The Science of Trust

Knight Commission on Trust, Media, and American Democracy
The Aspen Institute

Prepared by:

Luke Chang, PhD

Director of the Computational Social and Affective Neuroscience Laboratory
Assistant Professor of Psychological & Brain Sciences

Dartmouth College
6207 Moore Hall
Hanover, NH 03755
luke.j.chang@dartmouth.edu

July 2017

The Science of Trust

The foundation of modern society is built upon our ability to successfully conduct cooperative social exchanges. This capability has facilitated the emergence of strategic coalitions, markets, and systems of governance. A core feature of social exchange is our capacity for trust, which has been described as the "lubricant of a social system" (Arrow, 1974) as a consequence of its ability to reduce transaction costs and increase information sharing (Dyer & Chu, 2003). Trusting that a partner will honor their agreement is critical for ensuring successful interpersonal, business, and political transactions. Trust law dates back to around the 13th century and is considered one of the most innovative contributions of the English legal system. Crusaders would leave their land to a colleague ("trustee") to protect and continue to pay feudal taxes. However, upon their return there was no formal law requiring the trustee to return the land to the crusader. The dispute would be settled by the King's Lord Chancellor, often a clergyman, who would side with the crusader arguing under the principle of "equity" that it would be unconscionable for the trustee to renege on the prior agreement. The trustee, who was the legal owner under common law, would then be compelled to return the land to the "beneficiary" crusader when requested (Avini, 1995). Consequently, countries with formal institutions that protect property and contract rights have stronger internal perceptions of trust and civic cooperation, which are associated with decreased rates of violent crime in neighborhoods (Sampson, Raudenbush, & Earls, 1997) and increased national economic growth (Knack & Keefer, 1997). Thus, trust is intimately tied to broader indicators of societal success such as crime rates, economic growth, and governance.

This paper will focus more narrowly on interactions between individuals and will review the scientific evidence supporting the psychological and neurobiological foundations for how our minds have developed the capacity to trust. This review begins with a brief discussion of how trust has been conceptualized and studied in the laboratory and then reviews the psychological and neurobiological research on trust with a particular focus on expectations and psychological value. Though this review focuses on interpersonal trust, the general conceptual framework applies more broadly to institutional trust.

What is trust?

Our mental capacity for social life has been a longstanding feature of the human mind dating back at least 5-10 million years. It has been theorized that there were likely strong selection pressures to adapt cognitive abilities to successfully navigate social exchanges. This includes cognitive abilities such as theory of mind—the ability to represent another's mental state, beliefs, or intentions; the ability to store exchange histories; and the ability to detect and remember cheaters (Cosmides & Tooby, 1992). However, the importance of social exchange extends beyond simply attaining self-interested goals, and includes fostering successful interpersonal relationships, as well as developing a positive reputation that will impact future exchanges and relationships. Relationships not only help to fulfill a basic social need to belong (Baumeister & Leary, 1995), but are also critical to our survival through their ability to assist in reproduction, protection, and resource sharing, as well as facilitate positive physical and mental health outcomes (Uchino, Cacioppo, & Kiecolt-Glaser, 1996).

Though trust has been studied for a number of decades in a variety of disciplines, it has been difficult to agree on a comprehensive definition. Early work in psychology examined trust from the perspective of an individual and characterized trust as the general belief about the degree to which other people or groups are likely to be reliable, cooperative, or helpful in situations (Deutsch, 1973; Rotter, 1971). This work importantly established two key aspects of trust—expectations about another's behavior, and perceptions of another's benevolent motivations. However, this work was limited in that it was mostly focused on identifying individual differences in general *perceptions* of trustworthiness rather than studying *behavioral* displays of trust. Future work extended the study of trust to interpersonal interactions. Rather than being a general belief about a partner's reliability, trust required two partners and an action—*I* trust *you* to do *X* (Holmes & Rempel, 1989; Rempel, Holmes, & Zanna, 1985; Simpson, 2007). This importantly allowed trust to vary across dyadic interactions rather than just subjective perceptions. Other theorists have additionally emphasized the importance of the trustor's willingness to be vulnerable (Rousseau, Sitkin, Burt, & Camerer, 1998; Scanzoni, 1979) and the trustee's ability to overcome self-interested temptations. In this paper, we will define trust as *the psychological state of assuming mutual risk with a relationship partner to attain an interdependent goal in the face of competing temptations*.

Trust is a dynamic state and evolves over the course of a relationship. Early stages of a relationship are focused on assessing a partner's trustworthiness level. Trustors must be willing to display vulnerability and endure a risk. Initial motivations to trust might include altruism (Andreoni & Miller, 2002), efficiency gains to the both parties (Charness & Rabin, 2002), concerns about disparities in outcomes (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999), or simply to gain additional information to assess their partner's trustworthiness (Bohnet & Zeckhauser, 2004). The trustee can demonstrate their level of trustworthiness by their willingness to overcome their own self-interest and take an action that fulfills an interdependent goal. As the relationship progresses, both parties become more confident in their ability to predict the other's behavior and develop a sense of security in the relationship. The relationship might be periodically tested in order to provide a continual assessment of trust levels (Simpson, 2007). At some point in the relationship, one person might end up betraying their partner, which eventually will lead to a dissolution of the relationship (King-Casas et al., 2008).

How is trust studied?

Research investigating trust has primarily utilized two different types of experimental methodologies. The oldest method relies on subjective self-report and is primarily used to assess individual differences in dispositions to trust via surveys such as the Interpersonal Trust Scale (Rotter, 1967) or the dyadic trust scale (Larzelere & Huston, 1980). The more popular approach uses economic games to study trust behavior in the context of a simple social interaction.

Economic games are mathematical descriptions of a social interaction that includes a set of actions available to multiple players and the payoffs resulting from those actions. Games specify the sequence of play, the information available to each player, and the actions available to each player at each stage of the game. Games can be repeated or single-shot. Repeated games are more akin to real life social interactions, in which individuals might have multiple repeated interactions. However, this can introduce additional psychological factors such as learning and reputation, which can be difficult to separate from trust. In economics, single-shot games have been thought to provide a more pure model of trust as they remove any long-term strategic thinking resulting from repeated interactions. To further control for additional social factors, these games are traditionally played double-blind, i.e., anonymously. Because this tradition of anonymous single-shot games has a consequence of removing many of the social motivations that might contribute to trust, it provides a model of “pure” trust behavior and a benchmark for which other social factors can be introduced and compared (Camerer, 2003).

Games can also take a simultaneous or sequential form. Simultaneous games require each player to make their decisions at the same time. For example, imagine a team of two criminals that are arrested, imprisoned, and interrogated independently with no means of communication between them. The prosecutor hopes to get both prisoners sentenced, but lacks sufficient evidence. Thus, they offer each prisoner an opportunity for a lesser sentence if they are willing to betray their partner. If both prisoners choose to trust each other and cooperate with each other, they will both serve a short sentence. If one chooses to defect and betray their partner’s trust, then they will receive a lesser sentence and their partner will be forced to endure the longest sentence. However, if both prisoners choose to defect, then they will both receive a longer sentence than had they both cooperated. This vignette describes the classic Prisoner’s Dilemma Game, which has a payoff structure of DC > CC > DD > CD where C is cooperate and D is defect (Figure 1A). A rational and self-interested player should choose the outcome associated with the highest overall payoff, thus one potential solution to the game is for both players to defect. However, in real life players tend to converge on mutual cooperation indicating they are frequently willing to trust a relationship partner even when they don’t know who they are and will never see them again.

Sequential games, in contrast, allow players to take turns during which they make a decision conditional on the actions of the other player. The investment game, commonly referred to as the Trust Game, was designed to model pure trust similar to the crusader land trust example described above (Berg, Dickhaut, & McCabe, 1995).

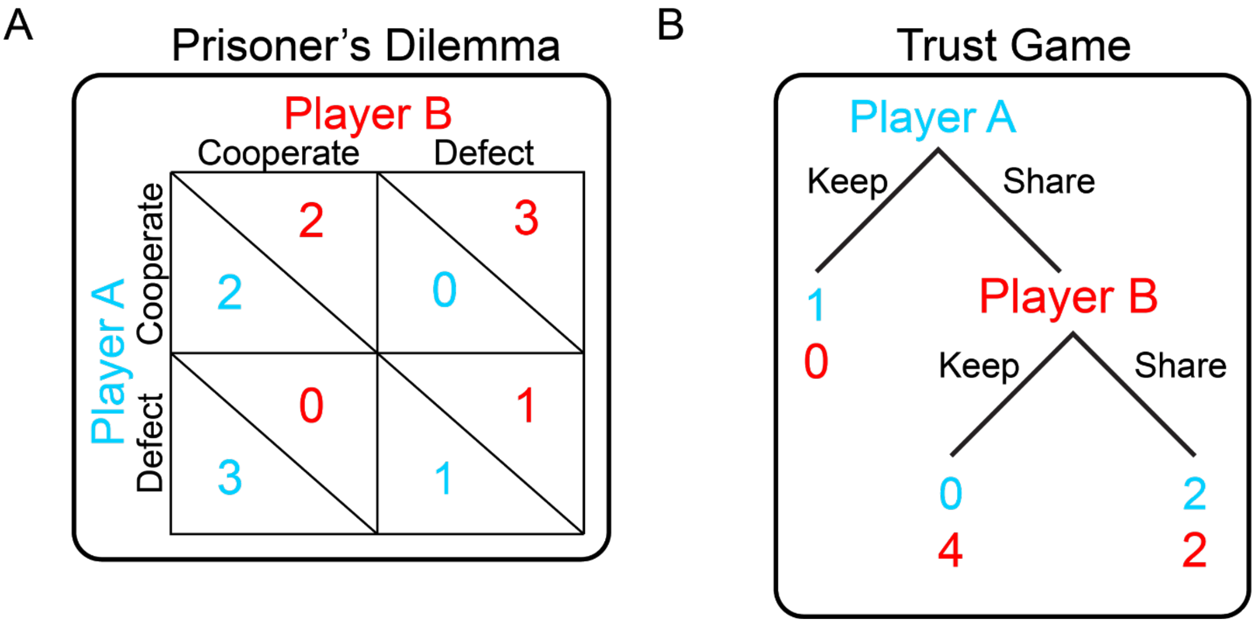

Figure 1. Example payoffs for decisions in two games involving trust. Player A's Payoffs are in Blue and Player B's Payoffs are in Red. Panel A depicts a classic Prisoner's Dilemma game. Panel B depicts a subgame of a classic investment game to emphasize the actions taken by each player in the game. The investment amount is multiplied by a factor of 4.

In this game, Player A is endowed with a sum of money (e.g., $1) and can choose to invest any amount of this endowment in their partner. The investment amount is multiplied by a factor predetermined by the experimenter (typically 3 or 4), and Player B then decides how much, if any, of this multiplied investment amount to return. A simplified version of the game is depicted in Figure 1B indicating two choices for each player to share or keep. This game is a particularly clean example of trust because only Player A endures any risk. By choosing to invest, Player A demonstrates a willingness to be vulnerable and expose their assets. Player B ultimately decides whether to honor or betray Player A's trust. This aspect of the game models Player B's willingness to forgo self-interested motivations by returning some amount of the multiplied investment back to Player A. Another interesting aspect of this game is that Player A can implicitly signal how much they expect Player B to return by the size of their investment (McCabe, Smith, & LePore, 2000). Investing all of the endowment (e.g., $10) sends a strong signal that Player A expects Player B to return about half of the multiplied endowment back, whereas investing a small portion of the endowment (e.g. $2) sends a signal that Player A does not expect Player B to return any money.

Why do players trust?

Most research on trust has used variants of the Prisoner's Dilemma and the Trust Game. An interesting question of course is *why would Player A choose to invest?* From a decision theoretic perspective, players should make decisions that maximize their expected gains minimize anticipated losses (von Neumann & Morgenstern, 2007). This allows for at least two possible theoretical explanations why Player A might trust Player B. The first is because Player A *expects* Player B to share (Chang, Doll, van't Wout, Frank, & Sanfey, 2010; Fareri, Chang, & Delgado, 2012). Using the example in Figure 1B, the expected value of choosing to invest can be formulated as the probability of sharing multiplied by the value of sharing plus the probability of not sharing multiplied by the value of not sharing, or $\Phi * 2 + (1-\Phi) * 0$, where Φ is Player A's belief about the probability of Player B choosing to share. Player A should invest if they have strong expectations of reciprocation and believe that the probability of Player B choosing share is greater than 50% (i.e., $\Phi > .5$). A second potential explanation is that Player A might receive some sort of additional *psychological value* if Player B chooses to share. Player A's psychological value might reflect an "other-regarding preference," i.e., their concern for Player B's payoff, or alternatively, the value they might receive from mutual cooperation. Here, the expected value can be formulated as $\Phi * (2 + \varphi) + (1-\Phi) * 0,$ where φ is the psychological value Player A receives from Player B choosing share. Thus, Player A should invest if $\Phi \geq .5$ and $\varphi > 0$ (Fareri, Chang, & Delgado, 2015). Both of these accounts of trust behavior are purely theoretical, which means this behavior can be predicted given a set of assumptions without running any experiments.

Developing expectations about trustworthiness

This theoretical analysis of the Trust Game indicates that expectations and psychological value are the primary variables that will impact trust. The next section of the paper examines how players develop beliefs about a relationship partner's trustworthiness and focuses on three primary mechanisms through which a player can learn about a partner's level of trustworthiness. First, they can learn by directly interacting with the partner and discovering the probability that the player will choose to reciprocate. This process involves reinforcement learning, the process by which a belief is incrementally updated after each interaction. Second, players might use some prior information as an indicator of trust. For example, do they look like they are a good person? Do they seem trustworthy? These types of judgments can be made very quickly and often outside conscious awareness. Third, players might learn about a partner's trustworthiness secondhand via conversing with another person who may have had direct experience interacting with the partner. This type of information is often conveyed through gossip.

Learning trustworthiness through reinforcement learning

Trust reflects our belief about the likelihood of a partner taking a specific action (Rotter, 1954, 1980). This generalized expectation enables us to infer whether another individual can be relied upon (Rotter, 1980). This inference enables us to predict their future behavior, thereby reducing our uncertainty about a potential relationship partner (Rempel et al., 1985). Most often, utilizing information about an individual's previous reciprocity is the best predictor of their future trustworthiness (King-Casas et al., 2005). In this way, trustworthiness reflects a dynamic belief about the likelihood of a relationship partner reciprocating (Chang et al., 2010).

The process of forming and updating these beliefs appears to recruit the basic learning architecture of the brain. According to reinforcement learning theory, we start with some expectations about the likelihood of a cue predicting a specific outcome (Rescorla, 1972). When an observed outcome is greater than we predicted, it produces a positive prediction error signal incrementally increasing our expectation for the subsequent learning event. In contrast, when an observed outcome is less than we predicted, a negative prediction error signal is produced incrementally decreasing our future predictions. This prediction error signal is the central mechanism through which any control theory system adapts (e.g., thermostat, cruise control, elevators, etc).

A seminal finding in computational neuroscience is that dopamine neurons located in the ventral tegmental area (VTA) of the brain in non-human primates, appear to fire in response to reward prediction errors (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). The frequency of firing increases when a larger reward is received than expected (positive prediction error signal), and decreases when a smaller reward is received than expected (negative prediction error signal). Importantly, as an animal learns how much reward to expect, these same neurons start to *predict* the amount of expected reward by firing in response to the reward cue that precedes the actual reward outcome. Outcomes that are correctly predicted result in no firing of dopamine neurons upon receipt of the reward. In humans, functional magnetic resonance imaging (fMRI) studies have found that prediction error signals correlate with activity in the nucleus accumbens (NAcc) a region in the ventral portion of the striatum that receives direct projections from the VTA. The anatomical locations of each of these regions can be seen in Figure 2.

Learning to trust a relationship partner in a repeated trust game appears to leverage this dopamine mediated learning system. Beliefs about the likelihood of a partner reciprocating are updated after receiving feedback about a partner's actions via prediction error driven reinforcement learning (Chang et al., 2010). A reinforcement learning model that utilizes prediction error can accurately predict subjective self-reported expectations after multiple interactions. In addition, NAcc activation at the time of outcome has been shown to correlate directly with model derived prediction error learning signals (Fareri et al., 2012; Rilling et al., 2002) and appears to propagate backwards to the earliest predictor of trust after repeated reciprocation (King-Casas et al., 2005; Kishida & Montague, 2012).

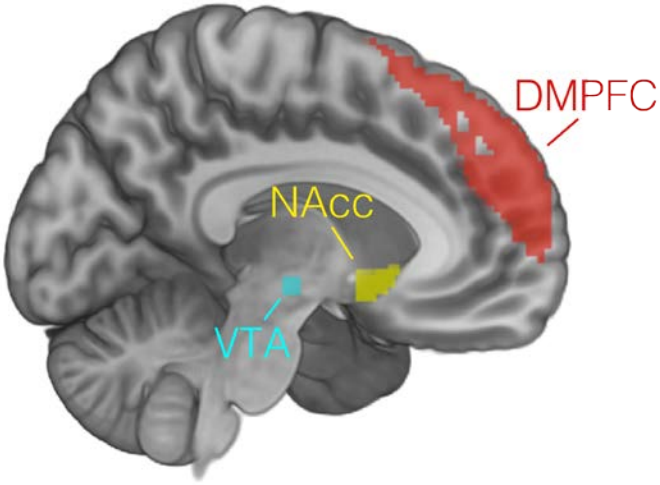

Figure 2. Neural basis of trust. Areas of the brain frequently associated with trust. VTA: ventral tegmental area, NAcc: nucleus accumbens also referred to as ventral striatum, DMPFC: dorsomedial prefrontal cortex.

Contextual information can impact trust learning

Both beliefs about trustworthiness and the process of learning whether to trust can be influenced by information outside of the specific trust context. First, external information can bias prior beliefs about an individual's trustworthiness. For example, facial expressions are processed very quickly (100-200 milliseconds) (Pizzagalli et al., 2002; Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006) and often outside conscious awareness (Winston, Strange, O'Doherty, & Dolan, 2002) convey rich social information. Individuals who are attractive or who appear happy are more likely to be viewed as more trustworthy (Scharlemann, Eckel, Kacelnik, & Wilson, 2001). These initial trustworthiness judgments can predict the amount of financial risk a person is willing to take in a Trust Game (van't Wout & Sanfey, 2008). Another source of prior information is an individual's group membership. For example, prior information about a partner's race can impact perceptions of trustworthiness and also the amount of trust one is willing to endure in a trust game. Participants with strong automatic implicit attitudes that white is associated with good and black is associated with bad invest substantially more money in Caucasian partners compared to African-American partners in a Trust Game (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011) irrespective of the participant's explicit racial attitudes. Beyond simply viewing a face, brief social interactions can also convey information about an individual's trustworthiness. When Player B is allowed to send a short message to Player A prior to making their investment decision, Player A is much more likely to trust Player B and Player B is more likely to reciprocate trust in return (Charness & Dufwenberg, 2006). Similarly, the ability to have a 30-minute social interaction prior to a repeated Prisoner's Dilemma Game can increase a participant's ability to accurately predict their partner's decision and thus the overall amount of cooperation (Frank, Gilovich, & Regan, 1993).

Second, prior information can change the way in which people learn and update beliefs about a relationship partner's trustworthiness. When given informational vignettes about a relationship partner's moral character, participants are more likely to invest money in partners that have previous evidence of "good" moral character compared to "bad" moral character (Delgado, Frank, & Phelps, 2005). Interestingly, this information can change how trustworthiness beliefs are updated after repeated interactions such that participants fail to update their beliefs that a "good" person is untrustworthy when the partner does not reciprocate trust. This finding is consistent with the notion of confirmation bias. Confirmation bias describes the psychological phenomenon of selectively shaping or identifying evidence that is consistent with a prior hypothesis and discounting information that is inconsistent with a prior belief (Nickerson, 1998). For example, imagine there is a political candidate that you like and support. According to confirmation bias, you will be less critical of any evidence that supports your beliefs about the candidate and might even attempt selectively seek out evidence that specifically supports your views (Jonas, Schulz-Hardt, Frey, & Thelen, 2001). In addition, you will likely be more critical of any evidence that counters your beliefs and may choose to avoid situations where your beliefs will be challenged, for example by clicking on a link to a story that might contain negative information about your favored candidate (Bakshy, Messing, & Adamic, 2015). This is a very problematic reasoning bias as it often results in self-fulfilling prophecies. Confirmation bias also appears to impact decisions to trust. In one study, participants interacted with each other in a virtual ball tossing task known as "cyberball" (Fareri et al., 2012). One of the partners frequently included the participant in the ball passing game, and the another frequently excluded the participant. In a subsequent trust game, the researchers found that participants used this prior information in their trust decisions even though it was in a completely new type of context (i.e., trust game vs cyberball). Interestingly, this information also impacted *how* participants learned to trust their partners. Participants were more likely to update their beliefs that their partner was trustworthy when they reciprocated if the partner frequently shared in the cyberball task. Alternatively, participants were more likely to update their beliefs that the partner was untrustworthy when they did not reciprocate trust if they also did not share in the cyberball game. A subsequent study found evidence that prior information about a partner in a repeated trust game was stored in the dorsomedial prefrontal cortex (DMPFC) and appeared to override reinforcement-based learning signals in the striatum consistent with the notion that they were using a model of the player to interpret behavior rather than trial-to-trial feedback (Fouragnan et al., 2013). This region has previously been associated with mentalizing operations such as representing another person's psychological state (Amodio & Frith, 2006).

Gossip can spread reputational information

The third way a player might learn about a partner's trustworthiness is from a third party via gossip. An agent's reputation is a belief about their overall dispositional trustworthiness shared among multiple other agents. Building a reliable and accurate belief about an individual's disposition requires many observations across multiple contexts. As discussed above, behavior in other contexts can lead to building a reputation, which can impact subsequent interactions in different contexts (Milinski, Semmann, & Krambeck, 2002). This information can be observed by many different agents and transmitted to each via communication in the form of gossip. Gossip can be defined as private evaluative comments about an absent third party. It provides a mechanism to vicariously learn about another agent's actions and reputation from secondhand information (Jolly & Chang, Under Review), though it does not appear to supplant direct first-hand experience (Sommerfeld, Krambeck, Semmann, & Milinski, 2007). Gossip is more likely to impact trust decisions, when multiple gossip sources are aggregated and the information is consistent (Sommerfeld, Krambeck, & Milinski, 2008). In laboratory experiments involving multiple agents playing trust games or the multi-person variant of the game known as the Public Goods Game, gossip has been shown to reflect accurate descriptions of the other player’s behavior (Jolly & Chang, Under Review; Sommerfeld et al., 2008, 2007) and increase overall levels of cooperation in the game (Feinberg, Willer, & Schultz, 2014; Jolly & Chang, Under Review; Sommerfeld et al., 2008; Wu, Balliet, & Van Lange, 2016). The sharing of gossip can take the form of a trust game in itself. Revealing one’s source for private gossip is generally viewed as a betrayal of trust. Cultivating trusting relationships will ultimately lead to greater access to reputational information from a broader social network.

Psychological value of trust

In addition to expectations, psychological value is another mechanism that can impact decisions to trust in the Trust Game. The next section of the paper examines different factors that can impact the psychological value of trust. First, mutual cooperation can be rewarding. Second, betrayal of trust can result in negative psychological value and make it less likely to trust even if there is an expectation of reciprocation. Third, we are often able to infer a relationship partner’s motivations and receive value from reciprocating a relationship partner’s good intentions.

Reciprocation of Trust is Rewarding

In general, we value fairness and appreciate when our relationship partners reciprocate our trust. Multiple studies have reported evidence of a reward signal in the ventral striatum when a relationship partner exhibits fairness (Tabibnia, Satpute, & Lieberman, 2008) and chooses to honor an investment in the trust game (Fareri et al., 2015; Phan, Sripada, Angstadt, & McCabe, 2010). Dopamine is complicated as it is not only involved in prediction-error driven learning as previously described, but also encodes the desire (Berridge & Robinson, 1998) and expectation of both reward (Fiorillo, Tobler, & Schultz, 2003) and hedonic pleasure (Sharot, Shiner, Brown, Fan, & Dolan, 2009). Moreover, the dopamine system is also intimately connected to the opioid system, which is involved in processing the feeling of ‘liking’ associated with hedonic rewards (Berridge & Robinson, 1998; Leknes & Tracey, 2008). Interestingly, strong expectations of reciprocity in the trust game are associated with increased activation in the ventral striatum when learning that a relationship partner reciprocated. This is the opposite prediction of the learning account of trust described above, as predictions are matched rather than violated. One recent study attempted to integrate these two seemingly competing accounts of trust using a computational model and a repeated trust game where participants played with a close friend or a stranger (Fareri et al., 2015). This study found evidence supporting both accounts—that neural activity in the ventral striatum was involved in both learning to predict a partner's trustworthiness and also the feeling of reward when a relationship partner reciprocated. The degree of reward was stronger the closer the participants described they were to their friend. This suggests that the process of entrusting and fulfilling obligations is a rewarding endeavor and critical in building and maintaining social relationships.

Betrayal of trust damages relationships

In contrast, we dislike having our trust betrayed and feeling like we have been taken advantage of by a relationship partner. This aversion to betrayal can make it difficult to trust and appears to be a negative value signal beyond simple risk (Bohnet & Zeckhauser, 2004). Thus, betrayal aversion serves as a negative psychological value, which decreases the likelihood of choosing to trust. In general, once there has been a breach in a relationship, trust will quickly decay. In a repeated Trust Game, following a ruptured alliance, skilled Player Bs will attempt to "coax" Player A to resume their trust by returning a larger proportion a low investment than they normally would (King-Casas et al., 2008). This action can effectively repair a relationship and typically results in increased trust in future rounds. Patients with Borderline Personality Disorder with significant impairment in their ability to maintain healthy interpersonal relationships do not exhibit such a behavior and typically result in a dissolution of trust by the end of the game (King-Casas et al., 2008). In addition, dissolution of trust decays proportional to the degree to which an expectation of trust was violated. In one repeated trust game, participants who were perceived to be the most trustworthy by an implicit facial judgement, but acted untrustworthy by only reciprocating 20% of the time were given the least amount of money at the end of the game (Chang et al., 2010). Interestingly, these types of partners were given less money than a different partner that also only reciprocated 20% of the time but looked untrustworthy. This provides additional support that betrayal can result in a negative psychological value which ultimately subtracts from a financial payoff received from choosing to trust.

Benevolent Intentions

As an agent learns about the types of actions that their partner will take in a variety of contexts, they will eventually begin to develop a model of their partner's motivational tendencies. For example, over many repeated interactions Player A might learn that Player B will consistently keep any money invested in Player B and conclude that Player B is only motivated by their own self-interests. Alternatively, Player B might only reciprocate when other people are watching and keep the money in more private contexts. This suggests that Player B might be self-interested, but is also motivated to maintain a positive reputation. As the relationship develops after many interactions, a trustworthy Player B might begin to develop a motivation to maintain the relationship and will be willing to protect Player A's interests even at the expense of Player B's own financial outcome. These types of actions will dramatically increase trust and are essential for a positive relationship that might ultimately be mutually beneficial over a longer amount of time (King-Casas et al., 2008).

Several studies have sought to demonstrate evidence of this behavior. 'Tit-for-tat' is a classic strategy that capitalizes on this motivation and has been shown to be a highly effective algorithm in repeated prisoner's dilemma games. This strategy of conditional cooperation will reciprocate whatever action their partner took on the previous 1-2 trials. If Player A cooperates, then Player B will cooperate. If Player A chooses to defect, then Player B will defect. This strategy has been shown to be highly stable and successful, beating out several competing computer algorithms in a large tournament (Axelrod, 2006). In more complicated contexts, participants seem to track their partner's intentions and will reciprocate good intentions and punish bad intentions (Dufwenberg & Kirchsteiger, 2004; Rabin, 1993). Interestingly, this behavior does not appear to emerge developmentally until late adolescence (Sutter & Kocher, 2007) and its emergence parallels developmental changes in cortical thickness in the dorsomedial prefrontal cortex (DMPFC), a region known to be associated with mentalizing or representing another's state of mind (Sul, Guroglu, Crone, & Chang, In Press). This suggests that inferring a relationship partner's intentions requires a psychological process distinct from simply learning about their behavior in each context.

The dubious link between oxytocin and trust

The final section of this review briefly discusses some of the research investigating the biological foundation of trust. The bulk of this work has focused on hormones such as oxytocin and several of these studies have received considerable attention from popular media that has resulted in a broad popular consensus that oxytocin is a "moral molecule." However, this interpretation may be premature given the extant scientific evidence and should be interpreted with caution.

Oxytocin has received considerable attention as a potential biological mechanism for trust. It is a neuropeptide that is produced in the hypothalamus and released into the bloodstream at the pituitary gland and associated with many different types of species-specific social behavior (Donaldson & Young, 2008; Insel, 2010). For example, oxytocin is released during childbirth and signals several maternal behaviors such as lactation and maternal bonding. Oxytocin has also been demonstrated to be instrumental in monogamous pair-bonding in prairie voles (Young & Wang, 2004). A preference for a specific partner can be manipulated in female prairie voles by administering oxytocin and inhibited by blocking an oxytocin receptor. In humans, there has been intense interest in investigating the role of oxytocin in social behavior such as trust, empathy, and even as a treatment for social disorders such as autism and social anxiety, though the results from early randomized clinical trials have been disappointing (Anagnostou et al., 2012; Guastella, Howard, Dadds, Mitchell, & Carson, 2009). However, this work has been challenging as it is currently not possible to reliably validate oxytocin administration techniques and determine how well they can deliver oxytocin to the brain (Christensen, Shiyanov, Estepp, & Schlager, 2014; Kagerbauer et al., 2013; Striepens et al., 2013).

One highly-cited study found that intranasal oxytocin administration significantly increased the amount of money that male players invested in the trust game compared to placebo, but did not impact the amount of money returned by the trustee or players' behavior in a nonsocial version of the task (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). This suggests that oxytocin appears to increase trust, but not general preferences for altruism or risk.

While a causal effect of oxytocin on trust is intriguing, it is important for this finding to be replicated by multiple research groups before it can be considered scientific evidence. Unfortunately, after six attempts at replication using slightly different experimental designs, the results have largely been inconclusive (Barraza, McCullough, Ahmadi, & Zak, 2011; Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008; Ebert et al., 2013; Klackl, Pfundmair, Agroskin, & Jonas, 2013; Mikolajczak et al., 2010; Yao et al., 2014). One technique for quantifying the overall evidence for oxytocin's effect on trust is to aggregate the effect size across all of these studies using a statistical meta-analysis. This analysis pooled the results from 481 participants and yielded an effect size of 0.077, which is very small and not statistically significant (Nave, Camerer, & McCullough, 2015). Thus, considering the methodological uncertainty of whether oxytocin administration can actually enter the brain along with the null treatment effect in the meta-analysis, there does not appear to be a robust link between oxytocin and trust at this time, but future research is certainly warranted.

Summary

This paper provides a broad overview of the scientific research investigating trust published in the fields of psychology, economics, neuroscience, and biology. To summarize, trust can be defined as the psychological state of assuming mutual risk with a relationship partner to attain an interdependent goal in the face of competing temptations. Trust behavior can be studied in the context of interpersonal transactions using economic games designed as "pure" measurements of trust such as the Trust Game and Prisoner's Dilemma. In these contexts, the two main mechanisms impacting trust are expectations and psychological value. The brain systems supporting establishing trustworthiness extend core neural circuits responsible for reward-based learning via prediction error. More generally, learning to trust can be influenced by other social cues such facial expressions, as well as prior beliefs and reputational information. Finally, the psychological value of trust is a key component of why we trust at all and why trust can be so fragile: because of the positive rewards that it brings, and the pain betrayals impart.

While this paper has focused on trust from an interpersonal perspective, these ideas extend to broader institutions such as news, business, and government. There are three different types of relationships that can impact trust in institutions: the product, the point of contact, and the organization. For example, we might decide to trust an individual piece of news based on (a) the apparent veracity of the informational content, (b) the specific person delivering the news (e.g., a colleague, writer, anchorperson), or (c) where the news came from (e.g., a specific news station, media conglomerate, or government) (Williams, 2012). Our ability to trust the person delivering the news and in the broader organization are based on all of the same concepts discussed in the context of interpersonal trust. What is the overall reputation of the person or organization? What is the context? What are the motivations for delivering the news and what type of risk is being endured to deliver it? We are likely to trust people and organizations if we feel like they are reliable, and care about our interests. Organizations that are solely concerned with maximizing their own profit and self-interests will likely be perceived as less trustworthy than organizations that prioritize consumer satisfaction. In general, news organizations that have a specific agenda and are motivated to portray news that furthers their agenda will likely be viewed as less trustworthy than organizations that are solely motivated to present news accurately and free of bias. However, specific individuals who share the same agenda might develop a stronger perception of trust with this particular organization. This can be problematic when a trustworthy source provides inaccurate information, as the individual will be less critical and more likely to believe the information. Finally, our trust in a representative from the organization (news anchor, salesperson, customer service, CEO, etc.) can also strongly impact our trust in the broader organization. These individuals represent the broader interests, motivations and reputation of the company.

Acknowledgements

The author would like to acknowledge Eunice Lee, Eshin Jolly, and Andy Chen for their helpful comments on previous drafts of this manuscript.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews. Neuroscience*, *7*(4), 268–277.

Anagnostou, E., Soorya, L., Chaplin, W., Bartz, J., Halpern, D., Wasserman, S., … Hollander, E. (2012). Intranasal oxytocin versus placebo in the treatment of adults with autism spectrum disorders: a randomized controlled trial. *Molecular Autism*, *3*(1), 16.

Andreoni, J., & Miller, J. (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica: Journal of the Econometric Society*, *70*(2), 737–753.

Arrow, K. (1974). *The Limits of Organization*. citeulike.org.

Avini, A. (1995). Origins of the Modern English Trust Revisited Comments. *Tulane Law Review*, *70*, 1139–1164.

Axelrod, R. M. (2006). *The Evolution of Cooperation: Revised Edition*. Basic Books.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Barraza, J. A., McCullough, M. E., Ahmadi, S., & Zak, P. J. (2011). Oxytocin infusion increases charitable donations regardless of monetary resources. *Hormones and Behavior*, *60*(2), 148–151.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529.

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*(1), 122–142.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research. Brain Research Reviews*, *28*(3), 309–369.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, *55*(4), 467–484.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review*, *90*(1), 166–193.

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

Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, *61*(2), 87–105.

Charness, G., & Dufwenberg, M. (2006). Promises and Partnership. *Econometrica: Journal of the Econometric Society*, *74*(6), 1579–1601.

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Christensen, J. C., Shiyanov, P. A., Estepp, J. R., & Schlager, J. J. (2014). Lack of association between human plasma oxytocin and interpersonal trust in a Prisoner's Dilemma paradigm. *PloS One*, *9*(12), e116172.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, *163*, 163–228.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618.

Deutsch, M. (1973). The resolution of conflict. *New Haven, CT: Yale*.

Donaldson, Z. R., & Young, L. J. (2008). Oxytocin, vasopressin, and the neurogenetics of sociality. *Science*, *322*(5903), 900–904.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Dyer, J. H., & Chu, W. (2003). The Role of Trustworthiness in Reducing Transaction Costs and Improving Performance: *Organization Science*, *14*(1), 57–68.

Ebert, A., Kolb, M., Heller, J., Edel, M.-A., Roser, P., & Brüne, M. (2013). Modulation of interpersonal trust in borderline personality disorder by intranasal oxytocin and childhood trauma. *Social Neuroscience*, *8*(4), 305–313.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, *6*, 148.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(21), 8170–8180.

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science*, *25*(3), 656–664.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(8), 3602–3611.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, *14*(4), 247–256.

Guastella, A. J., Howard, A. L., Dadds, M. R., Mitchell, P., & Carson, D. S. (2009). A randomized controlled trial of intranasal oxytocin as an adjunct to exposure therapy for social anxiety disorder. *Psychoneuroendocrinology*, *34*(6), 917–923.

Holmes, J. G., & Rempel, J. K. (1989). *Trust in close relationships*. Sage Publications, Inc.

Insel, T. R. (2010). The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior. *Neuron*, *65*(6), 768–779.

Jolly, E., & Chang, L. (Under Review). Gossip drives vicarious learning and facilitates robust social connections.

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, *80*(4), 557–571.

Kagerbauer, S. M., Martin, J., Schuster, T., Blobner, M., Kochs, E. F., & Landgraf, R. (2013). Plasma oxytocin and vasopressin do not predict neuropeptide concentrations in human cerebrospinal fluid. *Journal of Neuroendocrinology*, *25*(7), 668–673.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, *321*(5890), 806–810.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83.

Kishida, K. T., & Montague, P. R. (2012). Imaging models of valuation during social interaction in humans. *Biological Psychiatry*, *72*(2), 93–100.

Klackl, J., Pfundmair, M., Agroskin, D., & Jonas, E. (2013). Who is to blame? Oxytocin promotes nonpersonalistic attributions in response to a trust betrayal. *Biological Psychology*, *92*(2), 387–394.

Knack, S., & Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics*, *112*(4), 1251–1288.

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

Larzelere, R. E., & Huston, T. L. (1980). The Dyadic Trust Scale: Toward Understanding Interpersonal Trust in Close Relationships. *Journal of Marriage and Family Counseling*, *42*(3), 595–604.

Leknes, S., & Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews. Neuroscience*, *9*(4), 314–320.

McCabe, K. A., Smith, V. L., & LePore, M. (2000). Intentionality detection and “mindreading”: Why does game form matter? *Proceedings of the National Academy of Sciences*, *97*(8), 4404–4409.

Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., & Luminet, O. (2010). Oxytocin makes people trusting, not gullible. *Psychological Science*, *21*(8), 1072–1074.

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the “tragedy of the commons.” *Nature*, *415*(6870), 424–426.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(5), 1936–1947.

Nave, G., Camerer, C., & McCullough, M. (2015). Does Oxytocin Increase Trust in Humans? A

Critical Review of Research. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *10*(6), 772–789.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, *2*(2), 175.

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences*, *107*(29), 13099–13104.

Pizzagalli, D. A., Lehmann, D., Hendrick, A. M., Regard, M., Pascual-Marqui, R. D., & Davidson, R. J. (2002). Affective judgments of faces modulate early activity (approximately 160 ms) within the fusiform gyri. *NeuroImage*, *16*(3 Pt 1), 663–677.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, *83*(5), 1281–1302.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality*. Retrieved from http://psycnet.apa.org/journals/psp/49/1/95/

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical Conditioning II: Current Research and Theory*. Retrieved from http://ci.nii.ac.jp/naid/10018337711/

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

Rotter, J. (1954). Social learning and clinical psychology. https://doi.org/10.1037/10788-000

Rotter, J. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665.

Rotter, J. (1971). Generalized expectancies for interpersonal trust. *The American Psychologist*, *26*(5), 443.

Rotter, J. (1980). Interpersonal trust, trustworthiness, and gullibility. *The American Psychologist*, *35*(1), 1.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review. Academy of Management*, *23*(3), 393–404.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science*, *277*(5328), 918–924.

Scanzoni, J. (1979). Social exchange and behavioral interdependence. *Social Exchange in Developing Relationships*, 61–98.

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal Of Economic Psychology*, *22*(5), 617–640.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Sharot, T., Shiner, T., Brown, A. C., Fan, J., & Dolan, R. J. (2009). Dopamine enhances expectation of pleasure in humans. *Current Biology: CB*, *19*(24), 2077–2080.

Simpson, J. A. (2007). Psychological Foundations of Trust. *Current Directions in Psychological Science*, *16*(5), 264–268.

Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings. Biological Sciences / The Royal Society*, *275*(1650), 2529–2536.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(44), 17435–17440.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7710–7715.

Striepens, N., Kendrick, K. M., Hanking, V., Landgraf, R., Wüllner, U., Maier, W., & Hurlemann, R. (2013). Elevated cerebrospinal fluid and blood concentrations of oxytocin following its intranasal administration in humans. *Scientific Reports*, *3*, 3440.

Sul, S., Guroglu, B., Crone, E., & Chang, L. (In Press). Medial prefrontal cortical thinning mediates shifts in other-regarding preferences during adolescence. *Scientific Reports*.

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior*, *59*(2), 364–382.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The Sunny Side of Fairness: Preference for Fairness Activates Reward Circuitry (and Disregarding Unfairness Activates Self-Control Circuitry). *Psychological Science*, *19*(4), 339–347.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Social Cognition*, *27*(6), 813–833.

Uchino, B. N., Cacioppo, J. T., & Kiecolt-Glaser, J. K. (1996). The relationship between social support and physiological processes: a review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin*, *119*(3), 488–531.

van ’t Wout, M., & Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803.

von Neumann, J., & Morgenstern, O. (2007). *Theory of Games and Economic Behavior*. Princeton University Press.

Williams, A. E. (2012). Trust or Bust?: Questioning the Relationship Between Media Trust and News Attention. *Journal of Broadcasting & Electronic Media*, *56*(1), 116–131.

Willis, J., & Todorov, A. (2006). First Impressions. *Psychological Science*, *17*(7), 592–598.

Winston, J. S., Strange, B. A., O’Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*(3), 277–283.

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016). Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. *Scientific Reports*, *6*, 23919.

Yao, S., Zhao, W., Cheng, R., Geng, Y., Luo, L., & Kendrick, K. M. (2014). Oxytocin makes females, but not males, less forgiving following betrayal of trust. *The International Journal of Neuropsychopharmacology / Official Scientific Journal of the Collegium Internationale Neuropsychopharmacologicum* , *17*(11), 1785–1792.

Young, L. J., & Wang, Z. (2004). The neurobiology of pair bonding. *Nature Neuroscience*, *7*(10), 1048–1054.

Young, L. J., & Wang, Z. (2004). The neurobiology of pair bonding. *Nature Neuroscience*, *7*(10), 1048–1054.