What is trust and what are its origins?

Vinod Goel

York University, Canada

Abstract

1.0 What does it mean to trust

Let us begin by considering the following statements:

1. I trust my friend to pick me up from the airport.

2. I trust my car to stop when I press the brakes.

3. I trust the sun to rise tomorrow morning.

4. I trust the stock market will rise 100 points tomorrow.

While the term "trust" superficially makes sense in each of the above sentences, there seem to be at least three differences to note. The first difference is in the object of my trust, the second in the presence or absence of consent, and the third in the nature of the uncertainty involved. In the first case, the object of my trust is another human agent, someone with whom I have strong social bonds and he has consented to pick me up. When he arrives to pick me up my trust in him is justified and our social bonds strengthened. I will feel grateful and the need to reciprocate at some point. But what if he fails to do so? If he simply decides not to do it and play golf instead, my trust in him will be violated and I will feel betrayed (after all, I picked him up last month) and may well wish to punish him. However, if he fails to pick me up because he had an accident on route to the airport and is in the hospital, I will not feel betrayed or be upset with him. He intended to pick me up, was carrying out the intention, but was it unable to complete it due to factors beyond his control.

If my car brakes fail, I will be upset (irrespective of the reason). I may curse the mechanic who serviced them last week, and lose trust in him and feel that he cheated me… But it is strange to feel betrayed by and punish the car. In the case of the sun failing to rise tomorrow morning, there is not even a mechanic who can even be the object of any feelings of betrayal. Insofar as the stock market is the sum of the independent self-interested actions of numerous individuals (with no concerns for my interest), there is again no notion or object of betrayal when it does not go up 100 points. When my brakes do stop the car, or the sun comes up in the morning, I simply take it for granted. When the stock market rises hundred points, I consider myself a financial genius. The issue of gratitude or reciprocation does not arise in these nonsocial cases. There is nothing to reciprocate. These examples suggest that the term "trust" is being used in very

different senses in example 1 and examples 2-4. If we take example 1 as the natural or literal case of trust, it suggests that the objects of trust are (consenting?) human agents and when trust is justified it results in feeling of gratitude and reciprocity, and when it is broken, feelings of violation and betrayal.

There are also differences in the nature of the uncertainty involved. In the case of the sun rising tomorrow morning the uncertainty is minimal. In the case of mechanical failure of my brakes, the uncertainty is determined in part by my maintenance schedule and the state of wear and tear of my brakes. In so far as I am prepared to drive the car, I must believe it to be on the low side. The movement of the stock market on any particular day is unpredictable, though there seem to be reliable long-term upward trends. In the case of my friend, the uncertainty is not a function of natural law, mechanical failure, or randomness. It is up to him what he does.

Now consider the following variations on example 1:

1a. I trust the taxi driver that I found on the Internet to pick me up from the airport.

1b. I trust the taxi driver that I have found on the Internet and have prepaid to pick me up from the airport.

In the case of 1, the trustee is a friend, someone with whom I already have social bonds and perhaps social capital that I expect to be reciprocated. This sets up social obligations on his part and leaves me vulnerable to loss of previously invested social capital (I picked them up last week). In 1a and 1b the trustee is a *stranger*, and I have no social bonds with them. In 1a I have no pre-existing capital, social or otherwise invested with him/her. I certainly expect them to show up for their personal gain (payment), but if they do not, I am disappointed and annoyed, not necessarily betrayed, and will call another taxi. "Trust" may not be the appropriate term to use in this case. In 1b, however, I do have pre-existing capital with the individual (having paid them beforehand). This prepayment sets up an obligation on their part and makes me vulnerable to loss if they do not fulfill the obligation.

These examples further clarify the notion of trust. Trust need not involve friendship and social bonds and social capital. One can trust strangers with monetary capital in the context of legal and social norms. In both examples 1 and 1b I am entrusting capital (social and monetary respectively) to the trustee in return for an obligation that they consent to. Notice that in both

cases I am vulnerable because I transfer capital (social or monetary) prior to the performance of the obligation and there is always uncertainty whether the obligation will be fulfilled. In my friend's case, that uncertainty is bridged by our social bonds, while in the case of the taxi driver legal norms provide some assurance. If they arrive to pick me up, my trust is justified, and I will reciprocate (by doing something for my friend for calling this taxi driver again in the future). If my trust is violated, I will feel betrayed by their failure to fulfill their obligation.

It is also interesting to note that in each of the above cases I can replace the word "trust" with the word "believe." Beliefs can be true or false depending upon the state of affairs in the world. Beliefs do not set up expectations and obligations and result in feelings of gratitude for betrayal. This suggests that while the term "belief" may be appropriate examples 1a, 2, 3, 4, I actually mean something beyond belief in examples 1 and 1b.

Given these considerations, we can begin with the following definition of trust:

Trust is an *Intentional state* that the trustor freely directs at another individual or group, with the expectation that

- the trustee will recognize and consent to an obligation that benefits the trustor due to a *prior* transfer of social or monetary capital to the trustee;
- the action of the trustee will be motivated/caused by the relevant or right intention (i.e. discharging of the obligation);
- the trustee has the ability to do otherwise, rendering the trustor vulnerable to loss;
- if the trustee does otherwise, the trustor stands to lose their invested capital + X (where X = feelings of betrayal) and gain invested capital + Y + Z (where Y = return on investment and Z = feelings of gratitude) if they do not do otherwise.

The first criterion involves the trustee recognizing and accepting (via consent) an obligation to the trustor resulting from the transfer of social and/or monetary capital. In the absence of consent I do not expect my friend to be at the airport to pick me up. Just because he has always picked me up in the past is not sufficient to assume he will pick me up today. The consent affirms he will pick me up, but it is the *obligation* that binds him.

The second criterion rules out cases where my friend forgets to pick me up from the airport but nonetheless arrives at the airport at the designated time and place to pick up another friend,

and upon seeing me, offers me a ride home. In this case he acted upon the intention to pick up the other friend and not me. Even though I do get my ride home, it is by sheer chance that I arrived at the same time and place as his other friend. I would be justified in feeling betrayed that he did not come to pick me up as I expected him to (and he said he would). However, my *belief* that my friend will pick me up is satisfied in both cases.

The third criterion is that the trustee has the ability to do otherwise. The ability to do otherwise introduces a special notion of risk or uncertainty into the expectation, different from random chance or mechanical failure. An individual may freely change their mind and not do what they said they would do at their discretion. They are after all free agents with their own goals and desires. This heightens the element of risk and uncertainty involved in the interaction. The psychological state of trust bridges this uncertainty gap and binds the trustor and trustee, allowing the former to place themselves in a position to be vulnerable to the free actions of the latter. If the individual I trust does not have the ability to do otherwise there is actually no need for trust. The issue does not arise. I can simply act on my belief that they will do what they said they would. This criterion brings in the notions of free will and the reactive attitudes (Strawson, 2008) as necessary properties of both the trustor and trustee (discussed below).

The fourth condition attaches a cost to me for the violation of my trust and a benefit to me for its fulfillment. The cost and benefit always have a non-monetized component: feelings of betrayal or gratitude. Imagine two situations in which you invest money. In the first case the investment company literature promises you that your capital will be safe and you will earn 10%. In the second case, the investment company literature says your investments can go both up and down. You invest $1000 with both companies. You end up losing 50% of your funds with both. The monetary loss is the same in both cases but there is a difference in how you feel and behave. This difference is what we refer to as betrayal (directed at the first company). In the case of a gain or benefit I will feel gratitude and do business with them again

On this account, trusting is more than beliefs (though it may of course involve the belief that the trustee is trustworthy). Trusting involves making oneself vulnerable to the actions of others; beliefs do not. Beliefs can be true or false. Trust is fulfilled or violated. The former is associated with feelings of gratitude (and willingness to trust again); the latter is associated with feelings of betrayal.

In the example of my friend picking me up from the airport, the individual is both known to me and is making a commitment to do something. In the examples of the prepaid taxi in 1b and the two investment companies above, I do not have a social/friendship relation with them, but they are making explicit representations to do certain things in the context of social and legal norms. Another notion of trust involves interactions between strangers with no explicit commitments stated and no social or legal norms in place. Interestingly, this has become the standard scenario for studying trust in the field of behavioral economics in the form of the "trust game"

The trust game (Berg et al., 1995) provides an operationalized behavioral definition that allows for direct measurement of trust. It is an economic cooperation game with two players who are strangers to each other. Both players are given an equal amount of money. The first player (trustor or investor) has the option of transferring some arbitrary portion of his money to the second player (trustee), with the understanding that the experimenter will triple any amount that is transferred. The trustee can then decide whether to keep all the funds or send a portion back to the trustor. If the first player decides not to transfer any funds to the second player, each player keeps the initial funds. However, given the tripling rule, the self-maximizing choice for the first player is to transfer all their funds to the second player, as long as the second player then transfers half the tripled amount (or at least more than they received) back to the first player. This way, both players come out ahead.

But there is a danger. What if the second player violates the trust/fairness and fails to reciprocate (keep everything for himself)? If funds are transferred, it is immediately self-maximizing (hence rational) for the second player to keep all the proceeds and not send anything back. In this situation, the self-maximizing outcome is distant for the first player, leaving them vulnerable and relying on fairness/reciprocation, while the self-maximizing outcome for the trustee is immediate and relies on cheating.

The initial transfer is a direct measure of trust that the first player is placing on the second player and the transfer back by the second player is a reciprocation of that trust (and a measure of trustworthiness). When the trustee proves untrustworthy and betrayal does occur, it is defined and measured in terms of "the pure disutility of experiencing or anticipating nonreciprocal trust" (Fehr, 2009). In this case the reduction in the trustor's overall utility will be greater than the

reduction in their monetized utility. This additional component is the feeling of being a "gullible sucker."

It is important to appreciate the differences in this behavioral measure of trust and of the examples of my friend or prepaid taxi driver picking me up from the airport. In the case of my friend there is an agreement or consent that creates an obligation for the trustee. I expect the trustee to perform the obligation due to the social bonds binding us. In the case of the prepaid taxi there is also an agreement leading to an obligation, but in this case the obligation can be reinforced by the institution of the law. In both cases I do need to trust the other party (i.e. to make myself vulnerable) to the possibility that they will fail to carry out their obligations. In both cases I will feel betrayal (but will have the opportunity to punish).

In the case of the trust game, the other player is a stranger so there are no social bonds between the trustor and the trustee. Furthermore, the rules of the game are such that the second player is not legally obligated to return any funds and can keep everything. But nonetheless there is an implicit obligation and a recognition of such by the trustee. The source of the obligation is that I am advancing them some funds (making myself vulnerable to loss) from which they can automatically earn a reward (without any additional effort on their part). In this circumstance the right or fair thing for them to do would be to share a portion with me. Without my initial advance, they would have nothing. In this scenario X percentage of players are willing to trust a stranger and transfer funds and Y percentage of players reciprocate by transferring funds back, proving some natural propensity for trusting and trustworthiness. In fact, most players choose to make a substantial transfer, and the transfer back made by the second player correlates with the amount of the initial transfer (Eimontaite et al., 2013; Fehr & Fischbacher, 2003). The failure of reciprocation results in feelings of being cheated and betrayed and a call for punishment on the transgressor (Fehr & Fischbacher, 2004).

These properties of the trust game require a modification to the first criterion of our above definition of trust. The criterion that "the trustee recognizes and consents to an obligation due to a prior transfer of social or monetary capital" can be changed to "the trustee recognizes an obligation due to a prior transfer of social or monetary capital." No explicit consent is required in the trust game. And implicit obligation automatically arises with the transfer of funds by the

trustor. It is created by our innate recognition of reciprocity. The issue then becomes whether the trustee will respond fairly and recognize and discharge the obligation or cheat.

In some sense, this trust game task does seem to provide a clean measure of trust, independent of any social or legal bonds. However, it is important to recognize potential complications when trustors advance funds for purely altruistic reasons independent of expectations of reciprocity (Cox et al., 2008).

2.0 Some findings from the Trust Game

Many hundreds, if not many thousands, of studies of trust have been undertaken using the trust game or some variation of it. Among these findings are that trusting is more than just risk-taking, reason-based factors can modulate trust, emotions can modulate trust.

2.1 Trusting is more than just risk-taking:

Earlier we noted that trusting is more than overcoming risk/uncertainty aversion. We encounter uncertainty in the natural, mechanical, and random worlds, as noted in examples 2, 3, and 4, but the issue of trust only arises in the social world. Trust bridges socially constituted risks. Variations on the trust game allows us to empirically test the claim.

If trusting was just risk-taking there should be no difference in the behavior of the trustor to the trustee (in terms of funds transferred) irrespective of whether the trustee was another human or a computer, as long as the expected utility of trusting was the same in the two cases.

Comparison of the trust game with an identical decision-making game where the trustee is replaced by a random lottery (Bohnet et al., 2008) calculated and compared minimum acceptable probability in the two cases. If the minimum acceptable probability is different in the two cases, then one may infer a difference between risk aversion alone and trust aversion. Bohnet et al. (2008) report higher minimum acceptable probability for the Trust Game than another equivalent decision problem, suggesting that there is more to trust than simple risk aversion.

In another study they used a variation of the trust game (Risky Dictator Game) where there is a trustee but they cannot make the decision on the amount to return. It is a random assignment beyond their control. In this case there can still be socially constituted risks but betrayal is presumably ruled out because the trustee cannot do otherwise. The difference in minimum

acceptable probability in the regular trust game and the risky dictator game can be considered a measure of betrayal aversion and is a significant result found across several countries (Bohnet et al., 2008).

These results reinforce the third criterion of our definition and emphasize that trust requires a trustee susceptible to what Peter Strawson (2008) called the reactive attitudes. The reactive attitudes are "natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions." (p. 344)… They include gratitude, resentment, forgiveness, love, and hurt feelings (p. 342).

2.2 Reason-based factors can modulate trust

Any knowledge, beliefs, and perceptions about the trustworthiness of the other player (i.e., their reputation) modulates the initial transfer amount (i.e. trust) as well as the returned amount (i.e. reciprocation) (Berg et al., 1995). If there are repeat trials of the game, with the trustor and trustee alternating, then trustors become even more "generous," because they know they will be at the receiving end in the next trial and will have to deal with the consequences of their reputation for defecting or cooperating.

Beliefs about the trustworthiness of other members—known as their *reputation*—is established by their history of choices in previous interactions. It is transmitted either directly (through firsthand knowledge of previous choices) or indirectly, via language or some other system of symbols. Either way, potential trustors are less likely to trust (i.e., more likely to punish) those who have previously violated fairness norms. In repeated game interactions, where both individuals know that the other will have knowledge of their past interactions, cooperation rates rise dramatically. In fact, individuals are aware of the value of a good reputation and will expend resources to gain one. In an experimental situation where an individual has the possibility of developing a positive reputation, the cooperation rate rises from 37% to 74%. Reciprocity and a good reputation reinforce each other (Gächter & Falk, 2002).

2. 3 Emotions Can Modulate Generosity of Trustors

The amount transferred by the trustor can be manipulated by reasons other than levels of trust and anticipated reciprocation. If certain beliefs are instilled in the trustor about the trustee, such as, the trustee is dying of cancer this will increase the amount of the initial transfer to the trustee

for altruistic reasons. If the trustor is led to believe that the trustee has recently insulted them, the amount of the initial transfer will decrease, again having to do with anger and resentment (Eimontaite et al., 2013, 2019).

In these manipulations the instilled beliefs—albeit with important emotional components—modulated the outcome of the choice. Beliefs do not even have to be explicitly instilled. A manipulation whereby the two players spend a few minutes silently looking at each other increases generosity in single-shot games compared to totally anonymous interactions (Bohnet & Frey, 1999). This process allows not only for the humanization of the other player but also for identification for future interactions (i.e., it raises concerns about reputation).

Participants are also sensitive to facial expression bias. Ewing et al. (2019) carried out the trust game with children and adults and manipulated facial expressions of the trustee along the lines of mild happiness, anger, and neutral expressions. They reported that all participants were more generous to happy partners, followed by neutral expressions, with angry being last.

3.0 Origins of Trust: Cost-benefit analysis versus innate trait

What are the origins of trust? Is it a rational cost-benefit cognitive calculations based upon our beliefs, previous interactions with the individual, and the goal of maximizing resources. Or is it an evolutionary trait or predisposition drawing upon upon noncognitive mechanisms. The two answers need not be mutually exclusive. The former could be built on top of the latter.

One way of answering this question is seeking answers to the following five questions: (1) is trust universally available in all human societies or appears selectively? (2) is it available on other branches of the phylogenetic tree? (3) does it appear early in children prior to extensive development? (4) is it underwritten by implicit, automatic low-level mechanisms? (5) is it possible to trace specific subcortical neural circuitry and neurochemistry devoted to it (and, where relevant, find homologous behavior and circuitry in other species)?

3.1 Is Trust Universal?

There is evidence that levels of trust vary across and within societies (Finuras, 2023; Zak, 2008) but we know of no society in which trust does not exist. Johnson and Mislin (2011) review the literature on trust games from a dozen plus countries and report similar results.

3.2 Is trust uniquely human or does it exist on other branches of the phylogenetic tree?

Trust would seem to be a prerequisite for cooperative behavior. There are many species ranging from ants to wolves to chimpanzees to humans that cooperate in social groups. Is it appropriate to invoke trust in each of these cases? The short answer to this question is no.

Trust requires the trustor to go out on a limb and make the first move by making themselves vulnerable to loss. Why would we do this? What systems need to be in place to justify making the first move? An organism would need to possess the following traits: (1) personal benefit maximization; (2) recognition of fairness and contingent reciprocity in oneself and projection of it onto conspecifics; (3) ability to detect cheaters; and (4) ability to punish cheaters. The presence of these traits may predispose organisms to trust strangers with a probability of slightly over 50%. The probability will increase substantially in the case of kinship bonds, social bonds, and (in the case of humans) legal bonds.

If this is the case, then contingent reciprocity and cheater detection and punishment would need to be prerequisites for the emergence of trust. This is a high bar to overcome because these traits come with high encephalization demands. For example, among the most highly socially organized cooperative organisms are bees and ants. While we lack substantive understanding of the proximal mechanisms of cooperative behavior in bees and colonies, we can be certain that they lack the neural resources for these prerequisite traits and their cooperation can be explained much more simply in terms of mechanisms such as chemical bonding (McCreery & Breed, 2014). These chemical bonds obviate the need for something like trust to bring about cooperation.

But how about on branches closer to us on the phylogenetic tree? Do chimpanzees engage in trusting behavior? The answer here is not clear cut. On the one hand there are data indicating that chimpanzees do not engage in a calculated or contingent reciprocity, but on the other hand there is also some data to indicate a certain form of trust.

Chimpanzees live in social groups and cooperate with unrelated partners. Males cooperate in patrolling territory, hunting, sharing food, grooming, and joint mate guarding, and even form within-group coalitions for aggressive actions against other members. Despite all these prosocial behaviors, in experimental settings they will not volunteer to help another familiar but unrelated

individual obtain food, even at no cost to themselves (Silk et al., 2005). In one study, they show a small increase (5.7%) in sharing food with a familiar individual who has groomed them within the past two hours compared to an individual who has not groomed them within this time period (de Waal, 1997). In other studies, they fail to show that their altruism is conditional on reciprocity.

Sarah Brosnan and her colleagues (2009) carried out an experiment on captive chimpanzees to determine whether they would more readily share food with a partner from their home group who had shared food with them on previous trials versus partners who had not shared food with them. Individuals familiar with each other were tested in pairs. One individual was offered a choice between two options: (a) deliver a food reward to themselves and another equal one to the other individual (prosocial behavior) or (b) deliver a food reward to themselves and nothing to the other individual (selfish behavior). On the next trial, the other individual was offered the same choice. The trials were repeated a number of times. Interestingly, the choices individuals made were not affected by the choices that their partners made in previous trials. That is, any food cooperation was not contingent on previous interactions. Several leading primatologists now agree that reciprocal altruism (and hence cheater detection) is nonexistent among nonhuman animals, even including nonhuman primates (de Waal & Brosnan, 2006; Stevens & Hauser, 2004). It may be something unique to humans.

Frans de Waal and Sarah Brosnan (2006) propose three levels of reciprocity, of which the first two can be found among nonhuman animals and the third seems exclusive to humans: symmetry-based, attitudinal, and calculated or contingent reciprocity. The simplest form, symmetry-based reciprocity (i.e., “we are friends”), requires that both parties behave similarly with each other; it is based on existing relationships such as kinship, group membership, alliances, and similarity in age. It does not require scorekeeping. There is a very low degree of contingency. The altruistic behavior of meerkats, Belding’s ground squirrels, and vampire bats would fall into this category. By contrast, attitudinal reciprocity requires that an individual’s willingness to cooperate covary with the recent attitude of the partner (“if you are nice, I will be nice”). Both parties may not benefit simultaneously, but the requirement of scorekeeping is minimal. The contingency is immediate. The exchange is based on “general social disposition rather than specific costs and benefits” (Brosnan, de Waal, & Proctor, 2014, p. 24). The altruistic behavior of chimpanzees

reported here would fall in this category. Finally, in calculated or contingent reciprocity (Trivers's reciprocal altruism), individuals expect reciprocation of at least equal value, though allow for significant time lags. If expectations are violated, cheaters will be punished.

Prerequisites for contingent reciprocity identified by Trivers (1971) are (i) the benefit to the recipient greater than the cost to the donor, (ii) opportunity for repeated interaction, (iii) reasonably flat dominance hierarchies, and (iv) the cheater detection mechanism. Hauser and others (Stevens & Hauser, 2004) have also noted that full-fledged contingent or calculated reciprocity requires quite sophisticated cognitive abilities, such as recognition of individuals, memory of previous events, scorekeeping, numerical discrimination, and even temporal discounting. The only robust examples of it occur in humans. If this is the case, it would seem that trust may be largely a human trait.

But on the other hand, there is evidence for something akin to trust in chimpanzee social interaction. Engelmann et al. (2015) carried out a variation of the economic trust game with chimpanzees. Chimpanzees are placed in two rooms/cages separated by a space which contains a long table-apparatus joining the two cages. On the table is a trolley system containing two separate wells baited with preferred and unpreferred food. Each well is on a trolley attached to a rope which can be reached by the first animal (trustor) in one of the rooms. The animal can pull on the rope attached to the trolley with the food well containing the unpreferred food and access and consume it. However, to access the well baited with the preferred food the animal must pull on the rope of the other trolley but this sends the baited well to the other chimpanzee (the trustee). The other chimpanzee can only access and eat half of the food. They can then pull on their rope and send the trolley back to the original chimpanzee, or not. It is reported that (i) the trustor chimpanzee pulled the trust rope significantly more often in the test condition (where there was a chimpanzee in the other cage) than in the control condition (where there was no chimpanzee in the other cage); (ii) the trustor's trust increased as a function of previous trials if the partner reciprocated (proved trustworthy); (iii) trusting decisions were reciprocated 32% of the time in the first study and 58% of the time in the third study. It is possible that attitudinal reciprocity is sufficient to explain these data.

There is some inconsistency to discuss and resolve here… One important difference between this version of the trust game in the human version is that there is near zero cost to reciprocating

because the trustee cannot physically access half of the food in the preferred food well… Even then they only reciprocate 58% of the time in repeat (positively reinforced) trials…

3.3 Development of trust in children

Is trust exhibited by children early in life? If it is, it would strengthen the view that it is a primitive innate trait. A study by Vanderbilt et al. (2011) addresses the development of trust in children 3 to 5 years of age in the context of informants who gave advice about the location of hidden stickers. Some of the informants were known to the children to be "helpers" and others "tricksters" based on previous observations. The three-year-olds trusted (i.e. followed the advice of) helpers and tricksters indiscriminately at very high levels (95% and 91% respectively). The four-year-olds trusted the helpers 69% of the time and tricksters 60% of the time. Five-year-olds trusted helpers 71% of the time and trusted tricksters 51% of the time. These results suggest a predisposition to trust indiscriminately in young children which becomes more selective with cognitive development, in particular the ability to make theory of mind inferences. A study by Jaswal et al. (2010) tested 3-year-olds using the same paradigm and came to the same conclusion.

Add several more paragraphs…

3.4 Brain Systems of Trust

Add several paragraphs on brain systems…

3.5 Biology of trust

3.5.1 Genetics

Might there be a genetic component to trust? One way of approaching this question is by examining monozygotic (identical) twins and dizygotic (fraternal) twins. Monozygotic twins share the same set of genes whereas dizygotic twins do not. If there was a genetic component to trust than one would expect higher correlation in the trust behavior of the former as compared to the latter.

Cesarini et al. (2008) carried out a trust game study with monozygotic and dizygotic twin populations in the USA and Sweden. They report a heritability estimate of trust of 10% in Swedish subjects and 20% in American subjects; for trustworthiness they report 17% in US

subjects and 18% in Swedish subjects. They conclude that "a significant proportion of variance in trust is due to heritability" (p. 3723).

Another twin study by Sturgis et al. (2010) investigated the genetic basis of trust. But unlike the Cesarini et al. (2008) they do not use the trust game but rather a questionnaire consisting of four questions: (a) "I believe most people are basically well-intentioned," (B) "I believe that most people will take advantage of you if you let them," (C) "I think most of the people I deal with are honest and trustworthy," and (D) "my first reaction is to trust people." They also conclude that "the majority of the variance in a multi-item trust scale is accounted for by an additive genetic factor… The environmental influences experienced in common by sibling pairs have no discernible effect" (p. 205).

In the next section we will discuss the role of oxytocin in trust behavior. There is an oxytocin receptor gene located on chromosome 3p25 (Inoue et al., 1994) and it comes in multiple variations. These variations seem to correspond to differences in empathy and stress reactivity. It was therefore hypothesized that they may have a role to play in individual differences in trust behavior. To test this hypothesis Reuter et al. (cited in (Riedl & Javor, 2012)) administered a trust game and a risk game to participants screened for variance of the gene. They reported that participants who possess a particular variant of the oxytocin receptor gene are more trusting than those who possess alternate variant of the gene. However, these participants do not differ in their risk-taking behavior and are not themselves more trustworthy, indicating that the gene variant does not result in a general increase in prosocial behavior. They concluded "individual differences in the proclivity to trust are influenced by variations in the oxytocin gene." It must also be noted that another study (Apicella et al., 2010) found no genetic differences in the trust game. However, given some limits of the experimental design they were not prepared to rule out a role for oxytocin in explaining variations in trust behaviors.

These studies are consistent with the evolutionary story in implicating a genetic component to our predisposition to trust.

3.5.2 Hormones and neurotransmitters

Nonapeptides such as oxytocin and arginine vasopressin, produced in the hypothalamus, are both hormones and neurotransmitters and are known to play an important role in prosocial

behaviors such as pair bonding and maternal nurturing in mammals like rats and prairie voles (Fleming & Rosenblatt, 1974; Insel & Young, 2001). In humans, oxytocin is known for its role in inducing labor and stimulating milk flow in nursing mothers. If these nonapeptides are enhancing prosocial behaviors, they may also have a modulatory effect on trust in humans.

The first study to test this possibility in humans with the trust game was undertaken by Zak et al. (2005). They tested oxytocin levels from participants right after game completion and found that trustees who felt trusted (i.e. received a greater transfer of funds) had on average 41% higher oxytocin levels than a control group (who received a random allocation from a mechanical source). Furthermore, these trustees returned more money back to the trustor compared to those receiving a random transfer from a mechanical source (53% vs. 18%), thus proving more trustworthy. These results suggest that when people are trusted their brains release oxytocin which in turn affects trustworthiness. These results again reinforces the distinction between social and nonsocial risk-taking behaviors. Oxytocin affects the former but not the latter. There were no oxytocin related effects on the trustor participants.

In a follow-up study Kosfeld et al. (2005) carried out the trust game with trustors and trustees randomly assigned to a placebo or oxytocin condition, where oxytocin was administered as a nasal spray. They report that oxytocin did not increase trustees' willingness to send back money. Trustors who received oxytocin transferred 17% more money to the trustee compared to the control condition, even though there was no increase in trustors' beliefs about the level of back transfer. Of 29 participants 13 (45%) exhibited maximal trust in the oxytocin group (i.e. transferred all their funds) compared to 6 in the control group. However, not all trustors in the oxytocin condition transferred funds. So, one can conclude that exogenously introduced oxytocin enhances trust. However, how it does so is less clear.

Four possibilities were considered: (a) oxytocin makes trustors more optimistic, increasing their beliefs about the level of pay back transfer; (b) oxytocin increases generosity (i.e. prosocial behavior) in trustors; (c) oxytocin increases risk seeking behavior in trustors; and (d) oxytocin reduces aversion to betrayal. There experimental design allowed them to rule out hypotheses a-c and conclude that exogenously introduced oxytocin reduces aversion to betrayal. However, other studies (Barraza & Zak, 2009) have reported a correlation between oxytocin and generosity.

Baumgartner et al. (2008) carried out a follow-up study focusing on participants' trust and risk-taking after receiving small insignificant transfers back from the trustee. These subjects continued to transfer funds. Baumgartner et al. concluded that oxytocin selectively deactivates neural circuitry in the amygdala and midbrain regions involved in fear responses.

It is also important to note that some of these studies such as Kosfeld et al. (2005) have failed replication (Declerck et al., 2020) and there are many mixed results in the literature. The appropriate conclusion here is that while the oxytocin hypothesis is tantalizing, the exact role of oxytocin in trust behavior is not yet understood.

Returning to our question about the origins of trust, the answers to these five questions suggest a picture along the following lines: Trust, along with personal benefit maximization, recognition of fairness and contingent reciprocity in oneself and projection of it onto conspecifics, ability to detect cheaters, and ability to punish cheaters form the basis of cooperation in human groups and societies. It is an innate predisposition (with individual differences) that can be modulated by social norms and rational cost-benefit analysis.

4.0 Conclusion

Expand on how this relates to Tethered Rationality….

References

Apicella, C. L., Cesarini, D., Johannesson, M., Dawes, C. T., Lichtenstein, P., Wallace, B., Beauchamp, J., & Westberg, L. (2010). No Association between Oxytocin Receptor (OXTR) Gene Polymorphisms and Experimentally Elicited Social Preferences. *PLOS ONE*, *5*(6), e11153. https://doi.org/10.1371/journal.pone.0011153

Barraza, J. A., & Zak, P. J. (2009). Empathy toward Strangers Triggers Oxytocin Release and Subsequent Generosity. *Annals of the New York Academy of Sciences*, *1167*(1), 182–189. https://doi.org/10.1111/j.1749-6632.2009.04504.x

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. https://doi.org/10.1016/j.neuron.2008.04.009

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bohnet, I., & Frey, B. S. (1999). Social Distance and Other-Regarding Behavior in Dictator Games: Comment. *American Economic Review*, *89*(1), 335–339. https://doi.org/10.1257/aer.89.1.335

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, *98*(1), 294–310. https://doi.org/10.1257/aer.98.1.294

Brosnan, S. F., Silk, J. B., Henrich, J., Mareno, M. C., Lambeth, S. P., & Schapiro, S. J. (2009). Chimpanzees (Pan troglodytes) do not develop contingent reciprocity in an experimental task. *Animal Cognition*, *12*(4), 587–597. https://doi.org/10.1007/s10071-009-0218-z

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of Sciences*, *105*(10), 3721–3726. https://doi.org/10.1073/pnas.0710069105

Cox, J. C., Friedman, D., & Sadiraj, V. (2008). Revealed Altruism. *Econometrica*, *76*(1), 31–69. https://doi.org/10.1111/j.0012-9682.2008.00817.x

de Waal, F. B. M. (1997). The Chimpanzee's service economy: Food for grooming. *Evolution and Human Behavior*, *18*(6), 375–386. https://doi.org/10.1016/S1090-5138(97)00085-8

de Waal, F. B. M., & Brosnan, S. F. (2006). Simple and complex reciprocity in primates. In P. M. Kappeler & C. P. van Schaik (Eds.), *Cooperation in Primates and Humans: Mechanisms and Evolution* (pp. 85–105). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-28277-7_5

Declerck, C. H., Boone, C., Pauwels, L., Vogt, B., & Fehr, E. (2020). A registered replication study on oxytocin and trust. *Nature Human Behaviour*, *4*(6), 646–655. https://doi.org/10.1038/s41562-020-0878-x

Eimontaite, I., Nicolle, A., Schindler, I. C., & Goel, V. (2013). The effect of partner-directed emotion in social exchange decision-making. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00469

Eimontaite, I., Schindler, I., De Marco, M., Duzzi, D., Venneri, A., & Goel, V. (2019). Left Amygdala and Putamen Activation Modulate Emotion Driven Decisions in the Iterated Prisoner's Dilemma Game. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00741

Ewing, L., Sutherland, C. A. M., & Willis, M. L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology*, *55*(8), 1694–1701. https://doi.org/10.1037/dev0000747

Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of the European Economic Association*, *7*(2–3), 235–266. https://doi.org/10.1162/JEEA.2009.7.2-3.235

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), Article 6960. https://doi.org/10.1038/nature02043

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Finuras, P. (2023). Social Identity and Trust: An Evolutionary Perspective. *Journal of Intercultural Management and Ethics*, *6*(3), 5–17. https://www.ceeol.com/search/article-detail?id=1199997

Fleming, A. S., & Rosenblatt, J. S. (1974). Maternal behavior in the virgin and lactating rat. *Journal of Comparative and Physiological Psychology*, *86*(5), 957–972. https://doi.org/10.1037/h0036414

Gächter, S., & Falk, A. (2002). Reputation and Reciprocity: Consequences for the Labour Relation. *The Scandinavian Journal of Economics*, *104*(1), 1–26. https://doi.org/10.1111/1467-9442.00269

Inoue, T., Kimura, T., Azuma, C., Inazawa, J., Takemura, M., Kikuchi, T., Kubota, Y., Ogita, K., & Saji, F. (1994). Structural organization of the human oxytocin receptor gene. *Journal of Biological Chemistry*, *269*(51), 32451–32456. https://doi.org/10.1016/S0021-9258(18)31656-9

Insel, T. R., & Young, L. J. (2001). The neurobiology of attachment. *Nature Reviews Neuroscience*, *2*(2), 129–136. https://doi.org/10.1038/35053579

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young Children Have a Specific, Highly Robust Bias to Trust Testimony. *Psychological Science*, *21*(10), 1541–1547. https://doi.org/10.1177/0956797610383438

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889. https://doi.org/10.1016/j.joep.2011.05.007

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673–676. https://doi.org/10.1038/nature03701

McCreery, H. F., & Breed, M. D. (2014). Cooperative transport in ants: A review of proximate mechanisms. *Insectes Sociaux*, *61*(2), 99–110. https://doi.org/10.1007/s00040-013-0333-3

Riedl, R., & Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics*, *5*(2), 63–91. https://doi.org/10.1037/a0026318

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J., & Schapiro, S. J. (2005). Chimpanzees are indifferent to the welfare of unrelated group members. *Nature*, *437*(7063), 1357–1359. https://doi.org/10.1038/nature04243

Stevens, J. R., & Hauser, M. D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences*, *8*(2), 60–65. https://doi.org/10.1016/j.tics.2003.12.003

Strawson, P. F. (2008). *Freedom and Resentment and Other Essays*. Routledge. https://doi.org/10.4324/9780203882566

Sturgis, P., Read, S., Hatemi, P. K., Zhu, G., Trull, T., Wright, M. J., & Martin, N. G. (2010). A Genetic Basis for Social Trust? *Political Behavior*, *32*(2), 205–230. https://doi.org/10.1007/s11109-009-9101-5

Vanderbilt, K. E., Liu, D., & Heyman, G. D. (2011). The Development of Distrust. *Child Development*, *82*(5), 1372–1380. https://doi.org/10.1111/j.1467-8624.2011.01629.x

Zak, P. J. (2008). The Neurobiology of Trust. *Scientific American*, *298*(6), 88–95. https://www.jstor.org/stable/26000645

Zak, P. J., Kurzban, R., & Matzner, W. T. (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behavior*, *48*(5), 522–527. https://doi.org/10.1016/j.yhbeh.2005.07.009