EMBO reports

science & society

Trust in the brain

Neurobiological determinants of human social behaviour

Michael Kosfeld

What determines how humans interact socially? Why do we sometimes cooperate but at other times refuse to act cooperatively? Why are some people willing to trust a stranger, whereas others are mistrustful? Why do some people take risks to achieve their goals, whereas others prefer to stay on the safe side? Answering such questions will eventually help both social scientists and biologists to unravel the mechanisms that guide the social interactions of *Homo sapiens*, with potentially wide-ranging implications for human life.

In the past few decades, social scientists have made much progress in understanding the mechanisms of human social interaction and cooperation. Their main analytical tool, game theory, has been useful in widely varied disciplines such as economics, psychology, sociology and political science. Owing to the nature of these fields, most of the questions that social scientists ask focus on the external social, environmental and institutional determinants of human behaviour; the internal biological mechanisms underlying and regulating individual decision-making, however, have been treated more or less as a 'black box'.

At the same time, the neurosciences have long avoided explicitly modelling human social interactions. Interestingly—and I believe fortunately—this has changed recently. With the emergence of new fields such as neuroeconomics and social cognitive neuroscience, scientists have begun to look into the black box that guides human decision-making in social contexts. Their main approach is to combine methods from both the social sciences and neuroscience to understand how the human brain generates decisions. Game theory is now used to analyse the human brain in conjunction with a variety of scientific techniques, including functional magnetic resonance imaging, transcranial magnetic stimulation and intranasal administration of neuropeptides. In this article, I provide some insights into this interdisciplinary research by describing a particular game—the trust game—that has been used successfully to study the neurobiology of human social behaviour.

Game theory dates back to the late 1920s and has its roots in an article by the Hungarian mathematician John von Neumann, who subsequently published his book on the *Theory of Games and Economic Behavior* with Oskar Morgenstern in 1944. By providing an elegant mathematical language, game theory quickly became a powerful tool for analysing social situations in which two or more individuals interact.

Game theory regards—and therefore models—a social situation as a strategic game. Such a game consists formally of three elements: the players who interact with each other; a set of available actions for each player; and a so-called pay-off function for each player, which models the player's individual preferences such as his or her valuations of all possible outcomes of the game. The key feature of a strategic game is the fact that, although each player makes individual decisions according to his or her own interests, the behaviour of all players determines the final outcome. Therefore, the possible outcomes that an individual player can achieve depend not only on the player's own behaviour but also on the other players with whom he or she interacts. The players' behaviour reaches

With the emergence of new fields such as neuroeconomics and social cognitive neuroscience, scientists have begun to look into the black box that guides human decision-making in social contexts

equilibrium if no player has an incentive to deviate from his or her supposed behaviour. In 1950, the American mathematician John F. Nash proved that such an equilibrium exists for every game—although players might sometimes have to randomize over their actions. His proof made possible thousands of applications of game theory within the social sciences, and he later shared the 1994 Nobel Prize in economics for his contribution.

An important example of a strategic game is the so-called trust game for two players. The first player is the investor; the second player is the trustee. At the outset of the game, both players have an endowment of, in this example, 12 points—where each point is equivalent to real money. The investor decides first. He can transfer any amount, x, between 0 and 12 points, to the trustee. On the way to the trustee, the investor's transfer is tripled, so the trustee receives $3x$ points. Therefore, if the investor decides to transfer 8 points, the trustee receives 24 points. After the investor has made a decision, the trustee is informed about the investor's transfer. The trustee can then return any feasible amount back to the investor—that is, any amount, y, between 0 and $12 + 3x$ points. After the trustee has made a decision, the game is over and each player earns the number of points held in his or her hands. Therefore, the investor earns $12 - x + y$ points, and the

trustee earns $12 + 3x - y$ points. By returning twice the investor's transfer ($y = 2x$), the trustee could ensure that both players earn the same amount: $12 + x$. For example, if the investor transfers 12 points to the trustee and the latter returns 24 points, both players have doubled their initial endowment and earn 24 points. However, the trustee also has the option of returning nothing. If the investor transfers 12 points and the trustee returns 0 points, the investor earns nothing whereas the trustee earns 48 points.

The investor's decision in this game can be interpreted as a decision of trust—hence the name of the game. The investor must trust that the trustee will return a sufficiently large amount—at least as large as the original transfer—for the game to be profitable. At the same time, the trustee's behaviour can be seen as a measure of trustworthiness: the larger the trustee's back-transfer, the higher his or her trustworthiness towards the investor. If each player acts to maximize his or her own individual monetary income, however, the trustee has no reason to return any positive amount. The interaction ends after the trustee's decision, which means that the investor has no opportunity to sanction the trustee for misbehaviour. The game's equilibrium prediction is therefore that the trustee returns 0 points to the investor, regardless of the transfer received, and that the investor does not transfer any positive amount but retains all of his or her endowment. Although this is an equilibrium, the resulting outcome is clearly inefficient as no gains are realized: both players earn 12 points, but they could have earned twice as much if only the investor had trusted and the trustee had behaved in a trustworthy manner.

The game nicely captures the fundamental dilemma of trust in human society. According to sociologists Niklas Luhmann and John Coleman, trust is a risky decision whereby the trusting person risks being exploited, yet hopes that his or her trust will be rewarded by trustworthy behaviour (Luhmann, 1968; Coleman, 1990). In addition, trust increases the efficiency of the social interaction. If people trust and trust is rewarded, everyone is typically better off compared with a situation in which no one trusts and people act in an untrustworthy manner. Economist Kenneth Arrow therefore described trust as "an important lubricant of a social system" (Arrow, 1974). However, trust can, and sometimes will, be exploited. The risk involved in the trusting decision might therefore never be resolved completely. To achieve any potential benefits from the social interaction, the trusting person must always overcome his or her aversion to the risk of being exploited. The key question is how.

…there is a simple hypothesis about what might have a key role in the human brain in steering the decision to trust another human: oxytocin

The equilibrium prediction in the trust game—under the standard assumption that players maximize their individual incomes and no sanctions are possible—conjectures that there will be no trust, as any trust will be exploited. However, casual observations indicate that many people trust even in one-off situations and that trustworthy behaviour often reciprocates this trust. These observations are also corroborated by experimental evidence showing that the equilibrium prediction under the standard assumption fails to a large extent. Berg *et al* (1995) were the first to show in a laboratory experiment with real monetary stakes that subjects in the role of the investor transfer on average half of their initial endowment and receive average back-transfers of about the same size.

Different experiments have further identified conditions that affect the level of trust, such as repeated interaction, the possibility of sanctions or the activation of social norms. Although these studies have improved our understanding of the social—that is, external—determinants of trust, an important question still remains: what is it in the brain that makes humans trust each other? Although the question of analysing the trust game at the biological level might sound complex, there is a simple hypothesis about what might have a key role in the human brain in steering the decision to trust another human: oxytocin.

Oxytocin is a neuropeptide that is produced in the hypothalamus and stored in the posterior pituitary. It is released into the bloodstream, but is also widely distributed throughout the central nervous system. Apart from its classic physiological functions of stimulating smooth muscle contractions in the mammary myoepithelium during milk ejection and in the uterus during labour, oxytocin is known to promote social behaviour, including bonding, maternal behaviour and sexual behaviour. A number of animal studies have demonstrated that an intact intracerebral oxytocin system mediates the ability to form normal social attachments (Carter, 1998; Insel & Young, 2001). Furthermore, studies with human subjects showed that oxytocin reduces anxiety and the neuroendocrine responsiveness to social stressors, indicating that it acts as a central regulator of stress-protective social behaviour (Heinrichs *et al*, 2003).

As trust is an essential form of social behaviour that shares many characteristics with approach and bonding behaviour—a risky 'first step' towards another individual—the existing evidence on oxytocin function has sparked the hypothesis that the neuropeptide might increase the willingness of humans to show trust towards others. But how can we test this? Although the trust game provides a method for measuring individual levels of trust in humans, more is needed in order to examine the impact of brain oxytocin on human trust. Fortunately, neuropeptides such as oxytocin have been shown to pass through the blood–brain barrier after intranasal administration (Born *et al*, 2002), which provided a suitable method for testing the hypothesis in a double-blind placebo-controlled experiment (Kosfeld *et al*, 2005).

…oxytocin increased the subjects' willingness to trust their trustee, but did not make them more optimistic about the latter's trustworthiness

At the beginning of the experiment, the subjects received either a single intranasal dose of oxytocin or a placebo. Fifty minutes after substance administration, the subjects played the trust game against four different partners with the single modification that the investors could only make transfers of 0, 4, 8 or 12 points. The partners were matched randomly and no two subjects interacted twice. Each subject played the role of either the investor or the trustee in all four interactions. All interactions were anonymous—the subjects did not know the identity of the people with whom they were matched. In addition, the investor received no feedback about the trustee's decision in the four interactions. The subjects were fully informed about the structure of the experiment, including the rules of the game, the matching of different partners and the payment procedure at the end of the experiment. Each point in the experiment was worth 0.40 Swiss francs, so the subjects handled an endowment of 4.80 Swiss francs in each game.

The results of the experiment were intriguing. Of the 29 oxytocin-group investors, 45% transferred the maximum amount of 12 points in each interaction. By contrast, only 21% of the placebo-group investors did so. The average transfer made by the oxytocin-group investors was 9.6 points compared with 8.1 points by the placebo-group investors. Interestingly, the investors' expectations about the back-transfer from the trustee did not differ between the oxytocin and placebo recipients. Therefore, oxytocin increased the subjects' willingness to trust their trustee, but did not make them more optimistic about the latter's trustworthiness.

The key strategy is to combine game theory with state-of-the-art techniques that have been developed to analyse the decision-making process in the human brain

The results indicate that oxytocin somehow helps humans to overcome their natural aversion to uncertainty with regard to the behaviour of others. An important question, of course, is whether this effect is specific to the uncertainty in social interactions or also occurs in non-social situations. If the latter is true, these findings might hold not only for trust but also for more general situations of risk and uncertainty, such as the risk of stock-market trading.

To see whether the effect is specific, we conducted a second so-called risk experiment, in which investors faced the same choices as in the trust game but a random mechanism determined the outcome. The investors in the risk experiment were again in a risky situation, but this time the source of the risk was not another person's uncertain behaviour. Everything else in the risk experiment was identical to the trust experiment. No difference was found in the investors' behaviour between the oxytocin and the placebo groups. Therefore, we can conclude that the effect of oxytocin is, indeed, specific to trust and the willingness to take risks in social situations. As the risk experiment showed, oxytocin does not affect human attitudes towards risk and uncertainty in non-social contexts.

The discovery that oxytocin increases trust in humans is likely to have important clinical implications for patients with mental disorders associated with social dysfunctions, such as social phobia or autism. Social phobia ranks as the third most common mental health disorder after depression and alcoholism; sufferers are severely impaired when interacting socially with others and are often unable to show even basic forms of trust towards others. Given the results of the trust study, the administration of oxytocin in combination with behavioural therapy might yield positive effects for the treatment of these patients—particularly in light of the additional stress-protective and anxiolytic effects of the neuropeptide. The fact that oxytocin can easily be administered intranasally clearly facilitates future clinical applications.

At the same time, the results from these experiments might also raise fears of abuse, ranging from unscrupulous employers or insurance companies using oxytocin to induce trusting behaviour in their employees or clients, to dishonest car salesmen spraying customers with the hormone before praising the qualities of their fraudulent products. Fortunately, most of these fears are baseless, because the surreptitious administration of a substantial dose of oxytocin—for example, through air conditioning, food or drinks—is technically impossible, and forced nasal administration is likely to raise the recipient's level of distrust, therefore overriding any positive effect.

Eventually, the discovery that oxytocin promotes trust might even be useful for protecting consumers from the manipulation strategies of marketing departments. As these organizations use various techniques for inducing trust, a fascinating question is which of their approaches might trigger the endogenous release of brain oxytocin by cleverly designed stimuli.

New research in neuroeconomics and social cognitive neuroscience shows that scientists are beginning to uncover the neurobiological mechanisms that underlie human decision-making in social contexts. The key strategy is to combine game theory with state-of-the-art techniques that have been developed to analyse the decision-making process in the human brain (De Quervain *et al*, 2004; King-Casas *et al*, 2005; Singer *et al*, 2006; Knoch *et al*, 2006). How far this truly interdisciplinary research will ultimately bring us in understanding how the brain regulates human social behaviour remains to be seen. For the time being, however, anyone interested in human decision-making and social interaction is well advised to keep an eye on the field and its remarkable progress.

REFERENCES

Arrow KJ (1974) *The Limits of Organization*. New York, NY, USA: Norton

Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econ Behav* **10:** 122–142

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

Carter CS (1998) Neuroendocrine perspectives on social attachment and love. *Psychoneuroendocrinology* **23:** 779–818

Coleman JS (1990) *Foundations of Social Theory*. Cambridge, MA, USA: Belknap

de Quervain DJ, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, Fehr E (2004) The neural basis of altruistic punishment. *Science* **305:** 1254–1258

Heinrichs M, Baumgartner T, Kirschbaum C, Ehlert U (2003) Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biol Psychiatry* **54:** 1389–1398

Insel TR, Young LJ (2001) The neurobiology of attachment. *Nat Rev Neurosci* **2:** 129–136

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

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

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

Luhmann N (1968) *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*. Stuttgart, Germany: F. Enke

Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, Frith CD (2006) Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439:** 466–469

Michael Kosfeld is Assistant Professor of Behavioural Economics at the University of Zurich, Switzerland.
E-mail: kosfeld@iew.unizh.ch

doi:10.1038/sj.embor.7400975