What Human Trust Is and Is Not: On the Biology of Human Trust

Sviatoslav Braynov
Computer Science Department, University of Illinois at Springfield
sbray2@uis.edu

Abstract

The paper presents an overview of the empirical evidence and current research in neuroscience, behavioral economics, and biology showing that human trust is deeply biologically grounded and different from our intuitive understanding of trust as volitional, rational, and conscious activity. Empirical evidence conclusively shows not only that human trust is more complex than mere risk taking, but also that it is implemented in different brain regions and influenced by deferent neurochemistry. The paper raises important research questions as to whether it is trust that we really model and how to formalize the non-calculative, biologically-driven elements of trust.

Introduction

Research on trust and autonomous systems traditionally conceptualizes a world populated by rational intelligent agents guided by critical reason and sound judgment, ignoring or ruling out the powerful role biology plays in human behavior. Many of our likes, dislikes, preferences, choices, social affiliations, and even moral and decision values are intrinsically affected by our brain neurochemistry, which underlines cognition, and in turn, is affected by cognition.

The biological nature of human trust is further complicated by the intricate and intense feedback loops between brain and body. Many human decisions are based on preattentive information processing performed outside the conscious brain and are determined or highly influenced by visceral signals. For example, according to the theory of interoception (Craig 2002), a vast amount of internal bodily sensations are processed in the brain, consciously or unconsciously, affecting our feelings, emotions, rationality, and self-consciousness. In many situations, human consciousness acts as a mere bystander observing an unconscious decision already made and acted upon. A typical example are Libet's (1983, 1985) experiments showing that unconscious neuronal processes in the brain prepare a volitional action approximately 300 milliseconds before a conscious decision is made to perform the action. In other words, unconscious cerebral processes precede and potentially cause a conscious intention, which is later felt to be consciously decided by the subject.

The paper presents an overview of the empirical evidence and current research in neuroscience, behavioral economics, and biology showing that human trust is deeply biologically grounded and different from our intuitive understanding of trust as volitional, rational, and conscious activity. We show that human trust is different from calculative risk taking because trust is often beyond volitional control and is influenced by subconscious brain circuits and hormones. The biological nature of human trust also suggests that humans cannot always trust computers the same way they trust other humans.

Trust versus Risk Taking

The simplest and most common model of trust used in game theory, neuroscience, and psychology is the trust game shown in Figure 1. This is a one-shot game between a trustor and a trustee, in which the trustor moves first and chooses between trusting and not trusting. If the trustor decides to trust, the trustee can either honor or abuse trust. In order to qualify for a trust game, it must satisfy the following requirements:

- Placing trust in the trustee puts the trustor at risk. By moving first, the trustor becomes fully dependent on the trustee's decision which affects both the trustor and the trustee. The trustor receives a gain G if the trustee honors trust and suffers a loss L if trust is abused. Note,

the top payoff in Figure 1 goes to the trustor and the bottom payoff goes to the trustee.

- There is a temptation to abuse trust, i.e., the trustee is better off by abusing trust. Here, α represents the payoff increase the trustee receives from abusing trust. In other words, trustworthiness goes against the trustee' self-interest, i.e., honoring trust benefits the trustor at a cost for the trustee.
- Both the trustor and the trustee are made better off from trusting compared to not trusting, i. e., G > 0 and A > 0.

Although many researches often object that the trust game is too simplistic to capture numerous aspects of trust, such as social norms, personal relations, communications and many others, that is precisely the point of the game—it requires pure trust, i.e., trust without any external factors that would promote or affect trust.

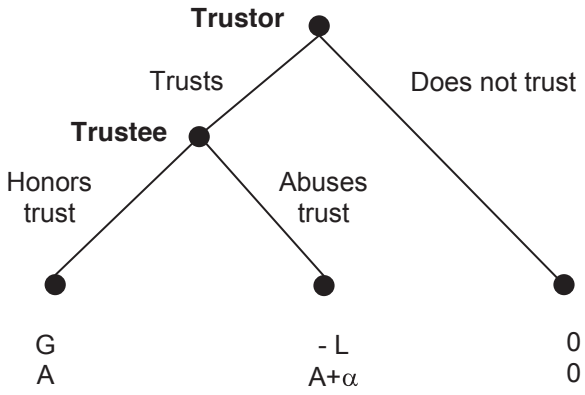

Figure 1. The trust game

The fundamental assumption in economics, game and decision theory is that people act in their own self-interest by choosing a course of action with the highest payoff. Any behavior that deviates from self-interest is viewed as irrational. In the context of the trust game, the only subgame perfect Nash equilibrium is the one in which the trustor does not trust and the trustee abuses trust. That is, if the payoff-maximizing trustee is given the chance to play, he should abuse trust because abusing trust yields a higher payoff than honoring trust: A+α > A. The trustor anticipates this and thus chooses not to trust. Such an outcome is socially (Pareto) inefficient because both the trustor and the trustee would have been better off had trust been placed and honored.

The problem with the game-theoretic solution of the trust game is that it cannot explain how trust emerges in short social interactions, such as the trust game. Experiments in behavioral game theory (Camerer 2003) show that, contrary to the game-theoretic predictions, people do place and reciprocate trust in one-shot games with anonymous counterparts. This problem has led to the development of alternative game-theoretic models of trust. For example, Kreps (1990) was one of the first to develop a model in which trust is developed as reputation building in a sequence of games. Braynov (2001, 2006) has shown that repeated trust games do not necessarily require complete trustworthiness. The trustee can alternate between honoring and abusing trust at some predetermined frequency which is high enough to sustain acceptable payoffs for the trustor.

While the theory of repeated games is useful for explaining how trust emerges and becomes stable in repeated interactions, it fails to provide valuable insights into how trust appears in the absence of repeated interactions, reputation, contracts, punishments, threats, social norms, and other enforcement mechanisms. Limited attempts have been made to explain trust using fairness and inequality aversion (Fehr and Schmidt 1999; Rabin 1993). The theory of fairness is motivated by the fact that people tend to behave nicely toward those who treat them nicely, and behave meanly towards people who are mean to them. One shortcoming of this theory is that it relies on people's judgments of whether others are nice or mean. Such judgments are often biased because of their reliance on individual perceptions, beliefs, predispositions, and cultural norms. Moreover, trust and fairness are different concepts, which cannot be substituted for each other.

Berg, Dickhaut, and McCabe (1995) were among the first to show that game theory does not accurately portray how real people play the trust game, and that trust and reciprocity do occur in one-shot trust games. They conducted multiple experiments with the following version of the trust game, known as the investment game. Two participants are randomly and anonymously matched, one as an investor and one as a trustee. Both participants receive a $10 show-up fee. The investor can send some, none, or all of his $10 to the anonymous trustee. Both the investor and the trustee are informed that the experimenter triples the amount the investor transfers to the trustee. After receiving the transfer, the trustee decides how much of the tripled money to keep and how much to send back to the investor. The investment game is shown in Figure 2, where t denotes the amount sent by the investor and r denotes the amount returned by the trustee, with $0 \le t \le 10$ and $0 \le r \le 40$. t could be used as a measure of the investor's trust and r as a measure of the trustee's trustworthiness. Obviously, the investment game is strategically equivalent to the trust game, and therefore, the unique Nash equilibrium prediction for the investment game is for the investor to send zero money. This prediction is rejected by multiple experiments conducted by Berg, Dickhaut, and McCabe showing that 30 out of 32 investors sent money ($5.16 on average) to their respective trustees and 11 of these 30 transfers resulted in paybacks

greater than the amount sent. The results of Berg, Dickhaut, and McCabe were further confirmed and expanded by various different studies (Camerer 2003; Chapter 2.7). Johnson and Mislin (2011) collected and analyzed data from 162 replications of the investment game conducted by different researchers, involving 23,924 participants from 35 countries. Their analysis indicates that that the amount sent is significantly affected by whether the game is played with a human or a computer counterpart, with subjects sending less money in the investment game when interacting with a computer.

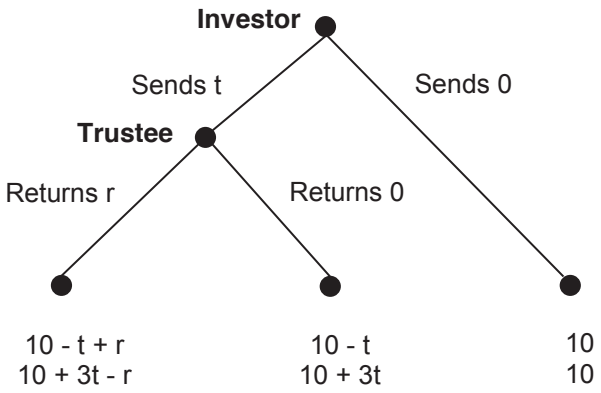

Figure 2. The investment game

The main reason why classical game theory cannot provide a realistic account of trust is the assumption that our behavior is completely volitional, i.e., it is entirely determined by our mind, which calculates the benefits and costs of each action available to us and weighs these benefits by the probability of their occurrence. According to classical game theory, we are disembodied walking computers that do nothing but optimize, thereby ignoring the complex human biology that underlines and determines the structure and the function of the brain.

Williamson (1993) was one of the first to distinguish between trust and risk taking. He refers to pure trust as personal trust and to risk taking as calculative trust. Calculative trust agrees with the theory of rational choice, according to which the trustor decides whether to trust the trustee on the basis of expected utility calculations. Personal trust, on the other hand, is not calculative and is based on feelings, personal relations, and the absence of conscious monitoring. According to Williamson, "The practice of using "trust" and "risk" interchangeably should therefore be discontinued." The same distinction has been made by Dunn (1988) who traces the philosophical roots of trust with respect to political agency. Dunn differentiates between trust as passion and trust as a modality of human action, with the main differences between the two being that passion is based on feelings whereas modality of action is based on conscious decision making and risk taking.

All these findings suggest that there is more to trust than just risk taking. If trust was reducible to risk taking, then humans would have never exhibited spontaneous trust in anonymous strangers and would have never chosen to trust their anonymous counterparts in the investment game in the absence of any trust-building mechanisms, such as social affiliations, norms, reputation, contracts, guilt, etc.

Another argument supporting the difference between trust and risk taking is that humans can certainly choose their own actions under risk, while they often do not have control over and cannot choose their own feelings. If trust was always a conscious decision, then how would one explain the fact that people can instantaneously form perceptions of trustworthiness just by looking at other people's faces? Deciding whether an unfamiliar person is trustworthy is an important and frequently made decision in social situations, which is greatly affected by automatic judgments of perceived facial trustworthiness. Willis and Todorov (2006) have shown that one hundred milliseconds of exposure to neutral faces is sufficient for people to make judgments of facial trustworthiness. Using functional magnetic resonance imaging (fMRI), Engell, Haxby, and Todorov (2007) have demonstrated that such decisions of trustworthiness are unconsciously and automatically made in a brain region known as the amygdala.

Oftentimes, people do behave as predicted by game theory and calculate the expected value of trust by taking into account the future benefits of cooperation, the risk of defection, and the history of previous interactions. Calculative trust, however, represents only one side of the coin. The other side of trust, not less important, is the automatic, instinctive, and subconscious trust, which is driven by our biological nature and is beyond the control of our mind. We argue that even calculative trust is affected by human biology because projecting and weighing future risks is inevitable "felt" by the human body, triggering various reactions in the body and the brain. Recent research in neuroscience and psychology (Coates 2012) shows that when we calculate and project future risks, we do much more than thinking about it. Our body and the brain, expecting an action, start preparing for it physically. In other words, trust is either biologically driven or at least affected by human biology. There is no pure calculative trust entirely based on pure reason and cold mind.

Trust and the Unconscious Brain

The main difference between calculative and non-calculative trust is in the brain mechanisms underlying them. Krueger et al. (2007) used hyperfunctional magnetic

resonance imaging (hyperfMRI) to study how strangers play a repeated non-anonymous alternating-role investment game with the same partner. They showed that calculative trust and non-calculative trust activate different brain regions. Calculative trust selectively activates the ventral tegmental area (VTA), a brain region closely related to the dopamine-based reward system and the evaluation of expected and realized rewards. Non-calculative trust, on the other hand, preferentially activates the septal area (SA), a region linked to social attachment and the release of oxytocin, a neuropeptide that influences brain activity and promotes not only trust and trustworthiness but also pair bonding and social attachment. Recent studies (Zak 2012) have dubbed oxytocin "the moral molecule" because of the important role it plays in promoting trust in social interactions.

Another difference between calculative and non-calculative trust found by Krueger et al. (2007) is that calculative trust is more cognitively costly to maintain than non-calculative trust and requires the constant activation of the paracingulate cortex (PcC), a brain area used to represent people's thoughts, beliefs, and the mental states of other people. In other words, calculative trust requires constant mentalizing and inferring each other's mental states and intentions. The maintenance of non-calculative trust, on the other hand, is based on attachment and requires less cognitive processing. Non-calculative trust activates PcC in the initial steps of trust development when the partner's trustworthiness needs to be verified. After the partner's trustworthiness has been established, the use of PcC is extinguished and replaced by the activation of SA, thereby producing more accurate and faster trust decisions than calculative trust. In general, Krueger et al. (2007) showed that non-calculative trust allows partners to achieve higher levels of synchronicity in their trust decisions and cooperate more often than calculative trust that tends to produce higher errors in inferring the partner's trusting intention and results in greater variance in cooperative decisions. In other words, non-calculative trust proves to be faster, less cognitively expensive, and more efficient in promoting cooperation than calculative trust.

Another important feature that sets non-calculative trust apart from calculative trust is that non-calculative trust could be unconscious or emotionally driven due to the involvement of the amygdala, part of the limbic system playing an important role in emotional reactions and memory modulation. Winston et al. (2002) have shown that the amygdala is critically involved in the assessment of facial trustworthiness. Viewing untrustworthy faces activates the amygdala regardless of whether the judgments are made consciously or unconsciously. Moreover, trust is increased when the amygdala is damaged (Adolphs et al. 1998).

The involvement of the amygdala in trust assessment can be linked to the central role it plays in emotional processing functions, such as fear extinction and anxiety. It seems reasonable to assume that suppressing the amygdala activity would limit fear of trust betrayal in social interactions and would increase trust in humans. Kosfeld et al. (2005) were the first to show that intranasal administration of the neuropeptide oxytocin causes a significant increase of trust among humans. More specifically, they have demonstrated that players who received oxytocin infusions were more trusting than a placebo control group in the investment game. The effect of oxytocin on trust is due to the ability of oxytocin to decrease risk aversion and increase readiness to bear social risks. The main contribution of Kosfeld et al. is to show that the brain distinguishes between social trust and non-social risk taking. Oxytocin increases willingness to take risks only in social interpersonal interactions. When the investors played the investment game with a machine implementing a random strategy, the investors' behavior did not differ between the oxytocin and the placebo group. In other words, oxytocin promotes social non-calculative trust and has no effect on monetary risk taking.

Kirsch, Esslinger, and Chen (2005) linked the effect of oxytocin on trust to the amygdala. They used fMRI to study how oxytocin modulates the activation of the amygdala in response to fear-inducing visual scenes and facial expressions. Their experiments showed that oxytocin suppresses the activation of the amygdala and the functional connectivity between the amygdala and the brainstem effector sites for fear response, thereby promoting social trust and encouraging social risk taking. Kirsch, Esslinger, and Chen have also found that the reduction in amygdala activation was more pronounced for socially relevant stimuli than for less social relevant scenes. This confirms the findings of Kosfeld et al. (2005) that oxytocin promotes trust by suppressing fears arising from social interactions, not fears in general.

Another piece of relevant evidence comes from the study of Baumgartner et al. (2008), which uses intranasal administration of oxytocin in an investment game to show that oxytocin affects trust only when it also dampens amygdala activity. In the experiment, subjects receiving placebo reduced their trust after they were informed that the trustee did not pay back in about 50% of the cases. On the other hand, subjects receiving oxytocin showed no change in their trust after they were presented with the same information. These findings are consistent with previous studies showing that oxytocin increases one's willingness to take social risks in interpersonal interactions.

Trust and Hormones

Hormonal effects on human behavior have been especially noticeable with respect to trust and trustworthiness. It has been repeatedly shown that higher oxytocin levels are associated with trustworthy behavior (Zak 2005a; Zak 2012; Heinrichs and Domes 2008).

Surprisingly, Zak et al. (2005a, 2102) have shown that the relation between trust and oxytocin is bidirectional in the trust game—not only does oxytocin increase human trust in social interactions but trust has a positive feedback effect on oxytocin. Oxytocin levels are higher in trustees that receive money from the trustor as a result of trusting intention relative to unintentional money transfer. In other words, oxytocin increases the trusting behavior of the trustor and makes him more willing to transfer money to the trustee. The very fact of being trusted raises the trustee's oxytocin levels, making him even more trustworthy!

Bos et al. (2010) conducted a placebo-controlled experiment showing that testosterone, a steroid hormone associated with competition and dominance, acts as an antidote to oxytocin and decreases interpersonal trust. The experiment involved the administration of testosterone or a placebo to female subjects who were subsequently asked to evaluate the trustworthiness of a series of human faces shown in photographs. The females who received testosterone showed a significant overall reduction in trustworthiness ratings compared with the placebo group.

Zak et al. (2005b) reported similar results on the effect of testosterone on trust in a trust game. When male subjects were distrusted, they experienced elevated levels of a derivative of testosterone called dihydrotestosterone, which promotes aggression and boosts the desire for physical confrontation. The increase in dihydrotestosterone was directly proportional to the amount of distrust experienced by the male subjects. Moreover, the relationship between distrust and dihydrotestosterone was stronger in men than in women.

The difference between trust and monetary risk taking becomes more pronounced if one takes into account the effect of testosterone on human behavior. Whereas testosterone decreases trust in social interaction, it increases appetite for financial risks. In general, testosterone is released into the body during moments of competition, risk taking, and triumph. Coates (2008, 2012) sampled testosterone levels in traders in the City of London and found that higher levels of testosterone in the morning correlate with higher financial profits at the end of the day. When traders make money, their testosterone levels rise, and as a result, successful traders go into the next day primed with even higher levels of testosterone, helping them to succeed again, and creating a winning streak marked by elevated confidence and appetite for risk. At some point of the winning streak, the upward spiral of testosterone begins to have an opposite effect on traders by making them overconfident, convinced of their own invincibility, and reckless. Eventually, too much testosterone leads to huge losses and financial failures.

Trusting Humans versus Trusting Computers

Since the outcome of strategic interactions, such as the trust game, is determined by the joint action of both players, it seems quite natural that a player's decision depends on what he believes about the beliefs, desires, and intentions of the other player. The ability to mentalize, infer, and understand implicitly or explicitly other people's mental states is known as theory of mind (ToM), and it often relies on the activation of the paracingulate cortex (PcC). McCabe et al. (2001) used fMRI in the investment game to test the hypothesis that trust and reciprocity require ToM. Their study reported that PcC is more active when subjects are playing a human than when they were playing a computer following a fixed probabilistic strategy.

The results suggest that although computers can model and simulate mental states, such as beliefs, desires and intentions, people do not perceive computers as having locus of mentality and performing ToM in the trust game.

Similar results were reported by Baumgartner et al. (2008) who showed that oxytocin increases trust in investment games against humans and has no effect on trust in games against computers. In the experiment, subjects receiving a placebo decreased their trust after being informed that the trustee did not pay back in about 50% of the cases, whereas subjects receiving oxytocin showed no change of behavior after they were presented with the same information. When the game was played against a computer who implements a fixed random strategy, both subjects in the placebo group and the oxytocin group did not change their trust after receiving the feedback information. Therefore, it seems that oxytocin promotes trust if interpersonal interaction and social risks are involved, and has no effect on trust if nonsocial risks are involved. The results are consistent with previous studies (Kosfeld et al. 2005) showing that oxytocin increases trust between humans, not trust between humans and machines. The aggregated data from the investment game also show that people tend to trust computers less than they trust other people (Johnson and Mislin 2011).

Conclusions

Based on the empirical evidence and current research in neuroscience, behavioral economics, and biology, we can conclude that:

- Human trust is often beyond volitional control, initiated and supported by automatic and unconscious responses in both the brain and the body.
- Even when trust involves an explicit calculation of the expected risks, benefits, and costs of cooperation, trusting decisions are implicitly influenced by subconscious brain circuits, body signals, and somatic markers. In other words, human trust is intrinsically biological and does not exist in the form of pure reason.
- Humans cannot always trust computers the same way they trust other humans. Since computers do not have the subconscious and the hormonal mechanisms that affect bond formation, social attachments, and social risk aversion, human interaction with computers is structurally and functionally different from human trust. Interaction with computers activates different brain regions and lacks many features of human trust. Trust in computers is often reducible to simple risk taking under uncertainty.

These findings give rise to important research questions as to whether it is trust that we really model and how to formalize the non-calculative, biologically-driven elements of trust.

References

Adolphs, R., Tranel, D., Damasio, A. 1998. The Human Amygdala in Social Judgment. *Nature* 393:470–474.

Baron-Cohen, S. 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MIT Press.

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(4):639-650.

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

Bos, P., Terburg, D., van Honk, J. 2010. Testosterone Decreases Trust in Socially Naïve Humans. *Proceedings of the National Academy of Sciences* 107(22): 9991-9995.

Braynov, S., Sandholm, T. 2001. Contracting with Uncertain Level of Trust. *Computational Intelligence* 14(4):501–514.

Braynov, S. 2005. Trust Learning Based on Past Experience. *In the Proceedings of the IEEE Knowledge Intensive Multi-Agent Systems* 197-201.

Camerer, C., F. 2003. *Behavioural Game Theory: Experiments on Strategic Interaction*. Princeton University Press.

Coates, J., Herbert., J. 2008. Endogenous Steroids and Financial Risk Taking on a London Trading Floor. *Proceedings of the National Academy of Sciences* (16):6167–6172.

Craig, A., D. 2002. How Do You Feel? Interoception: The Sense of the Physiological Condition of the Body. *Nature Reviews Neuroscience* 3(8): 655-66.

Decety, J., Grèzes, J. 2006. The Power of Simulation: Imagining One's Own and Other's Behavior. *Brain Research* 1079:4–14.

Dunn, J. 1988. Trust and Political Agency. In Gambetta, Diego (ed.) *Trust: Making and Breaking Cooperative Relations* 73-93.

Engell, A., Haxby, J., Todorov, A. 2007. Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience* 19(9):1508–1519.

Fehr, E., Schmidt, K. 1999. A Theory of Fainess, Competition and Cooperation. *Quarterly Journal of Economics* 114:817-868.

Heinrichs, M., Domes, G. 2008. Neuropeptides and Social Behavior: Effects of Oxytocin and Vasopressin in Humans. In I. Neumann and R. Landgraf (Eds.*) Progress in Brain Research* 170:1–14.

Johnson, N., Mislin, A. 2011. Trust Games: A Meta-Analysis. *Journal of Economic Psychology* 32 (5):865–889.

Kirsch, P., Esslinger, C., Chen, Q. 2005. Oxytocin Modulates Neural Circuitry for Social Cognition and Fear in Humans. *The Journal of Neuroscience* 25 (49):11489–93.

Kosfeld, M., Heinrichs, M., Zak, P., 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. *Proceedings of the National Academy of Sciences of the USA* 104(50):20084-20089.

Libet, B. 1985. Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *The Behavioral and Brain Sciences* 8:529-566.

Libet, B., Gleason, C., Wright, E., Pearl, D. 1983. Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness Potential): The Unconscious Initiation of a Freely Voluntary Act. *Brain* 106:623-642.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T. 2001. A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange. *Proceedings of the National Academy of Sciences of the USA* 98(20):11832-11835.

Rabin, M. 1993. Incorporating Fairness into Game Theory and Economics. *The American Economic Review* 83:1281-1302.

Williamson, O. 1993. Calculativeness, Trust, and Economic Organization. *Journal of Law and Economics* 36(1):453-486.

Willis, J., Todorov, A. 2006. First Impressions: Making up Your Mind After a 100-ms Exposure to a Face. *Psychological Science* 17:592-598.

Winston, J., Strange, B., O'Doherty, J., Dolan, R. 2002. Automatic and Intentional Brain Responses During Evaluation of Trustworthiness of Faces. *Nature Neuroscience* 5:277–283.

Zak, P., Kurzban, R., Matzner, W. 2005a. Oxytocin is Associated with Human Trustworthiness. *Hormones and Behavior* 48(5): 522-527.

Zak, P., Borja, K., Matzner, W., Kurzban, R. 2005b. The Neuroeconomics of Distrust: Sex Differences in Behavior and Physiology. *American Economic Review Papers and Proceedings* 95:360-364.

Zak, P. 2012. *The Moral Molecule: The Source of Love and Prosperity*. Dutton.