Measuring Trust in Social Neuroeconomics: a Tutorial

Jan B. Engelmann

A brief outline of the field of Neuroeconomics

Neuroeconomics is a relatively recent research area that is based on the amalgamation of various disciplines, including experimental economics, experimental psychology and cognitive neuroscience. The combination of methods from these fields has allowed researchers to design experiments investigating brain function during decision-making. Recent experiments in neuroeconomics have significantly furthered our understanding of the neural mechanisms involved in economic and social decisions. On the one hand, recent advances in neuroeconomics have lend support to the validity of economic models by demonstrating that parameters central to economic models, such as decision-utility (e.g. Knutson et al., 2005) and reinforcement learning (Montague et al., 2004) are represented in the brain. On the other hand, they have offered refinements of well-established views held by traditional economics, such as the rational-agent model. Against the assumption that economic behavior is purely rational, a wealth of converging evidence from behavioral and neuroimaging data underline the role of emotions in decision-making in both financial (McClure et al., 2004), as well as social (Sanfey et al., 2003) settings.

Since the current paper is directed at an audience outside of the fields that constitute neuroeconomics, we briefly review some of the core methodologies employed within neuroeconomics and related disciplines to investigate the neural mechanisms of decision-making before we discuss recent advances on the neural correlates of social preferences using the example of trust.

Methodologies commonly employed in Neuroeconomics

One fundamental building block of experiments in neuroeconomics is the behavioral paradigm that is employed to investigate decision-making. Behavioral paradigms are typically taken from experimental economics and psychology. Examples include the Trust Game (e.g. Berg et al., 1995; Kosfeld et al., 2005), in which participants are faced with the decision to entrust an anonymous person with their money, or the risky choice task (e.g. Engelmann and Tamir, 2009), in which participants choose between lotteries with different levels of risk and real financial consequences. Behavioral paradigms are then adapted for use with methods from cognitive neuroscience, such as non-invasive brain imaging and stimulation, which require specific experimental design considerations related to timing and randomization of events in order to optimize statistical analyses. Common neuroscience methods employed to study the brain in neuroeconomics include functional Magnetic Resonance Imaging (fMRI), which allows researchers to observe brain function during decision-processes, transcranial magnetic stimulation (TMS), through which a temporary lesion in a targeted brain region can be generated, and pharmacological manipulations, which can temporarily alter the amount of a targeted neurotransmitter within the central nervous system.

By far the most commonly used tool to investigate the neural basis of economic and social decision-making is fMRI. This procedure involves placing participants inside an MRI scanner, which provides a record of neuronal activity in form of the Blood Oxygen Level-Dependent (BOLD) response within the brain while participants make decisions in the context of an experimental paradigm. The BOLD response is reflective of the amount of oxygen that is being delivered to neurons through the blood stream. Changes in blood oxygenation levels lead to localized changes in the ferromagnetic properties of blood that can be visualized by the MRI scanner. Specifically, an increased BOLD signal in a certain brain region is indicative of an increase in the amount of oxygen within the blood stream, which in turn is reflective of an increased demand for energy made by active neurons (e.g. Raichle and Gusnard, 2002). Researchers can then couple the data provided by the fMRI scanner, the BOLD signal, with behavioral data obtained from participants' task performance that reflects economic and social preferences to localize regions involved in producing behavior using statistical analyses based on the General Linear Model. To accomplish this feat, recent experiments have used model-based approaches to investigate brain function (O'Doherty et al., 2007). In such sophisticated analyses, quantitative computational models estimate parameters reflective of participants' behavior that are relevant to the cognitive processes underlying economic decision-making. These variables can then be used to make predictions about what patterns of activation a given brain region should follow if it is involved in relevant computations that lead to the observed behavior. Specifically, brain regions whose neural signals show significant correlation with parameters derived from computational models are likely involved in the computations that produce the decisions of interest. Using such model-based fMRI analyses, O'Doherty et al. (2003) have demonstrated that neurons in the ventral striatum make specific predictions about the reward contingencies of the environment in the form of the reward prediction error[1] (O'Doherty et al., 2003), a finding that had previously been demonstrated using invasive electrophysiological recordings in monkeys (Schultz et al., 1998).

One drawback of fMRI is that it provides correlational information, meaning that brain regions identified to be involved in the decision-process by this approach only correlate with the observed behavior that was performed inside the scanner. While considerable evidence indicates that regions identified via this method are involved in the computations that underlie behavior (Logothetis et al., 2001), causal inferences in the form of *brain region X causes behavior Y* can only be made in the presence of converging evidence. Such converging evidence is commonly provided by studies from patients with localized brain lesions, that, after removal of brain tissue, show aberrant performance on decision-making tasks (e.g. Bechara et al., 1996). It is, however, often difficult to come by patients with focal damage to brain regions of interest and this approach suffers from its own drawbacks. Recent technological advances have made non-invasive stimulation techniques that can be applied to healthy human participants, such as Transcranial Magnetic Stimulation (TMS) and transcranial Direct Current Stimulation (tDCS) available to researchers in neuroeconomics. Such brain stimulation approaches can provide converging causal evidence that directly implicate brain regions in decision-processes.

Transcranial Magnetic Stimulation (TMS) involves applying high-intensity magnetic pulses via a magnetic coil that is strategically placed over the scalp to alter the excitability of neurons within specific regions of cortex. The transient magnetic pulse penetrates

[1] Human and animal studies have repeatedly demonstrated that midbrain dopamine neurons process rewarding stimuli, such as food or money. Expectation modulates activity within these regions: when a rewarding stimulus is unexpected, firing rates increase; when it is expected, midbrain dopamine neurons do not change their firing rates, and when an expected reward does not occur, a depression in firing rates is observed. Such neuronal activity patterns closely follow predictions from reinforcement learning theory about how an agent learns environmental contingencies to maximize rewards (for review see Montague et al., 2004).

through scalp and skull to generate an electrical field that can either stimulate or disrupt neurons within targeted areas of cortex. The specific effects of TMS on the excitability of cortical neurons depends on variations in stimulation parameters, including intensity, frequency and repetition amount of the stimulation (Fitzgerald et al., 2002). Using specific stimulation parameters, researchers can create a localized, temporary and fully reversible brain lesion and study the effects of disabling activity within a region of interest on behavior. Furthermore, temporally extended effects on cortical excitability can be achieved via repetitive TMS (rTMS), in which trains of pulses are delivered to the same brain region for several minutes. This approach has yielded important insights about the role of dorsolateral prefrontal cortex (DLPFC) in social decisions (Knoch et al., 2006). Using low-frequency rTMS, Knoch et al. (2006) disrupted activity within the right DLPFC, a region previously implicated in executive control and inhibition of prepotent responses. Participants then played the Ultimatum Game, in which they were asked to either accept or reject unfair offers made by another participant (see below for a more detailed outline of the Ultimatum Game). In the context of this game, accepting an unfair offer can be interpreted as following mostly selfish impulses that maximize payoffs, while rejecting such an offer is reflective of punishing norm violators, a decision that is costly to the subject. Knoch et al. showed that participants whose right DLPFC was disrupted accepted significantly more unfair offers compared to placebo stimulation, indicating that the neural computations performed in DLPFC are important for controlling self-interest and punishing unfair offers.

Pharmacological manipulations involve administering a drug challenge using compounds known to change the availability of the targeted neurotransmitter within the central nervous system and subsequently testing the effects of altered brain chemistry on behaviors of interest. Neurotransmitter systems that are commonly targeted are the dopamine system, for instance via systemic administeration of the dopamine agonist L-DOPA (e.g. Pleger et al., 2009; Eisenegger et al., 2010), and the oxytocin system (e.g. (Kosfeld et al., 2005; Baumgartner et al., 2008). These systems have repeatedly been implicated in core decision processes, with dopaminergic brain regions being involved in signaling the rewarding value of stimuli, while oxytocin is involved in mediating social behaviors. A recent investigation of the role of serotonin (5-HT) in social decision-making demonstrated that decreased serotonin levels within the central nervous system increased rejection rates of unfair offers in the Ultimatum Game (Crockett et al., 2008). Together with a known role of serotonin in impulsive aggression, these findings underline the importance of emotion regulation in ultimatum bargaining.

Finally, recent investigations in neuroeconomics have begun to examine the role of genetic predispositions in economic and social decision-making (e.g. Kuhnen and Chiao, 2009; Eisenegger et al., 2010), as well as aberrant decision-making in clinical populations, including patients with depression, schizophrenia and anxiety disorders (e.g. Roiser et al., 2010).

Social Neuroeconomics investigates the neural basis of social preferences

Traditional economics assumes that economic behavior is primarily based on material self-interest, such that rational decision-makers should maximize their own payoffs. Considerable evidence from a multitude of experiments, however, argues against this simplifying assumption (for review see Kahneman, 2003). Some of the most striking examples in support of this notion are found during strategic interactions (e.g. Fehr and Gachter, 2002). Experiments using games in which one player's actions have a direct impact on the payoffs of other participants, have repeatedly demonstrated that people exhibit clear «social preferences». This means that player's decisions are other-regarding, taking into account the well-being of other players even though, in the context of strategic interactions, such choices are costly and therefore not in agreement with their own self-interest.

To illustrate how social preferences can affect behavior, consider for example results from the following oft-cited experiment employing the Ultimatum Game (Güth et al., 1982). Two anonymous players interact in the Ultimatum Game, the «proposer» and the «responder». The proposer is allocated a certain sum of money, say 10 monetary units (MU), with the task to split the money as he wishes. If the responder accepts the split, both participants earn their respective amounts. If, however, the responder rejects the split, neither of the participants earns any money. Traditional economics would argue that the responder should accept any amount offered by the proposer that is greater than 0 MU, as this strategy would maximize his payoff. However, in the experiment by Güth and colleagues, responders tend to reject small offers of 2 MU, a 20:80 split in favor of the proposer, about half of the time. Such findings indicate that responders are willing to pay 2 MU, and sometimes even more, to punish the proposer for his unfair offer. These results have been replicated in a multitude of experiments using the Ultimatum Game and generalize to different experimental settings (for reviews see Camerer, 2003; Fehr, 2009). Such behavior is easily interpreted as an expression of social preferences that goes against the notion that behavior is motivated purely by self-interest. Because of their ubiquity, social preferences have been incorporated as social utilities in formal models of social preferences by assigning subjective values to other's well-being (for review see Fehr, 2009).

Measuring trust

Similar results have been obtained using a related experimental setup, referred to as the Trust Game. In a typical version of the Trust Game, originally introduced by Berg et al. (1995), two anonymous players called the «investor» and «trustee» interact by sequentially exchanging monetary amounts as follows: The investor is allocated a certain amount of money by the experimenter, say 10 MU, and is asked to send any amount from her endowment to the trustee. Known to both participants, the amount transferred by the investor is then tripled by the experimenter. The trustee's role is to decide how to share her endowment with the investor, that is, how much money to send back. The amount sent by the investor is taken to reflect trust-taking, as she voluntarily makes herself vulnerable by placing her resources at the disposal of the trustee. Her motive for doing so is that the social risk she took could increase her financial wellbeing, if her trust is reciprocated. The amount returned by the trustee, then, is a measure of prosociality and trustworthiness.

In the most commonly used version of the Trust Game, referred to as one-shot games, players interact with a different individual on each trial in order to avoid reputation building. In such settings, according to predictions made by the self-interest assumption of traditional economics, investors should keep all their money to themselves, because they cannot expect any return from a purely self-interested individual. Trustees, on the other hand, should not transfer back any money, because in a one-shot game no financial gain is obtained from repaying the investor. The results reported by Berg et al. (1995), however, paint a different picture. They find that investors in their experiment sent more than half their initial endowment on average and that about 95% of the investment was repaid by trustees. These results have been replicated across a multitude of cultures (for review see Camerer, 2003), indicating that people generally trust and are trustworthy. Together, these results underline the notion that social preferences interact with self-interest in the production of behavior.

What does the Trust Game measure? The following criticism in regard to the external and ecological validity of the one-shot Trust Game is commonly raised: Behavior in the Trust Game does not reflect trust as understood by the general public or the social sciences. Trust is a multifaceted concept that is significantly richer than the behavior measured in the Trust Game. For instance, it is a fundamental building block of most interpersonal relationships such as friendships, marriages, and parent-child relations, but is also crucial for economic transfers. Trust is also established and maintained by social, legal and economic sanctions and can be furthered by communication. All these elements seem to be absent in the behaviors measured by the Trust Game.

In one way, these criticisms are correct. In order to investigate trust experimentally, researchers have to reduce this complex concept to observables, that is, facets of behavior that can be measured quantitatively. This, however, is a desirable feature of the Trust Game, as it provides a clean measure of trust that is free of confounding variables, such as reputation-building, contractual pre-commitments and the potential of punishment. Compared to face-to-face interactions, which would be more naturalistic, anonymity controls for trustworthiness inferences from facial features that are typically made within a fraction of a second (Todorov et al., 2009). Research has demonstrated that such inferences are influenced to a great degree by facial attractiveness (Stirrat and Perrett, 2010) and have little to no predictive power about the actual trustworthiness of a person (*personal communication with Charles Efferson*). Furthermore, repeated interactions between two players would allow individuals to form a reputation, which is driven by the individual's self-interest to increase future payoffs. Therefore, experiments that ensure anonymity between players and employ one-shot games provide the cleanest measure of trust (Fehr, 2009).

Despite the above limitations of the experimental approach, recent experiments in social neuroeconomics provide evidence in support of the notion that the Trust Game captures relevant aspects of a broader definition of trust. Indeed, considerable evidence underlines the social nature of the decision-process involved in trust-taking in the context of the Trust Game, which can be summarized as follows: (1) behavioral and brain responses during trust-taking are distinct from risk-taking and (2) trusting decisions in the context of Trust

Games recruit neural circuitry commonly implicated in generating models of other's mental states, referred to as Theory of Mind (ToM). Taken together, converging evidence from social neuroeconomics indicates that participants interpret the Trust Game as a strategic environment, underlining the ecological validity of the Trust Game.

The Neurobiology of Trust

1. Trust-taking is distinct from risk-taking. In the context of the Trust Game, a decision to trust could inherently be a decision to take a risk, be it a social risk. That is to say that an investor who decides to transfer some or all of her endowment to the trustee forfeits a sure payoff and risks losing some or all of it at the benefit of potentially increasing her final payoff, depending on the decision of the trustee. One question that arises from this notion is whether the brain makes a distinction between trust-taking within the Trust Game and risk-taking in a similarly framed choice setting. If trust-taking and risk-taking produce differential behaviors and are processed by separate brain systems, this would underline the social nature of trust-taking in the context of the Trust Game. Two recent experiments using the neuropeptide oxytocin (OT) have shed light on this question.

Oxytocin is a uniquely mammalian neuropeptide that is synthesized in the hypothalamus. It is released into the bloodstream via the pituitary gland, where it acts as a hormone that is important for partuition and lactation (Burbach et al., 2006). Of high relevance to social neuroscience is the fact that the hypothalamus can also release OT within the central nervous system, where it acts as a neurotransmitter. OT binding sites are found throughout the brain, but regions showing the highest OT receptor density are located within the limbic brain and include the ventral striatum, nucleus accumbens and amygdala (e.g. Hammock and Young, 2006). There is now considerable evidence implicating OT in social behaviors and cognition, including maternal behavior and pair bonding, social recognition, as well as social motivation and sexual behavior (for review see Skuse and Gallagher, 2009). For instance, in rodents it has been demonstrated that administration of an OT antagonist, which blocks the action of OT in the central nervous system, decreases maternal behavior (van Leengoed et al., 1987) while direct infusion of OT into the brain leads to increased maternal behavior in animals known to be non-maternal (Pedersen et al., 1982). Similarly, central infusion of OT promotes pair bonding in monogamous voles (Williams et al., 1992), while a blockage of OT action within the brain decreases pair bonding (Cho et al., 1999). In humans, it has repeatedly been demonstrated that intranasal administration of OT, which increases OT concentrations within the brain (Born et al., 2002), facilitates social cognition. A number of studies have used this approach, demonstrating that OT improves performance of difficult discriminations of facial expressions in healthy humans (Domes et al., 2007a) as well as in patients with autism spectrum disorder (ASD) (Guastella et al., 2010), a neurodevelopmental disorder chararcterized by grave social deficits (for review see Frith and Frith, 2003). Neuroimaging studies have demonstrated that the social effects of OT are likely mediated via the amygdala (Kirsch et al., 2005; Domes et al., 2007b), a region that is important for social cognition and affective processing.

Kosfeld et al. (2005) conducted a version of the Trust Game with two groups, one group that received intranasal administration of synthetic oxytocin and a control group that received inactive placebo administration. Their results indicate that oxytocin administration led to increases in trust-taking, as investors in the OT group had significantly greater average transfers relative to the placebo control group. To investigate whether oxytocin effects are specific to trust-taking in social settings or whether there is a more generalized effect on risk-taking, the study included a risk task, in which participants faced the same choices, except that they knew that the amount of money returned was determined by a random mechanism implemented by a computer instead of a human counterpart. Interestingly, despite the fact that the choice context was the same, oxytocin had no effect on average transfer rates in this condition. Finally, the effect of oxytocin could be to increase prosocial inclinations in general, in which case, increased transfer rates should not only be observed for investors, but also for trustees. No effect of oxytocin for trustees was observed, indicating that the effect of oxytocin was specific for trust-taking. Taken together, results from Kosfeld et al. (2005) indicate that oxytocin increases trust-taking, and that these effects are specific and do not generalize to risk-taking, nor lead to increased prosocial behaviors. Because of oxytocin's known role in facilitating social behaviors, the authors interpreted these results as indicating that OT's effects on trust-taking are mediated via reducing social anxiety, such as betrayal aversion (Bohnet and Zeckhauser, 2004).

A more recent neuroimaging investigation by Baumgartner et al. (2008) provides supporting evidence for this conclusion. Similar to the experiment by Kosfeld et al. (2005), this experiment included two groups of participants, one that received oxytocin and a placebo control group. While undergoing fMRI, participants played both the Trust Game and a risk game in which choices are equivalent to the Trust Game, except that reinforcement was based on a random algorithm implemented by a computer instead of choices made by another player. Importantly, about halfway through the experiment participants were informed that their decisions to trust and take risks were not returned on half the trials. Behavioral results indicate that, while group differences before feedback were not statistically significant, behavioral adaptation to feedback was significantly affected by oxytocin. Specifically, average transfers increased after feedback in the OT group and decreased significantly in the placebo group. No such differences were observed in the risk game, indicating that these effects are specific to social risks taken in the Trust Game. These results are consistent with the notion put forward by Kosfeld et al. (2005) that OT reduces fear of social betrayal. Neuroimaging findings from Baumgartner et al. (2008) lend further support to this notion. They demonstrate an increase in activity in the amygdala during trust-taking in the postfeedback phase relative to prefeedback in the placebo group, but no such effect in the OT group. The effect of OT was therefore to decrease activity in the amygdala, which was associated with increased trust-taking even after participants were informed of betrayal. Taken together with an extensive literature implicating the amygdala in fear processing and emotional relevance detection (e.g. Phelps and LeDoux, 2005), as well as studies showing that OT decreases fear responses by modulating activity in the amygdala (Kirsch et al., 2005; Domes et al., 2007b), these findings are consistent with the notion that OT reduces social fear by decreasing reactivity of the amygdala. Since such effects are specific to social tasks, OT administration likely leads to a reduction of social anxiety, such as betrayal aversion (Bohnet and Zeckhauser, 2004).

2. Trust-taking recruits neural circuitry involved in perspective-taking. The differential effects of OT on behavior in risk and Trust Games indicate that there is a specifically social aspect to trust-taking that is mediated via the social neuropeptide OT. A further factor that plays an important role in game theoretical paradigms, such as the Trust Game, is the ability to reason about other people's mental states, also termed *Theory of Mind (ToM)*. When interacting with other human players in the context of economic games, knowledge of another player's state of mind, that is his level of fairness, current emotional state and how he might react to my behavior, can help predict the opponent's future actions and provide a competitive advantage. Even in one-shot games, in which such knowledge cannot be obtained, strategic thoughts concerning the perspective of an opponent have been shown to occur (Costa-Gomez et al., 2001). Such strategies are absent when outcomes of decisions are determined by random processes, such as in the context of the risk game, or when playing against a computerized opponent.

Social perspective taking has been extensively studied by previous research. Developmental studies using the false-belief task in children have demonstrated that the ability to infer another's perspective does not fully develop until the age of 5 (Baron-Cohen et al., 1985) and is impaired in patients with autism spectrum disorder (ASD) (e.g. (Baron-Cohen et al., 1994). Neuroimaging studies employing such tasks have implicated a specific network of brain regions in the capacity to infer other's mental states, including the medial prefrontal cortex, superior temporal sulcus and temporoparietal junction (for reviews see Amodio and Frith, 2006; Hein and Knight, 2008).

Supporting evidence for the notion that the Trust Game does indeed measure social aspects of trust would be provided by results from behavioral and neuroimaging experiments underlining the notion that investors engage in perspective-taking. In one of the first neuroimgaing investigations of trust-taking by McCabe et al. (2001), participants played a number of games that included the Trust Game outlined above. On half of the trials, participants interacted with another human player, while on the other half, the opponent participants faced was a computer. Results from this study indicate that participants could be distinguished based on two distinct strategies they employed to solving this task: (1) cooperators that considered the well-being of other players and (2) non-cooperators. Interestingly, these different strategies led to differential brain activation patterns. Only cooperators showed significant activation in mPFC, specifically the anterior paracingulate cortex, when interacting with a human compared to a computer counterpart. Given the known involvement of mPFC in mentalizing (Amodio and Frith, 2006), the authors concluded that generating a model of the other player's mental state plays an important role in producing cooperative behavior.

A number of related studies lend support to this notion. A Positron Emission Tomography (PET)[2] study by Gallagher et al. (2002) showed activation in anterior paracingulate cortex when participants played a game of «stone-paper-scissors» against a human compared to a computer opponent. Similarly, Rilling et al. (2004)

[2] Positron Emission Tomography is a neuroimaging method that uses radioactive tracers injected into the blood stream. Using specific tracers such as fludeoxyglucose, this method can provide images reflective of the energy consumption in the brain.

demonstrated activations in anterior paracingulate cortex, as well as superior temporal sulcus and posterior cingulate cortex when participants played the Ultimatum Game or the related Prisoners Dilemma Game against a human compared to a computer opponent. Together, these results demonstrate that fundamentally different choice strategies are employed when participants interact with human compared to non-human partners in various game-theoretical settings. Neurobiological evidence showing that central nodes within the Theory of Mind network are intimately involved in decision-making when participants interact with other human players lends support to the notion that social choice involves the generation of models representing another agent's mental state.

In repeated Trust Games, the same participants interact with each other over several rounds of the Trust Game by sequentially placing both participants in the roles of the trustee and the investor. This approach allows participants to form a reputation and thus increases the ecological validity of the experiment. Forming a reputation is similar to building a mental model of the trustworthiness of the other player that likely involves mentalizing. Various experiments have employed this approach in combination with hyperfunctional fMRI, which provides a simultaneous record of neuronal activity from multiple brains. A recent study by King-Casas et al. (2005) showed that activity within the trustee's caudate nucleus encodes her intention to trust on the next trial. Specifically, during trials immediately before trustees decided to increase trust, signal in the caudate nucleus increased after the investor's decision was revealed. Importantly, as participants formed a model of the other player, this response shifted forward in time, such that in later rounds of the Trust Game, an increased peak predicting the intention to trust on the next trial was observed even before the investor's decision was revealed. A similar temporal shift was observed in cross- and within-brain correlations between activity in the caudate nucleus, as well as mid and anterior cingulate cortex, a region that is adjacent to the anterior paracingulate cortex. Together, these results indicate that caudate responses transitioned from reactive to anticipatory as a mental model of the other player was formed in a fashion that resembles the reward prediction error. Corroborating evidence for this notion was provided by a separate behavioral study, in which the authors showed that trustees predictions about the magnitude of investor's backtransfers significantly improved after playing about eight rounds of the Trust Game with the same partner. These results indicate that trustees build a mental model of their partner's trustworthiness, whose predictions become more accurate with increasing experience. Neurobiological evidence implicates a brain network involving caudate nucleus and anterior cingulate cortex in mediating this process.

A follow-up study using the same experimental setup demonstrated that specialized regions within cingulate cortex encode decisions of others and oneself (Tomlin et al., 2006), with regions in mid-cingulate most responsive to the subject's own decision, while anterior and posterior regions represented the other subject's choice. Importantly, in non-social control experiments that kept the visual and motor requirement constant, no such activations were observed within cingulate cortex. Together, these findings demonstrate that the cingulate cortex performs social agency computations to perform distinctions between actions made by the self and others, implicating this region in generating mental models of other players.

Finally, Krueger et al. (2007), using a multiround Trust Game, demonstrated activation of the paracingulate cortex during trials in which participants decided to trust. An interesting feature of this study was the observation that two types of relationships developed in the course of repeated interactions between two players: (a) in nondefector relationships trustees never defected on investors' decisions to trust while (b) in defector relationships players experienced several breaches of trust throughout the experiment. Interestingly, activity in the paracingulate region differentiated between such relationships, showing significantly greater activation during trust-building in non-defectors compared to defectors. These results indicate that an engagement of the mentalizing network during the relationship-building phase of the experiment led to greater levels of cooperation.

Taken together, results reviewed above lend support to the notion that participants interacting with human opponents recruit central nodes of the Theory of Mind network, such as STS and mPFC, while such activations are absent when interacting with non-intentional entities, such as a computer. Evidence from multiround Trust Games indicates that such activation is involved in generating models of other player's mental states. To return to our original question, these results therefore lend further support to the notion that trust-taking in the context of the Trust Game is perceived as a social act that involves generating models of other players' mental states and their social preferences, even in the context of one-shot games.

Conclusions

Social interactions are inherently complex and can be influenced by a multitude of factors. In order to be able to experimentally investigate such complex interactions, the experimental approach employed in social neuroeconomics requires a reduction of complicated concepts, such as trust, to observable facets of behavior that can be quantitatively measured. On the one hand, this leads to experimental tasks that measure very specific behaviors that, at first glance, do not appear naturalistic nor easily generalize to real-world interpersonal interactions. On the other hand, the experimental approach allows for observation of quantifiable behaviors that can be coupled with methods from neuroscience to enable investigations of brain function. Importantly, via the example of the Trust Game we have demonstrated that, despite the reductionist experimental approach employed in neuroeconomics, the main features of the behavior of interest are preserved. Specifically, participants interpret the Trust Game as a social environment that requires strategic interactions with another player that are distinct from simple risk-taking and involve generating models of other player's social preferences. Removal of features that we cannot directly measure but might systematically influence our results is necessary to obtain clearly interpretable data free from confounding factors. While this leads to limited insights achieved by any given experiment, there is always the possibility to perform follow-up experiments that investigate additional features of the behavior of interest.

I would to thank Grit Hein, Christian Ruff and Friederike Meyer for comments on earlier versions of this manuscript.

References

Amodio DM., Frith CD. (2006): Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7: 268-277.

Baron-Cohen S., Leslie AM., Frith U. (1985): Does the autistic child have a «theory of mind»? Cognition 21: 37-46.

Baron-Cohen S., Ring H., Moriarty J., Schmitz B., Costa D., Ell P. (1994): Recognition of mental state terms. Clinical findings in children with autism and a functional neuroimaging study of normal adults. Br J Psychiatry 165: 640-649.

Baumgartner T., Heinrichs M., Vonlanthen A., Fischbacher U., Fehr E. (2008): Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron 58: 639-650.

Bechara A., Tranel D., Damasio H., Damasio AR. (1996): Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. Cereb Cortex 6: 215-225.

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

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

Born J., Lange T., Kern W., McGregor GP., Bickel U., Fehm HL. (2002): Sniffing neuropeptides: a transnasal approach to the human brain. Nat Neurosci 5: 514-516.

Burbach JP., Young WJ., Russel JA. (2006): Oxytocin: Synthesis, secretion and reproductive functions. In: Physiology of reproduction (Neil JD., ed), pp 3055-3128. Amsterdam: Elsevier.

Camerer CF. (2003): Behavioral Game Theory: Experiments in Strategic Interaction. Princeton, NJ: Princeton University Press

Cho MM., DeVries AC., Williams JR., Carter CS. (1999): The effects of oxytocin and vasopressin on partner preferences in male and female prairie voles (Microtus ochrogaster). Behav Neurosci 113: 1071-1079.

Costa-Gomez M., Crawford VP., Broseta B. (2001): Cognition and behavior in normal-form games: an experimental study. Econometrica 69: 1193-1235.

Crockett MJ., Clark L., Tabibnia G., Lieberman MD., Robbins TW. (2008): Serotonin modulates behavioral reactions to unfairness. Science 320: 1739.

Domes G., Heinrichs M., Michel A., Berger C., Herpertz SC. (2007a): Oxytocin improves «mind-reading» in humans. Biol Psychiatry 61: 731-733.

Domes G., Heinrichs M., Glascher J., Buchel C., Braus DF., Herpertz SC. (2007b): Oxytocin attenuates amygdala responses to emotional faces regardless of valence. Biol Psychiatry 62: 1187-1190.

Eisenegger C., Knoch D., Ebstein RP., Gianotti LR., Sandor PS., Fehr E. (2010): Dopamine receptor D4 polymorphism predicts the effect of L-DOPA on gambling behavior. Biol Psychiatry 67: 702-706.

Engelmann JB., Tamir D. (2009): Individual differences in risk preference predict neural responses during financial decision-making. Brain Res 1290: 28-51.

Fehr E. (2009): Social Preferences and the Brain. In: Neuroeconomics: Decision Making and the Brain (P.W. Glimcher CC, R.A. Poldrack, E. Fehr, ed), pp 215-231. London, UK: Academic Press.

Fehr E., Gachter S. (2002): Altruistic punishment in humans. Nature 415: 137-140.

Fitzgerald PB., Brown TL., Daskalakis ZJ. (2002): The application of transcranial magnetic stimulation in psychiatry and neurosciences research. Acta Psychiatr Scand 105: 324-340.

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

Gallagher HL., Jack AI., Roepstorff A., Frith CD. (2002): Imaging the intentional stance in a competitive game. Neuroimage 16: 814-821.

Guastella AJ., Einfeld SL., Gray KM., Rinehart NJ., Tonge BJ., Lambert TJ., Hickie IB. (2010): Intranasal oxytocin improves emotion recognition for youth with autism spectrum disorders. Biol Psychiatry 67: 692-694.

Güth W., Schmittberger R., Schwarze B. (1982): An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3: 367-388.

Hammock EA., Young LJ. (2006): Oxytocin, vasopressin and pair bonding: implications for autism. Philos Trans R Soc Lond B Biol Sci 361: 2187-2198.

Hein G., Knight RT. (2008): Superior temporal sulcus--It's my area: or is it? J Cogn Neurosci 20: 2125-2136.

Kahneman D. (2003): Maps of bounded rationality: Psychology for behavioral economics. The American Economic Review 93: 1449-1475.

King-Casas B., Tomlin D., Anen C., Camerer CF,. Quartz SR., Montague PR. (2005): Getting to know you: reputation and trust in a two-person economic exchange. Science 308: 78-83.

Kirsch P., Esslinger C., Chen Q., Mier D., Lis S., Siddhanti S., Gruppe H., Mattay VS., Gallhofer B., Meyer-Lindenberg A. (2005): Oxytocin modulates neural circuitry for social cognition and fear in humans. J Neurosci 25: 11489-11493.

Knoch D., Pascual-Leone A., Meyer K., Treyer V., Fehr E. (2006): Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314: 829-832.

Knutson B., Taylor J., Kaufman M., Peterson R., Glover G. (2005): Distributed neural representation of expected value. J Neurosci 25: 4806-4812.

Kosfeld M., Heinrichs M., Zak PJ., Fischbacher U., Fehr E. (2005): Oxytocin increases trust in humans. Nature 435: 673-676.

Krueger F., McCabe K., Moll J., Kriegeskorte N., Zahn R., Strenziok M., Heinecke A., Grafman J. (2007): Neural correlates of trust. Proc Natl Acad Sci U S A 104: 20084-20089.

Kuhnen CM., Chiao JY. (2009): Genetic determinants of financial risk taking. PLoS One 4: e4362.

Logothetis NK., Pauls J., Augath M., Trinath T., Oeltermann A. (2001): Neurophysiological investigation of the basis of the fMRI signal. Nature 412: 150-157.

McCabe K., Houser D., Ryan L., Smith V., Trouard T. (2001): A functional imaging study of cooperation in two-person reciprocal exchange. Proc Natl Acad Sci U S A 98: 11832-11835.

McClure SM., Laibson DI., Loewenstein G., Cohen JD. (2004): Separate neural systems value immediate and delayed monetary rewards. Science 306: 503-507.

Montague PR., Hyman SE., Cohen JD. (2004): Computational roles for dopamine in behavioural control. Nature 431: 760-767.

O'Doherty JP., Hampton A., Kim H. (2007): Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104: 35-53.

O'Doherty JP., Dayan P., Friston K., Critchley H., Dolan RJ. (2003): Temporal difference models and reward-related learning in the human brain. Neuron 38: 329-337.

Pedersen CA., Ascher JA., Monroe YL., Prange AJ., Jr. (1982): Oxytocin induces maternal behavior in virgin female rats. Science 216: 648-650.

Phelps EA., LeDoux JE. (2005): Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48: 175-187.

Pleger B,. Ruff CC., Blankenburg F., Kloppel S., Driver J., Dolan RJ. (2009): Influence of dopaminergically mediated reward on somatosensory decision-making. PLoS Biol 7: e1000164.

Raichle ME., Gusnard DA. (2002): Appraising the brain's energy budget. Proc Natl Acad Sci U S A 99: 10237-10239.

Rilling JK., Sanfey AG., Aronson JA., Nystrom LE., Cohen JD. (2004): The neural correlates of theory of mind within interpersonal interactions. Neuroimage 22: 1694-1703.

Roiser JP., Stephan KE., den Ouden HE., Friston KJ., Joyce EM. (2010): Adaptive and aberrant reward prediction signals in the human brain. Neuroimage 50: 657-664.

Sanfey AG., Rilling JK., Aronson JA., Nystrom LE., Cohen JD. (2003): The neural basis of economic decision-making in the Ultimatum Game. Science 300: 1755-1758.

Schultz W., Tremblay L., Hollerman JR. (1998): Reward prediction in primate basal ganglia and frontal cortex. Neuropharmacology 37: 421-429.

Skuse DH., Gallagher L. (2009): Dopaminergic-neuropeptide interactions in the social brain. Trends Cogn Sci 13: 27-35.

Stirrat M., Perrett DI. (2010): Valid facial cues to cooperation and trust: male facial width and trustworthiness. Psychol Sci 21: 349-354.

Todorov A., Pakrashi M., Oosterhof NN. (2009): Evaluating faces on trustworthiness after minimal time exposure. Social Cognition 27: 813-833.

Tomlin D., Kayali MA., King-Casas B., Anen C., Camerer CF., Quartz SR., Montague PR. (2006): Agent-specific responses in the cingulate cortex during economic exchanges. Science 312: 1047-1050.

van Leengoed E., Kerker E., Swanson HH. (1987): Inhibition of post-partum maternal behaviour in the rat by injecting an oxytocin antagonist into the cerebral ventricles. J Endocrinol 112: 275-282.

Williams JR., Carter CS., Insel T. (1992): Partner preference development in female prairie voles is facilitated by mating or the central infusion of oxytocin. Ann NY Acad Sci 652: 487-489.

— Dr. Jan Engelmann arbeitet im Rahmen des Projekts «Vertrauen verstehen» an einer Habilitation mit dem Thema «Die emotionalen und neurobiologischen Ursachen von Vertrauen».