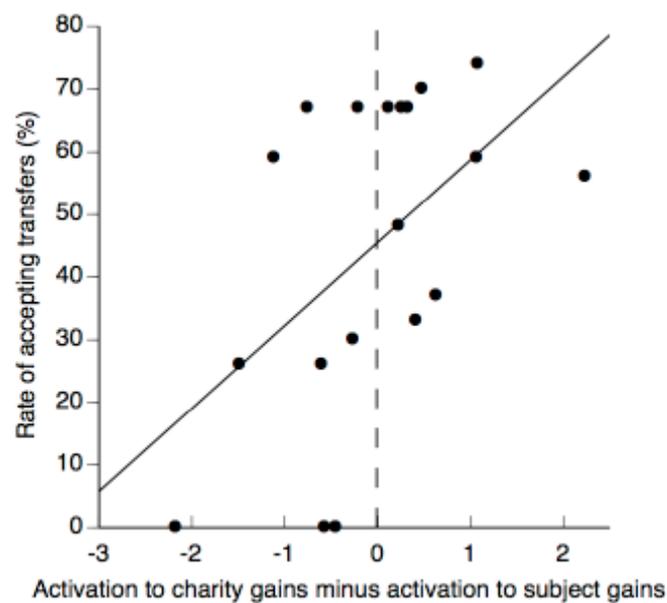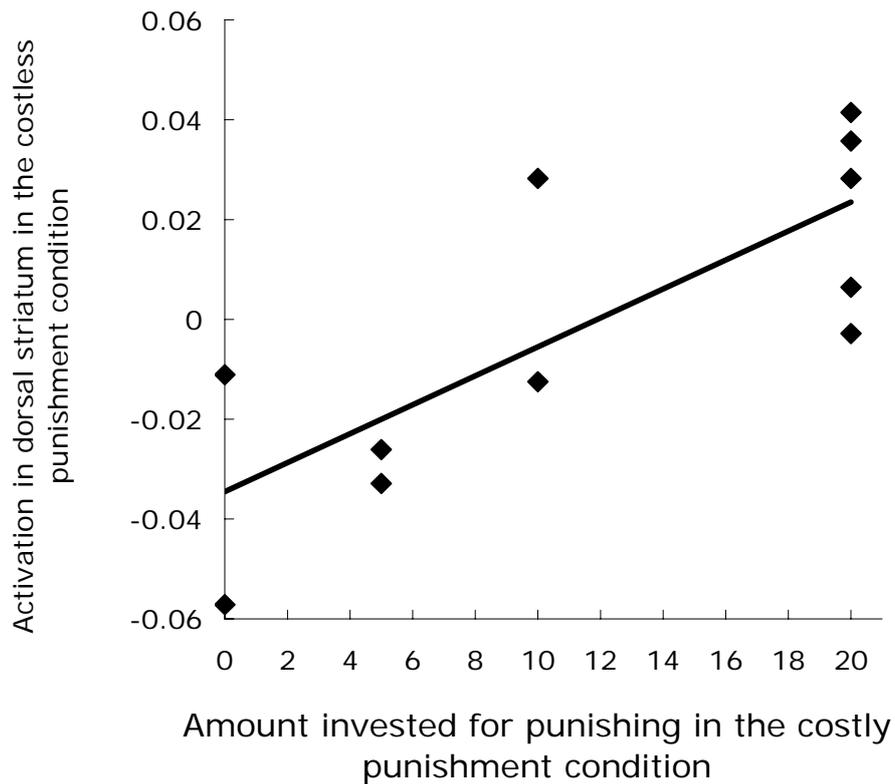
Figure 2. Predicting the frequency of subjects' altruistic choices with brain activations "out of treatment". **a**. In [27] subjects punished maximally when punishment of defection was costless but they still exhibit different activations in the dorsal striatum (y-axis). These activations predict how much subjects spend on punishment when it is costly (x-axis). **b**. In [30], the individual differences in striatal activations between the condition where subjects

just observe the charity receiving money (i.e., they cannot decide) and the condition where they themselves receive money can be interpreted as a measure of hedonic altruism (x-axis). Consistent with this interpretation, subjects with a higher difference in these activations accept transfers to charities more often when they have the freedom to do so (y-axis).

Open questions:

1. How does the brain relate decision utility to anticipated and experienced utility?

2. What is the relationship between moral emotions such as guilt and shame and moral behaviour? How can we measure and induce these moral emotions, and which behaviors are caused be them?

3. Which behaviors, emotions and neural mechanisms of human prosociality are unique, and which do we share with other primates?

4. What gene clusters are reliably linked to economic aspects of social behaviour?

5. Which computational models of brain activity predict both neural events and social behaviour correctly?

6. How are social disorders like autism, Asperger and Williams syndrome, social phobias, and anti-social personality disorder linked to differences in neural activation?

7. How is group membership perceived and processed neurally, and what are its implications?

8. How does extensive experience (e.g. experts in negotiation, or professional poker players) affect neural bases of social exchange?

9. What are the neural correlates of skill in strategic interaction? How does training and experience affect neural activity?

10. Many social institutions use social network connections (e.g., personal referrals) or agents (e.g., in bargaining); how does neural activity in these cases differ from personal interaction?