CHAPTER 8

Trust and Learning
Neurocomputational Signatures of Learning to Trust

Gabriele Bellucci and Jean-Claude Dreher

8.1 Introduction

Learning whom to trust or distrust is an important skill in building social relationships. This chapter focuses on learning dynamics underlying decisions whether to trust or distrust other partners in social interactions. Two main experimental paradigms have been employed to investigate how people learn whom to trust. One experimental setting refers to advice-taking paradigms (Yaniv & Kleinberger, 2000). In this type of paradigm, two partners interact with each other as adviser and advisee, respectively. Participants take in general the role of advisee and need to decide whether to trust the information provided by the advisers. Advice utilization operationalizes trusting and reciprocal behaviors in these paradigms. In general, two estimates are required from participants, before and after seeing the advice of an adviser, and the degree to which participants revise their opinion after receiving advice measures their willingness to trust the received information and hence, by proxy, its source. Participants' trust and reciprocity are further modulated in these paradigms by manipulating social characteristics of the advisers, such as their competence, confidence, and kindness (Biele et al., 2009, 2011; Hertz et al., 2017; Mahmoodi et al., 2018; Meshi et al., 2012; Toelch et al., 2014).

Another experimental setting is the economic game known as the investment (or trust) game (Berg et al., 1995) (see also Chapter 2). In a standard investment game, a player in the role of investor receives an initial endowment and needs to decide whether to share some of it with a partner

Gabriele Bellucci is supported by the Max Planck Society. Jean-Claude Dreher was funded by the IDEX-LYON from the University of Lyon (project INDEPTH) within the program Investissements d'Avenir (ANR-16-IDEX-0005) and by the LABEX CORTEX (ANR-11-LABX-0042) of the University of Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency (ANR), and grants from the ANR and the National Science Foundation, within the Collaborative Research in Computational Neuroscience program (ANR-16-NEUC-0003-01). Corresponding author: Gabriele Bellucci (gbellucc@gmail.com).

in the role of trustee. If any money is shared, the amount is multiplied (usually tripled) and passed on to the trustee. The trustee needs now to decide whether to share any portion of the multiplied amount back to the investor or keep the entire amount. The investor's decision is regarded as a trust decision, while the trustee's decision is regarded as reciprocity (Chaudhuri & Gangadharan, 2007; Csukás et al., 2008). Such a version of the investment game has mainly been used to study trust in reciprocity and cooperation, but lends itself to also investigate more strategic forms of behaviors (Camerer, 2003; Chaudhuri et al., 2002; Krueger et al., 2007, 2008).

In Section 8.2, we review evidence about the type of social information about others (i.e., their social characteristics and psychological traits) that functions as central determinant and predictor of trustworthiness impressions and trusting behaviors. In Section 8.3, we turn to examine the learning dynamics and computational mechanisms that unravel how people integrate these different pieces of social information about others to update their trustworthiness beliefs and revise their trusting behaviors. In Section 8.4, we review neuroimaging evidence on the neural underpinnings of these learning processes. Finally, in Section 8.5, we draw conclusions on the neurocomputational processes underlying learning to trust, draft possible insights for clinical research, and propose future directions for forthcoming investigations.

8.2 Shades of Trust

8.2.1 The Traits One Can Trust

Across disciplines, trust is defined as the willingness to be vulnerable to another on the basis of positive expectations of the other's intentions and behaviors (Rousseau et al., 1998). A similar multidisciplinary definition of trust underlines that trust occurs when no control over the other's behavior is possible (Mayer et al., 1995). The absence of control refers to cases in which an agent's decision outcomes depend on someone else's decisions that the agent cannot (or does not want to) monitor or regulate. Interactions where rules and monitoring activities are in place, such as contracts that regulate the expected behavior of another person and define deterrents for deviant behaviors, do not require trust. Importantly, when an agent eschews the imposition of regulatory frames, the ultimate decision to trust lies in the positive expectations of the other person's intentions. Information that drives an individual to focus on the possible bad

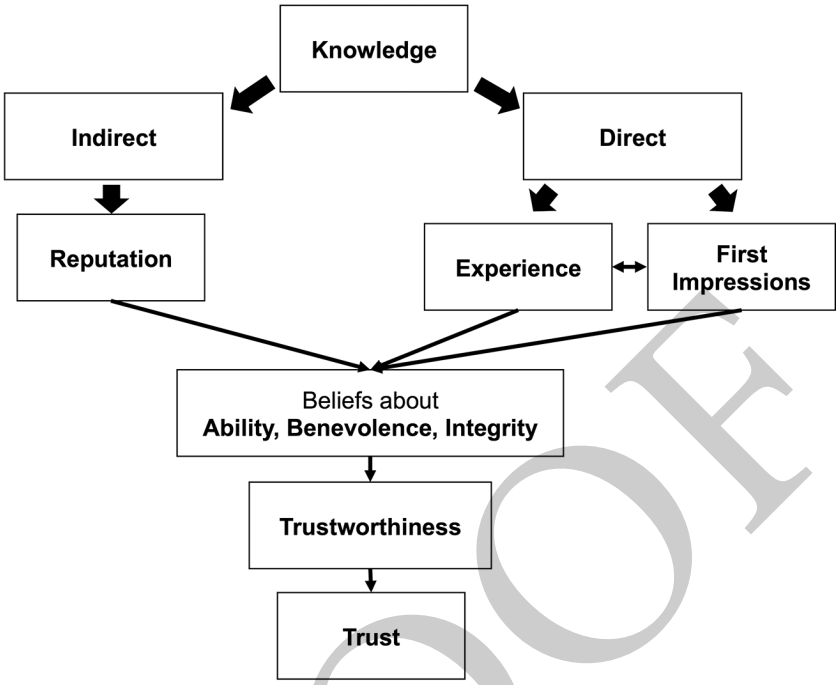

Figure 8.1 Schematic formation of trustworthiness impressions guiding trust behavior. Information leading to the formation of trustworthiness judgments is based on either direct or indirect information about the partner's character and behaviors. Indirect information refers mostly to the other's reputation. On the contrary, direct knowledge might vary in nature depending on how well the other person is known. Literature on trust can be roughly divided into studies on first impressions about others and repeated interactions (establishing a history of experiences with the other person). This information helps build beliefs about three main characteristics of the other person: her ability, benevolence, and integrity. Judgments about others' trustworthiness are formed on the basis of these beliefs and ultimately guide trust decisions in social interactions.

intentions of another, for instance, by simply calling the other person an "opponent," evokes negative expectations that corrupt trust (Burnham et al., 2000). Similarly, showing distrustful behavior toward another person elicits the expectation that the target of such distrustful behavior is herself untrustworthy and need to be avoided, or ostracized (Hillebrandt et al., 2011). However, how do people form such expectations?

Expectations might arise from indirect or direct knowledge (Figure 8.1). Individuals might form trustworthiness judgments based on a person's reputation, which mainly originates in indirect information received from

other people (see also Chapter 7). On the contrary, direct knowledge refers to information gathered from a more or less shallow, personal interaction with the other person. The literature on trust has mainly studied two forms of direct knowledge. First, individuals might rely on subjective (implicit) impressions about the other person's trustworthiness that are formed rapidly and effortlessly (Siegel et al., 2019; Todorov et al., 2009). Implicit trustworthiness impressions have reliably been shown to play a pivotal role in the decision to trust unknown others or strangers in single and anonymous interactions, where individuals do not know anything about the partner except for her physical appearance (e.g., facial trustworthiness) and/or some prior knowledge about her reputation (e.g., indirect reputation of being trustworthy) (Bellucci et al., 2020).

Second, positive expectations can emerge dynamically from experience with the partner during repeated social interactions that give individuals the opportunity to update their beliefs about the partner's character and behaviors (Hula et al., 2018). For instance, an individual can learn the benevolence of her partner over the course of multiple interactions from the partner's kind behavior (Ho & Weigelt, 2005), and be more likely to reciprocate if she learns that the partner has previously trusted (McCabe et al., 2003). Importantly, also in repeated interactions, implicit trustworthiness impressions formed at the beginning of the interaction with the partner are not completely discarded but slowly integrated into the learning dynamics underlying trusting interactions, influencing the final trustworthiness judgments resulting from the social interaction (Chang et al., 2010).

Now, the question arises as to what kind of information individuals seek, gather, and integrate in these different contexts to make inferences about the other person's trustworthiness. A classic model posits that at least three characteristics (or traits) are central to the formation of trustworthiness impressions and the final trust decision (Figure 8.1). These characteristics hence constitute factors of perceived trustworthiness or antecedents of trust, and include the following: ability, benevolence, and integrity (Mayer et al., 1995).

Ability refers to skills and competencies of the trustee that are taken into consideration before a trust decision is made. The trustee might want to use her skills to help the trustor reach his goals. In general, the trustor might be more or less dependent on the trustee's skills for his actions. This interdependency creates the conditions for trust, as the trustor needs to believe that the trustee's skills can ultimately benefit him and that he cannot easily and inexpensively do well without them. These skills and

abilities might be domain- or situation-specific, as individuals would rely on a doctor for a medical problem and ask a lawyer for a legal counseling and not vice versa. However, this restriction does not need to be valid in general, as competent individuals are perceived as domain-general authorities and might be asked for advice in other domains than the one they are known to be expert in. One reason for this is that individuals turn to experts not only because of their technical expertise in a specific field but also for their general, analytic skills that made them the experts they are in the first place. Hence, individuals might reach out to other-field experts to sample a different type of information from them, for instance, a methodological approach to make a specific decision, instead of requiring a specific solution to the problem at hand. Finally, individuals might also turn to experts just to receive a boost in their confidence.

Benevolence refers to the positive intentions and attitudes of the trustee. A trustee is benevolent to the extent to which she is willing to engage in actions that benefit the trustor despite their cognitive, physical, or monetary costs and beyond a strictly egocentric motivation for self-profit. With this respect, being altruistic and inclusive induces trustworthiness impressions in others that make the altruistic individual more likely to be trusted in social interactions, whereas excluding others for no ostensive reasons evokes impressions of a malevolent character that promote distrustful behaviors (Delgado et al., 2005; Frost et al., 1978). Benevolence is central to one's decision to trust (King-Casas et al., 2005). Signs that the other might have bad intentions or might not be well minded decrease one's willingness to trust, even when other information is available that would otherwise evoke trustworthiness impressions. For instance, individuals are more likely to be influenced by the advice of nonexperts with good intentions than by the advice of experts with likely bad intentions (McGinnies & Ward, 1980), suggesting that benevolence outweighs expertise when these types of information compete. Furthermore, unconditional kindness, but neither positive nor negative reciprocity, has strongly been associated with trusting behaviors (Thielmann & Hilbig, 2015). However, further studies are needed to better understand how individuals weight different trustworthiness sources for a decision to trust, as an expert is likely to be trusted despite a reputation of being untrustworthy if one's personal experience is inconsistent with the expert's reputation.

Finally, integrity refers to the extent to which the trustee adheres to a set of principles that the trustor finds acceptable. This factor closely relates to the moral dimension or the moral character of the trustee, and highlights

the trustee's behavioral consistency that reflects the trustee's congruence to a determined set of values. For example, individuals with strong moral characteristics are perceived as more trustworthy, trusted more, preferred as social partner, and are believed to be more likely to reciprocate trust (Everett et al., 2016). However, simple behavioral consistency is not sufficient, as a trustee might consistently act in a self-serving manner (see also the relationship between trust and behavioral predictability). Recently, it has been proposed that integrity is not only an important characteristic of the trustee but also a relevant characteristic of the trustor, although for different reasons. While in the trustee integrity signals trustworthiness, in the trustor it impels to trust. Previous studies found, for example, that the sense of compliance with a trust norm and the sense of respect for the other person predict individual trust (Dunning et al., 2014), suggesting a normative (moral) component inherent to trust decisions (Dunning et al., 2019).

It has been shown that these factors of trust and trustworthiness are partly captured by three personality traits. In particular, a model of personality, namely the HEXACO Personality Inventory, has been operationalized in a self- and observer-report instrument consisting of six dimensions – Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O) (Ashton & Lee, 2007; Ashton et al., 2004; Lee & Ashton, 2004). Studies using this inventory show that propensity to trust relies on positive expectations of social partners held by individuals who do not perceive others as exploitative (i.e., higher agreeableness), who feel optimistic in general (i.e., higher extraversion), and who are not themselves inclined to exploit others (i.e., higher honesty-humility) (Ashton et al., 2014).

8.2.2 The Distrusting Side of Control

In Section 8.2.1, we described factors that play a role in an individual's willingness to trust others. However, we have not yet touched on a related topic. In fact, it is not enough to identify which information is important for a trust decision. Most of the time, information is noisy or comes from individuals who are themselves untrustworthy or might have reasons to hide relevant information. To decrease decisional uncertainty originating from noisy sources, individuals might engage in different control strategies to check the other person's trustworthiness and boost one's confidence before trusting. But how much information is necessary to trust someone?

Is not seeking out more information to prove a source's reliability as such already a sign of distrust? The trade-off between exerting control (increasing one's confidence about the other person's trustworthiness) and giving up on control (blind trust) is extremely fragile.

Previous evidence has consistently shown that attempts to exert control signal untrustworthiness impressions and decrease others' trust. For example, requiring a partner to give up on control decreases the partner's willingness to trust, likely because it signals to the partner that she is not trusted (Das & Teng, 1998, 2001; Malhotra, 2004). That control behavior reflects untrustworthiness impressions was first investigated in a seminal work by Strickland (1958), where participants as supervisors needed to decide whether to monitor other participants playing the role of workers. Supervisors were incentivized to keep the workers' performance high over the course of the task but, at first, could only monitor in predetermined trials (monitored trials). However, after having received feedback about the history of the workers' performance in both monitored and nonmonitored trials at the end of the first part of the experiment, supervisors were allowed to freely choose whom to monitor in the second part. Results showed that workers who performed poorly in nonmonitored trials were judged as less trustworthy and monitored more often by supervisors.

Complementarily, individuals refrain from control behaviors to avoid inducing untrustworthiness impressions that might have a negative impact on a partner's decisions. For instance, in a recent study, participants had the opportunity to sample information about the history of their partners' reciprocal behavior before a decision to trust (Ma et al., 2020). In one condition, participants were told that the partner would be informed about how much they sampled (overt sampling condition), while in the other condition they were told that the partner would not know about their information-sampling behavior (covert sampling condition). Behavioral results showed that participants sampled less when sampling was overt and reported that they believed overt sampling information would make the partner's reciprocation less likely, suggesting that participants refrained from sampling to avoid inducing negative impressions.

Attempts to impose binding contracts have similar effects on trustworthiness impressions and trusting behaviors. Malhotra and Murnighan (2002) propose that trust and binding contrasts represent two mutually exclusive mechanisms of social behavior control and regulation. Trust is an informal mechanism of risk management for uncertainty reduction in social interactions, whereas binding contracts represent more a formal mechanism thereof. Importantly, while binding contracts facilitate

exchanges and allow for successful negotiations, their enforcement erodes the development of interpersonal bonds and the establishment of interpersonal trust, which emerges only in situations where the partner's good intentions can be put at test (e.g., when there are incentives not to cooperate). Importantly, enforcing binding contracts are detrimental to trust, while trust increases in situations regulated by nonbinding contracts. This is because nonbinding contracts allow for attributions of positive, dispositional attitudes, as the partner's behavior is not dictated by exogenous, contextual factors (like in binding contracts) but by her personal choices and personality. As nonbinding contracts enable inferences on personal attributions for cooperation, they provide a better basis for building interpersonal trust and control uncertainty in several social interactions (Malhotra, 2004; Pillutla et al., 2003).

This raises the question as to whether efforts to detect trustworthiness can be seen as attempts to predict a partner's behavior for uncertainty reduction, given that trustworthiness hints at an individual's behavioral consistency (Lewis & Weigert, 1985). For instance, if I know that my decisions' outcomes hinge on your well-minded behavior, information about your trustworthiness will decrease the outcome uncertainty associated with my decisions. All things being equal, if you are trustworthy, the outcome I expect is the one that will realize. Similarly, if I have reasons to believe that you are not to be trusted, I will probably gauge the chances to be pretty low that a particular outcome contingent on our joint decisions will realize. Hence, trust might be compared to a probability distribution over outcome occurrences, giving the idea that trust is a form of risk (Coleman, 1990; Deutsch, 1958; Luhmann, 1979). However, recent empirical results refute this equation between trust and risk.

For example, previous studies have found no relationships between risk preferences and trust (Ashraf et al., 2006; Berg et al., 1995) (see also Chapter 5). Moreover, risk-averse behaviors in trusting interactions are very different from risk-averse behaviors in gambling contexts, as they yield different individual aversion parameters that do not map onto each other (Fairley et al., 2016). Several studies have provided evidence on the differences between trust and risk decisions. For example, individuals have been shown to be more reluctant to take a chance on another individual than on a lottery that randomly determines decision outcomes (Bohnet & Zeckhauser, 2004). Other studies have suggested a similar difference but in the opposite direction. In particular, a reduced willingness to trust a partner in Bohnet and Zeckhauser (2004) might be due to the fact that

the partners had incentives to be untrustworthy (Snijders & Keren, 2001) and participants might have underestimated the proportion of those who would reciprocate (Dunning et al., 2019; Fetchenhauer & Dunning, 2009). In a recent study, individuals were found to be more likely to play a gamble when its outcomes depended on a partner with no incentives to be malevolent as opposed to chance (Bellucci et al., in press). These results accord with previous evidence that individuals are more likely to trust a partner who reciprocates frequently than playing with a slot machine that rewards with the same probability (Chang et al., 2010). Interestingly, the opposite effect was observed for untrustworthy partners. Specifically, participants are less willing to trust a partner who reciprocates infrequently than playing with a slot machine that rewards with the same probability (Chang et al., 2010). Overall, these results suggest that a partner's intentions are central to trust decisions in social interactions, which do not simply reduce to risk behaviors in nonsocial interactions.

A recent study extended these results. By eliminating incentives to be untrustworthy, participants were observed to have similar trust levels toward partners whose reputation was unknown and partners with a reputation of being trustworthy (Bellucci & Park, 2020). On the contrary, all untrustworthy partners were distrusted to a similar extent despite different degrees of untrustworthiness. These results suggest that people discriminate degrees of trustworthiness in a coarser way than risk probabilities – being more extreme in their trustworthiness perceptions. In this sense, trustworthiness impressions might be more discrete or even categorical, likely because another agent's behavior is traced back to a stable personality that is believed to bear a higher degree of consistency and coherence than agency-independent events. These behavioral observations have important implications for learning dynamics in social contexts.

These studies suggest that behavioral consistency, as a feature of an individual's trustworthiness, is pivotal for trust. Yet, trust cannot be reduced to a form of control behavior nor trustworthiness to predictability. This is because trust signals giving up on exerting control over other individuals. Indeed, people with low need to control others are perceived as more trustworthy, while attempts of control signal distrust (Frost et al., 1978). And finally, predictable peers are not necessarily trusted and the learning processes underlying the formation of trustworthiness beliefs about social partners manifest themselves as fundamentally different from learning patterns linked to the learning of outcome probabilities and event contingencies in nonsocial contexts (such as the risk domain).

8.3 Learning to Trust

8.3.1 Computational Models of Learning

The studies discussed so far provide evidence on behavioral differences across social and nonsocial contexts that hint at different underlying learning dynamics. Now, we turn to studies that analyze and formalize these learning dynamics (Cheong et al., 2017; Park et al., 2019). As previously mentioned, trustworthiness beliefs rely on indirect knowledge originating from another person's reputation, and from direct knowledge based on subjective first impressions or dynamic learning of the other's behavior through repeated experiences (Figure 8.2). This dynamic learning has been linked to reinforcement learning processes. Reinforcement learning describes how organisms learn by trial and error to predict and acquire rewards (Gershman & Daw, 2017). Reinforcement learning relies on prediction error (PE, or surprise), which signals the discrepancy between actual and expected rewards (Rudebeck et al., 2013; Tsuchida et al., 2010). The Rescorla–Wagner model formalizes learning as trial-by-trial updates of expectations according to the current prediction error (Rescorla & Wagner, 1972). Expectations (or predictions) of the obtainable reward

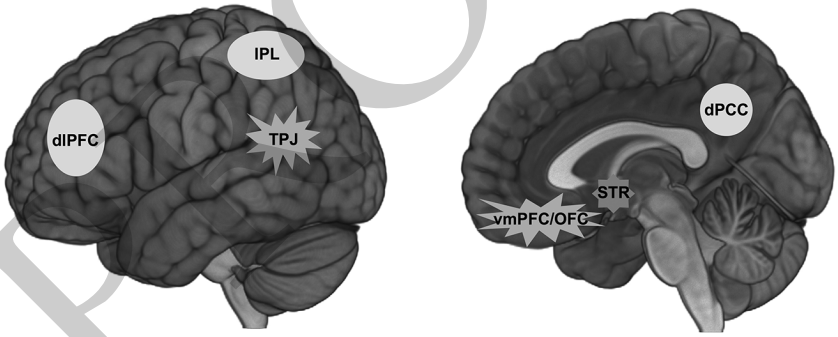

Figure 8.2 Brain regions involved in trust learning.
The STR underlies action–outcome associative learning allowing the identification of discrepancies in another person's behavior. Activity in the ventromedial prefrontal cortex/orbitofrontal cortex (vmPFC/OFC) encodes character trait information about the other. The temporoparietal junction (TPJ) represents current beliefs about the other's likely behavior. The inferior parietal lobule (IPL), dorsolateral prefrontal cortex (dlPFC), and dorsal posterior cingulate cortex (dPCC) integrate current feedback about others' trustworthiness with knowledge about their reputation to guide present and future decisions.

associated with a stimulus s are encapsulated by $V_t(s_t)$, that is, the value associated with a particular stimulus at trial t. Given that R_t describes the received reward at trial t, updates of reward expectations are formalized as follows:

$$PE_t = R_t - V_t(s_t), \tag{1}$$

where PE_t is the prediction error at trial t. PEs are smaller when the received reward is close to what is expected and larger when the received reward is far from what is expected. The PE can be thought of as the quantity that determines how much update is needed. The more we learn about the associative strength between a particular stimulus and its reward, the less learning occurs, as the expected reward approximates the actual reward. This implies that our expectation (prediction) of a reward R given a stimulus s will increase in accuracy over time with a concomitant reduction of discrepancy (error). Importantly, this learning process is not linear but hinges on individual learning parameters that determine the size of the update step and so affect the magnitude of the changes involved. Hence, the predicted value of a stimulus s on the next trial t is updated as follows:

$$V_{t+1}(s_t) = V_t(s_t) + \alpha * PE_t, \tag{2}$$

where α is the individual learning rate (usually between 0 and 1) updating the reward expectation. The value of a stimulus updated with large learning rates reflects the more recent history of received rewards, as reward expectations are updated by more heavily weighting the current reward. On the contrary, the value of a stimulus updated with small learning rates reflects the more remote history of rewards, as reward expectations are updated by more strongly weighting the expectations.

On the neural level, PEs are carried by dopaminergic neurons (Montague et al., 1996; Schultz et al., 1997). In particular, phasic dopaminergic responses in the midbrain, striatum (STR), and orbitofrontal cortex (OFC) show properties similar to reward PEs as described by the Rescorla–Wagner model (Hollerman & Schultz, 1998; Schultz, 2000; Schultz et al., 1997). Combining reinforcement learning models with neuroimaging, correlations between neural signals and model-based PEs have been observed in humans as well, such as in the ventral striatum (vSTR) (Dreher et al., 2006; O'Doherty et al., 2004), anterior insula (Preuschoff et al., 2006), hippocampus (Vanni-Mercier et al., 2009), OFC (Li et al., 2016; Metereau & Dreher, 2015), and midbrain (Howard & Kahnt, 2018). While reinforcement learning models were

initially employed to describe instrumental learning (i.e., Pavlovian conditioning), they have recently been used to characterize other forms of learning (Joiner et al., 2017), like learning another person's character traits (Biele et al., 2009; Chang et al., 2010; Delgado et al., 2005; Fouragnan et al., 2013).

In social interactions, learning is complicated by the intentional stance of the interacting partner. Hence, an agent does not only need to learn the associations between events to identify actions for reward maximization and loss avoidance, but it also needs to consider the intentions of the interacting partner to correctly weight the outcomes of the social interaction. Indeed, positive outcomes can hide bad intentions and negative outcomes can be the by-product of a well-minded action. Hence, in social contexts, events bear information about the partners' action utilities, which allow estimations of their interests, inferences on their intentions, and predictions of their future behavior. A well-known formalization of action utilities for social behaviors in simple two-person interactions is provided by the Fehr–Schmidt model (Fehr & Schmidt, 1999).

Within the Fehr–Schmidt framework, kind and unkind intentions are reflected by fair and unfair behaviors, which can be described as self- and other-centered inequity aversion. Individuals are inequity averse if they dislike outcomes that are perceived as inequitable (Fehr & Schmidt, 1999). Inequity aversion is self-centered when individuals care about their own material payoff relative to the payoff of others and increases with increasing disadvantageous inequity (the less the individual has relative to others). Inequity aversion is other-centered when individuals care about the payoff of others relative to their own material payoff and increases with increasing advantageous inequity (the more the individual has relative to others). Given this definition of fairness, which is exclusively based on the importance (or weighting) people ascribe to outcomes of joint behaviors in social interactions, fair and unfair behaviors can be formalized with a utility function. Let $x = x_1, \ldots, x_n$ denote the vector of outcomes (e.g., monetary payoffs) for specific actions of a set of n players indexed by $i \in \{1, \ldots, n\}$. The utility function of player $i \in \{1, \ldots, n\}$ is given by:

$$U_i(x) = x_i - \alpha_i \, max\left\{x_j - x_i, 0\right\} - \beta_i \, max\left\{x_i - x_j, 0\right\}, \qquad i \neq j. \quad (3)$$

As can be seen in Equation (3), the utility U_i of player i is a weighted sum of the utility gain from player i's outcome (i.e., x_i) and the utility losses from disadvantageous (i.e., $\alpha_i \, max\left\{x_j - x_i, 0\right\}$) and advantageous (i.e., $\beta_i \, max\left\{x_i - x_j, 0\right\}$) inequities. That is, when player i has much less than player j (i.e., $x_j > x_i$), the utility of player i's outcome is reduced by

disadvantageous inequity aversion (or envy), as individuals dislike having less than others. Similarly, when player i has much more than player j (i.e., $x_j < x_i$), the utility of player i's outcome is reduced by advantageous inequity aversion (or guilt), as individuals feel guilty for having more than others. In Equation (3), α_i and β_i are subject-specific parameters that capture individual differences in how much people value the disutilities from disadvantageous and advantageous inequities.

Despite this model neatly capturing mentalizing processes (e.g., how people think about others' fairness) in trusting interactions (Hula et al., 2015, 2018; Khalvati et al., 2019; Xiang et al., 2012), it still has important limitations. For instance, it does not take into consideration that the costs of mentalizing are cognitively nonnegligible (e.g., limited working memory), especially when mentalizing involves prospection (Na et al., 2019). To reduce such costs, individuals use trustworthiness as a safety signal that allows them to engage in costly inferences on a partner's behavior only when the partner's trustworthiness is low or unknown (Sperber et al., 2010; Wu et al., 2020). Hence, computational models attempting to formalize social learning and mentalizing dynamics need to reliably address such a trade-off.

8.3.2 The Endurance of a Good Reputation

In social interactions, individuals are mainly tasked with the challenge of learning another person's character. The character of another points to generally stable traits that allow reliable inferences on the other person's intentions and actions, and on the quality of the information she has and communicates (e.g., her credibility). Reliable group members are believed to be a source of accurate information (Gordon & Spears, 2012) and information sharing is associated with trust (Burt & Knez, 1995). In contrast, a bad reputation has a deleterious effect on beliefs about others' trustworthiness and credibility (McGinnies & Ward, 1980; Reichelt, Sievert, & Jacob, 2013; Weiner & Mowen, 1986). Similarly, lying, both in the forms of concealment of information and sharing of inaccurate information, calls for norm-enforcing behaviors such as punishment (Fehr & Fischbacher, 2003; Sánchez-Pagés & Vorsatz, 2007).

Because reliable, trustworthy, and credible individuals are likely to benefit the group as a whole by sharing accurate information, engaging in prosocial behaviors, or holding their word in task assignments and promises (Becker et al., 2017; Galton, 1907; Mellers et al., 2014; Sjöberg, 2009), societies have a strong interest of promoting and

reinforcing good, prosocial qualities. Reputation functions as a proxy for credit assignment based on a person's past behavior in a social group and represents an important mechanism for cooperation in heterogeneous, large-scale societies (Fehr & Fischbacher, 2004). Reputation works like a social tag to easily distinguish good cooperators from bad ones, especially when no prior information about the partners is provided. In this sense, it can be used as a prior for first interactions with other people that is subsequently updated based on new incoming information about the partner's behavior. Hence, with new information about a partner's character traits, individuals should be able to successfully update their trustworthiness beliefs about the partner. However, we will see that reputation strongly biases people's initial impressions and learning, even impacting neural responses to rewards.

In particular, individuals believe that positive traits are more frequent than negative traits and that positive traits are more easily lost than negative ones (Rothbart & Park, 1986). Consequently, the theory of trust asymmetry posits that a good reputation is more easily lost than gained and hence individuals are faster at adapting their behavior after feedback about another person's untrustworthiness than after feedback about another person's trustworthiness. This prediction is based on early evidence that trust is decreased more by negative events than increased by positive ones (Slovic, 1993). These results chime well with a general pattern of evaluations of good and positive feedbacks in humans, which suggests that negative information (e.g., negative events and monetary losses) loom greater than positive ones (e.g., positive events and monetary gains) (Kahneman & Tversky, 1979; Platt & Huettel, 2008; Tversky & Kahneman, 1992).

The theory of trust asymmetry relies on evidence that impressions about favorable traits (such as one's ability and integrity) require more instances to form than impressions about unfavorable traits and unfavorable traits are harder to lose. For example, a previous study has found that individuals more strongly distrust nonexpert advisers with a reputation of being expert than trust expert advisers with a reputation of being nonexpert (Yaniv & Kleinberger, 2000). Hence, individuals seem to more readily distrust those with a good reputation when current disconfirming information is provided. On the contrary, a bad reputation hampers attempts to regain trust, confirming that negative experiences with a social partner have greater influence than positive ones (Yaniv, 2006). Despite this apparently confirming evidence, another important phenomenon that generally goes overlooked is that impressions about others' traits that are more easily

formed are also harder to lose irrespective of the favorability of the trait (Rothbart & Park, 1986). Consequently, if it is easier to form a positive impression (maybe because of some prior beliefs or the particular situation and context), that impression will be also more enduring and harder to lose.

In the previous study by Yaniv and Kleinberger (2000), participants had to learn advisers' expertise from feedback about complicated factual knowledge and the advisers were deprived of intentionality, as participants believed the advisers were always communicating their best guess. A recent study tried to overcome this limitation by allowing advisers to deliberately decide whether to be honest or dishonest in advice giving (Bellucci & Park, 2020). Being honest was a reasonable behavioral strategy for advisers to build a good reputation that could have paid off in a subsequent interaction, where participants could repay the advisers for their advice-giving behavior. This established a social context that facilitated the formation of positive trustworthiness impressions. However, being dishonest was not disincentivized and represented a cognitively, less costly strategy. Results show that initial levels of trust were very high for all advisers, confirming that the experimental situation induced positive priors about the interacting partners. However, after a couple of trials (during the reputation-building phase) participants realized that some advisers were not honest and quickly adjusted their behavioral strategy accordingly. At the end of the reputation-building phase, participants could clearly distinguish advisers with a reputation of being dishonest from advisers with a reputation of being honest.

Now, later on in the experiment, advisers with a bad reputation began to show signs of honesty, while advisers with a good reputation turned dishonest. Notably, participants successfully revised their first impressions of the advisers with a bad reputation and trusted them increasingly more, but did not change their behavior toward the advisers with a good reputation, suggesting a learning impairment for the latter. A reinforcement learning model indicated a reputation-dependent asymmetry in the valuation of the advisers' honesty and dishonesty. Contrary to the theory of trust asymmetry, participants did not weight dishonesty more than honesty, and the dishonesty of the advisers with a good reputation was not weighted more than the dishonesty of the advisers with a bad reputation. On the contrary, a good reputation strengthened the valuation of honest behavior, whereas a bad reputation corroborated valuation of dishonest behavior (Bellucci & Park, 2020).

Virtually the same results were found by Siegel et al. (2018) in a moral decision-making task, where participants observed the moral decisions of

other co-players and rated their moral impressions of them. Participants learnt the reputation of the co-players who initially were either bad or good but then began to make decisions that were more or less moral than previously. Results showed that participants updated their moral impressions more for bad than for good co-players, suggesting that good impressions of a moral co-player did not optimally change when that co-player was less moral than previously, while bad impressions of an immoral co-player did not impair the accurate tracking of the co-player's morality (Siegel et al., 2018). Comparable results were found in a study by Fareri et al. (2012), where participants learnt the moral (first interaction) and trustworthy character (second interaction) of their partners in two consequential interactions. Despite similar trustworthy behaviors of the partners in the second interaction, participants were more likely to trust those partners who established a reputation of being moral in the first interaction. A reinforcement learning model indicated a reputation-dependent asymmetry in belief updating with trustworthiness beliefs being more likely updated after positive feedbacks for the partner with a good reputation but after negative feedbacks for the partner with a bad reputation. Finally, a recent study extended these results showing that such reputational bias can be observed only in social interactions with trustworthy partners but not in nonsocial contexts when playing with rewarding slot machines (Lamba et al., 2020).

These results suggest that the three antecedents of trust discussed in the previous paragraph contribute to trustworthiness impressions differently depending on the context. Moreover, traits closely related to the intentionality dimension are learnt faster and impact learning and trusting behavior more profoundly than traits associated with individual ability and expertise. However, to our knowledge, no study has directly compared how different antecedents of trust impact people's trustworthiness impressions in the same experiment. Future studies are thus necessary to understand whether and how different sources of information about others' traits are sampled and integrated for a decision to trust. In particular, computational research on how intentions of others are inferred and integrated into outcome valuations is still at an embryonic stage. Recent investigations implementing the Fehr–Schmidt model have revealed different levels of mentalizing sophistication. For example, a study by Xiang et al. (2012) estimated individual depth-of-thoughts (cognitive levels of sophistication of one's model of a partner's intentions) from trusting behavior in response to a partner's reciprocity. Results indicated at least three levels of depth-of-thought and showed that participants with the least sophisticated model of

the partner's intentions were less successful at learning the partner's reciprocity for efficient cooperation. By highlighting the importance of mentalizing processes for accurate learning and adequate behavior revision in social interactions, these findings call for more investigations on the intentionality aspect inherent to social learning dynamics and the interactions with the types of situations where they occur.

8.4 Brain Trust

8.4.1 Identifying Signs of Distrust

Different neuroimaging studies have addressed the question as to which brain areas are involved in learning a partner's trustworthiness during social interactions (see also Chapter 7). An early study by Delgado et al. (2005) investigated neural responses to feedback about behaviors of partners of different moral character traits. Results demonstrated striatal responses underlying both a decision to trust a partner and learning that the partner reciprocated. Further, higher activations in the vSTR (Figure 8.2) were observed for trust in the partner with a bad reputation and in the caudate for reciprocal behavior of the same partner. On the contrary, no differences in neural responses in these regions were observed for the partner with a good reputation. These differences in neural responses in the STR might underlie the behavioral patterns discussed in the previous paragraph. In particular, as individuals seem to optimally revise their impressions of partners with a bad but not a good reputation, recruitment of the STR might support such behavioral updating for more efficient learning by integrating the relevant information. The caudate nucleus might specifically underlie updating of behaviorally relevant information about others, as this region is recruited particularly during interactions with individuals with a bad reputation (Wardle et al., 2013), and is consistently engaged in information processing about another person's behavior in trusting interactions (Bellucci et al., 2017).

Similarly, a study by King-Casas et al. (2005) found the caudate to be involved in learning from a partner's reciprocal behavior. In particular, activations in the caudate peaked during the feedback phase at the beginning of the trusting interaction but shifted over time from the period of the revelation of the other person's behavior to the period prior to it. These findings seem to suggest that striatal activity tracks the partner's reciprocity over time and its temporal shift might represent a signature of learning dynamics, according to which the STR integrates information about the

other's reputation for belief updating in early states but signals predictions or inferences on the other's likely reciprocal behavior in later stages. Another study has provided first evidence for the selective role of the STR in learning others' reciprocity. In particular, neural responses to cooperative partners with a reputation for reciprocity were observed both in the STR and OFC, another important region for learning (Gottfried & Dolan, 2004; Phan et al., 2010; Rudebeck & Murray, 2014). However, greater striatal activations for the cooperative than the uncooperative or neutral partners were observed only in the STR. Moreover, these striatal activations were specifically driven by stronger neural responses to feedbacks on the reciprocity of the cooperative partner (Phan et al., 2010), suggesting that striatal neural responses were selective to information received about the behavior of partners with a reputation for reciprocity.

If striatal activity plays a role in belief updating about others' trustworthiness for behavioral adaptation, activity in the STR should specifically track value updating over time. In particular, within the framework of reinforcement learning models, striatal activity should reflect PE trustworthiness signals in response to the partner's trustworthy behavior. The temporal shift of striatal activity in King-Casas et al. (2005) points to a role for the STR in PEs about others' traits (e.g., social PEs). Another neuroimaging study provided first model-based evidence to this hypothesis (Fareri et al., 2012). Model-based PEs correlated with neural signal in the STR during the revelation of the other person's behavior, suggesting that the STR was integrating information about the partner's trustworthiness based on feedback about the partner's reciprocal behavior. Importantly, however, no differences in neural signals were observed for partners with different reputations. Hence, despite this evidence on the involvement of striatal activity in trustworthiness belief updating for behavior adaptation to others' reputation, it is unclear whether the STR specifically reflects an individual's expectations of the partner (e.g., her reputation) or rather more general computational processes, such as information integration processes related to action–outcome associative learning.

8.4.2 Tracking Trustworthiness

The evidence discussed so far suggests a role of striatal regions in learning dynamics during trusting interactions, which are also involved in learning of other characteristics of social partners such as their prosocial tendencies (Lockwood et al., 2016). Further, recent pharmacological evidence has provided converging evidence for a role of the dopaminergic system in

trusting behaviors (Bellucci et al., 2020). Given the high shared variance between trustworthiness and attractiveness – a primary reward (Bellucci et al., 2020; Stirrat & Perrett, 2010; Wilson & Eckel, 2006), it is not surprising that trustworthiness recruits brain areas known to encode several rewards (Aharon et al., 2001; O'Doherty et al., 2003; Pegors et al., 2015; Winston et al., 2007). Further, given the role of the STR in learning, the involvement of striatal structures during social learning seems to confirm investigations in the nonsocial domain. However, at least two limitations can be highlighted. First, most of this evidence comes from experimental paradigms, such as economic games, where "trust decisions" are inextricably intertwined with monetary rewards. Thus, paradigms and computational models that investigate belief updating in these experiments need to control for neural signatures associated with reward PEs, especially because those striatal activations were consistently observed during feedback phases where participants not only learn about the partner's behavior but also get to know how much they earned in the trial. Second, neural patterns in these brain structures are mostly associated with associative learning but not with other forms of learning such as language learning (Ekerdt et al., 2020; Finkl et al., 2020; Price, 2012; Zatorre, 2013) and might hence only represent forms of associative learning involved in social learning. Hence, it is unclear whether they are specific to social learning or rather reflect other processes woven into the social behaviors in those experimental paradigms, such as associative learning to track changes in the environment that might be caused by the other person's behavior.

If striatal neural patterns reflect learning dynamics related to understanding whether the partner is trustworthy, those same neural patterns should generalize to other experimental paradigms. Neuroimaging studies investigating trustworthiness learning in advice-taking paradigms, however, seem to provide a negative answer. For example, a recent neuroimaging study was able to isolate social evaluation signals related to another person's trustworthiness (learnt through the other's honesty and dishonesty) from nonsocial value signals related to rewards (i.e., winnings and losses) received during the feedback phase (Bellucci, Molter, & Park, 2019). Disentangling these two signals further allowed the investigation of neural signatures specifically related to trustworthiness representations and their modulatory effects on reward processing. Results showed that the STR and anterior cingulate cortex specifically encoded reward information, while feedbacks about the other person's trustworthiness were represented in the dlPFC, dorsal posterior cingulate cortex, and parietal cortex (e.g., inferior parietal lobule, IPL) (Figure 8.2). Importantly, neural signal

from these regions was able to predict reputation-dependent trust in the partner during a subsequent interaction, whereas neural signal from the STR was not informative of an individual's future trust (Bellucci, Molter, & Park, 2019). These findings represent the first evidence that other brain regions than striatal structures represent social information relevant to reputation-dependent trustworthiness beliefs.

Another work using a reinforcement learning model further indicates that social PEs about a partner's trustworthiness correlates with activity in the medial prefrontal cortex, TPJ, and posterior superior temporal cortex (Behrens et al., 2008) – important mentalizing brain regions (Grèzes et al., 2001; Koster-Hale et al., 2017; Saxe & Kanwisher, 2003). On the contrary, reward PEs correlated with neural activity in the STR and anterior cingulate cortex, suggesting a dissociation between trustworthiness updating signals and reward PE signals. Similar results were found in a similar paradigm by Diaconescu et al. (2017), despite the use of a different reinforcement learning model. A recent neuroimaging study extends these findings showing that social PEs correlate with neural activity in the medial prefrontal cortex (i.e., OFC) and TPJ (Figure 8.2). Importantly, OFC activity preferentially encoded information about another person's trustworthiness but not subjective trustworthiness impressions of the other person (Bellucci & Park, in press). On the contrary, stronger functional connectivity between the OFC and TPJ was associated with more favorable trustworthiness impressions, suggesting that the OFC entails positive character trait information that supports belief updating about another person's behavior in the TPJ for trustworthiness impressions formation.

Further evidence on these dissociable signals comes from results observed in a different paradigm investigating the neural correlates of strategic behavior in an inspection game (Hampton et al., 2008). In this study, pairs of participants played in a two-player strategic game in which opponents have competing goals. One participant played as employer and could either inspect (distrust) or not inspect (trust), while the other played as employee and could either work (be trustworthy) or shirk (be untrustworthy). Computational modeling results showed that participants were not only using representations of the opponents' future choices to guide their own choice, but were also incorporating knowledge of how one's own actions influenced the partner's strategy, that is, how much the partner was showing reciprocal cooperation, which is a central feature of trustworthiness signals (Mahmoodi et al., 2018). Moreover, at the time of outcome revelation, influence updates of the partner's inferred trustworthiness and reward PEs were correlated with different neural signatures. In particular,

trustworthiness update signals were found in the posterior superior temporal sulcus, while reward PEs were found in the vSTR. Importantly, also in this study, parameter values for trustworthiness updating and reward PEs were estimated from the same model, suggesting that they reliably captured independent and dissociable signals.

Finally, evidence on the role of mentalizing brain regions in trusting behaviors comes from brain network analyses of resting-state functional brain connectivity. Those brain regions that have been found to be consistently engaged by mentalizing tasks also cluster into an interconnected network at rest, known as the default-mode network (DMN; Alves et al., 2019; Ingvar, 1974; Raichle et al., 2001). Previous evidence has shown that resting-state functional connectivity bears predictive information about individuals' behavior (Rosenberg et al., 2016), personality traits (Adelstein et al., 2011; Kunisato et al., 2011), and social preferences (Hahn, Notebaert, Anderl, Reicherts, et al., 2015) (see also Chapter 12). Two studies provided evidence that whole-brain resting-state connectivity also predicts propensity to trust (Hahn, Notebaert, Anderl, Teckentrup, et al., 2015; Lu et al., 2019) but left open the question whether there is any specificity in the brain networks that preferentially represent information underlying trusting behavior. A recent study filled this gap by testing five classic resting-state networks, namely the default-mode, frontoparietal, sensorimotor, cinguloopercular, and occipital networks (Bellucci, Hahn, et al., 2019; Dosenbach et al., 2007, 2010). These networks have been associated with different functions, such as central-executive functions for the frontoparietal network, saliency for the cinguloopercular network, and mentalizing for the DMN. The study by Bellucci, Hahn, et al. (2019) provides evidence that the DMN was the only brain network able to predict individual decisions to trust an anonymous person. Given the need to build mental models of the other person during trust decisions (especially in anonymous settings that abound in occasions for betrayal) (Aimone & Houser, 2013; Van Overwalle & Baetens, 2009) and the role of the DMN in simulating an alternative perspective (Buckner et al., 2008), these results suggest that mentalizing brain regions are pivotal to an individual's propensity to trust, likely because they support simulations of the other person's mind for estimations of her likely future behavior (Fletcher et al., 1995; Van Overwalle, 2009).

In conclusion, these studies indicate specific roles for several brain regions (Figure 8.2). In particular, the STR might be responsible for action–outcome associative learning to identify discrepancies in the other person's behavior. Mentalizing brain regions support the decision-making

process in social contexts via simulations of thoughts and behaviors of others. The interplay between the TPJ and OFC underlies the formation of beliefs about others from character trait impressions, with the OFC encoding character trait information and the TPJ representing current beliefs about the other person's likely behavior. Finally, brain regions in the parietal (e.g., IPL), posterior cingulate, and prefrontal cortices (e.g., dlPFC) integrate current feedback about others' actions with other sources of information, such as previous beliefs and reputational knowledge, to guide decisions and prompt behavior change.

8.5 Conclusions

The results discussed in this chapter provide a first overview of the neurocomputational processes underlying trust learning as a form of social learning. Leveraging mathematical formulations of behavior, the core processes of social learning might be uniquely identified and described. Combining these mathematical parameterizations with neuroimaging techniques allows the investigation of the neural instantiations of those cognitive processes underlying social learning. These attempts not only contribute to a better understanding of social cognition in the healthy population but also help tackle the dysfunctioning processes in clinical disorders (Gromann et al., 2013, 2014; King-Casas et al., 2008; Lis et al., 2016; Maurer et al., 2018; Sripada et al., 2009; Xiang et al., 2012) (see also Chapters 16 and 17).

Despite the important advances achieved until now, a unifying neurocomputational theory of social learning is still lacking. To date, many studies investigating social learning have borrowed reinforcement learning and Bayesian models developed in other fields to study other forms of learning (e.g., associative learning). This concerns not only trust learning but also, for example, social dominance learning (Ligneul et al., 2016). Despite their ability to capture some cognitive processes in play in social interactions, their suitability to satisfactorily describe the complexities of social dynamics is yet to be proven. Processes such as strategic thinking, planning, and social comparisons are not formally captured by reinforcement learning and Bayesian models. On the contrary, other models like the Fehr–Schmidt model provide a neat formulation of social comparison computations but lack a framework for belief updating about the interacting agent in repeated interactions that could account for strategy change and behavior revision. A solution might be to use a Bayesian framework

based on partially observable Markov decision processes, as recently attempted (Hula et al., 2018; Khalvati et al., 2019; Park et al., 2019). In conclusion, while much progress has been achieved in the last few years, a lot of interesting and fascinating work still awaits future, investigative efforts in the field of computational social neuroscience.

References

Adelstein, J. S., Shehzad, Z., Mennes, M., et al. (2011). Personality is reflected in the brain's intrinsic functional architecture. *PLoS One*, 6(11), Article e27633. http://dx.doi.org/10.1371/journal.pone.0027633

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'Connor, E., & Breiter, H. C. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron*, 32(3), 537–551. http://dx.doi.org/10.1016/S0896–6273(01)00491-3

Aimone, J. A., & Houser, D. (2013). Harnessing the benefits of betrayal aversion. *Journal of Economic Behavior & Organization*, 89, 1–8. http://dx.doi.org/10.1016/j.jebo.2013.02.001

Alves, P. N., Foulon, C., Karolis, V., et al. (2019). An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings. *Communication Biology*, 2, Article 370. http://dx.doi.org/10.1038/s42003–019-0611-3

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, 9(3), 193–208. http://dx.doi.org/10.1007/s10683-006-9122-4

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2), 150–166. http://dx.doi.org/10.1177/1088868306294907

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, 18(2), 139–152. http://dx.doi.org/10.1177/1088868314523838

Ashton, M. C., Lee, K., Perugini, M., et al. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, 86(2), 356–366. http://dx.doi.org/10.1037/0022-3514.86.2.356

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 114(26), E5070–E5076. http://dx.doi.org/10.1073/pnas.1615978114

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, 456(7219), 245–249. http://dx.doi.org/10.1038/nature07538

Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., & Krueger, F. (2017). Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Human Brain Mapping*, 38(3), 1233–1248. http://dx.doi.org/10.1002/hbm.23451

Bellucci, G., Hahn, T., Deshpande, G., & Krueger, F. (2019). Functional connectivity of specific resting-state networks predicts trust and reciprocity in the trust game. *Cognitive, Affective & Behavioral Neuroscience*, 19(1), 165–176. http://dx.doi.org/10.3758/s13415–018-00654-3

Bellucci, G., Molter, F., & Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. *Nature Communications*, 10(1), Article 5184. http://dx.doi.org/10.1038/s41467–019-13261-8

Bellucci, G., Münte, T. F., & Park, S. Q. (2020). Effects of a dopamine agonist on trusting behaviors in females. *Psychopharmacology (Berl)*, 237(6), 1671–1680. http://dx.doi.org/10.1007/s00213–020-05488-x

(in press). Value computations under social uncertainty and serotonin.

Bellucci, G., & Park, S. Q. (2020). Honesty biases trustworthiness impressions. *Journal of Experimental Psychology: General*, 149(8), 1567–1586. http://dx.doi.org/10.1037/xge0000730

(in press). Neurocomputational mechanisms of cognitive biases in impression formation.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142. http://dx.doi.org/10.1006/game.1995.1027

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science*, 33(2), 206–242. http://dx.doi.org/10.1111/j.1551-6709.2009.01010.x

Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The neural basis of following advice. *PLoS Biology*, 9(6), Article e1001089. http://dx.doi.org/10.1371/journal.pbio.1001089

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4), 467–484. http://dx.doi.org/10.1016/j.jebo.2003.11.004

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124, 1–38. http://dx.doi.org/10.1196/annals.1440.011

Burnham, T., McCabe, K., & Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization*, 43(1), 57–73. http://dx.doi.org/10.1016/s0167–2681(00)00108-6

Burt, R. S., & Knez, M. (1995). Kinds of third-party effects on trust. *Rationality and Society*, 7(3), 255–292.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, 7(5), 225–231. http://dx.doi.org/10.1016/S1364–6613(03)00094-9

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. http://dx.doi.org/10.1016/j.cogpsych.2010.03.001

Chaudhuri, A., & Gangadharan, L. (2007). An experimental analysis of trust and trustworthiness. *Southern Economic Journal*, 73(4), 959–985. https://doi.org/10.2307/20111937

Chaudhuri, A., Sopher, B., & Strand, P. (2002). Cooperation in social dilemmas, trust and reciprocity. *Journal of Economic Psychology*, 23(2), 231–249. http://dx.doi.org/10.1016/s0167–4870(02)00065-x

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models in social neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior* (1st ed., pp. 229–244). John Wiley & Sons, Ltd.

Coleman, J. (1990). *Foundations of social theory*. The Belknap Press of Harvard University.

Csukás, C., Fracalanza, P., Kovács, T., & Willinger, M. (2008). The determinants of trusting and reciprocal behaviour: Evidence from an intercultural experiment. *Journal of Economic Development*, 33(1), 71–95. http://dx.doi.org/10.35866/caujed.2008.33.1.004

Das, T. K., & Teng, B.-S. (1998). Between trust and control: Developing confidence in partner cooperation in alliances. *Academy of Management Review*, 23(3), 491–512. http://dx.doi.org/10.5465/amr.1998.926623

(2001). Trust, control, and risk in strategic alliances: An integrated framework. *Organization Studies*, 22(2), 251–283. http://dx.doi.org/10.1177/0170840601222004

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. http://dx.doi.org/10.1038/nn1575

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, 2(4), 265–279.

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, 12(4), 618–634. http://dx.doi.org/10.1093/scan/nsw171

Dosenbach, N. U., Fair, D. A., Miezin, F. M., et al. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 104(26), 11073–11078. http://dx.doi.org/10.1073/pnas.0704320104

Dosenbach, N. U., Nardos, B., Cohen, A. L., et al. (2010). Prediction of individual brain maturity using fMRI. *Science*, 329(5997), 1358–1361. http://dx.doi.org/10.1126/science.1194144

Dreher, J. C., Kohn, P., & Berman, K. F. (2006). Neural coding of distinct statistical properties of reward information in humans. *Cerebral Cortex*, 16(4), 561–573. http://dx.doi.org/10.1093/cercor/bhj004

Dunning, D., Anderson, J. E., Schlosser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology*, 107(1), 122–141. http://dx.doi.org/10.1037/a0036673

Dunning, D., Fetchenhauer, D., & Schlösser, T. (2019). Why people trust: Solved puzzles and open mysteries. *Current Directions in Psychological Science*, 28(4), 366–371. http://dx.doi.org/10.1177/0963721419838255

Ekerdt, C. E. M., Kuhn, C., Anwander, A., Brauer, J., & Friederici, A. D. (2020). Word learning reveals white matter plasticity in preschool children. *Brain Structure and Function*, 225(2), 607–619. http://dx.doi.org/10.1007/s00429–020-02024-7

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, 145(6), 772–787. http://dx.doi.org/10.1037/xge0000165

Fairley, K., Sanfey, A. G., Vyrastekova, J., & Weitzel, U. (2016). Trust and risk revisited. *Journal of Economic Psychology*, 57, 74–85. http://dx.doi.org/10.1016/j.joep.2016.10.001

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, Article 148. http://dx.doi.org/10.3389/fnins.2012.00148

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425 (6960), 785–791. http://dx.doi.org/10.1038/nature02043

(2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8 (4), 185–190. http://dx.doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868. http://dx.doi.org/10.1162/003355399556151

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology*, 30(3), 263–276. http://dx.doi.org/10.1016/j.joep.2008.04.006

Finkl, T., Hahne, A., Friederici, A. D., Gerber, J., Murbe, D., & Anwander, A. (2020). Language without speech: Segregating distinct circuits in the human brain. *Cerebral Cortex*, 30(2), 812–823. http://dx.doi.org/10.1093/cercor/bhz128

Fletcher, P. C., Happe, F., Frith, U., et al. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, 57(2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience*, 33(8), 3602–3611. http://dx.doi.org/10.1523/Jneurosci.3086-12.2013

Frost, T., Stimpson, D. V., & Maughan, M. R. (1978). Some correlates of trust. *Journal of Psychology*, 99(1st Half), 103–108. http://dx.doi.org/10.1080/00223980.1978.9921447

Galton, F. (1907). Vox populi. *Nature*, 75(1949), 450–451. http://dx.doi.org/10.1038/075450a0

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68, 101–128. http://dx.doi.org/10.1146/annurev-psych-122414-033625

Gordon, R., & Spears, K. (2012). You don't act like you trust me: Dissociations between behavioural and explicit measures of source credibility judgement. *Quarterly Journal of Experimental Psychology*, 65(1), 121–134. http://dx.doi.org/10.1080/17470218.2011.591534

Gottfried, J. A., & Dolan, R. J. (2004). Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. *Nature Neuroscience*, 7(10), 1144–1152. http://dx.doi.org/10.1038/nn1314

Grèzes, J., Fonlupt, P., Bertenthal, B., Delon-Martin, C., Segebarth, C., & Decety, J. (2001). Does perception of biological motion rely on specific brain regions? *NeuroImage*, 13(5), 775–785. http://dx.doi.org/10.1006/nimg.2000.0740

Gromann, P. M., Heslenfeld, D. J., Fett, A. K., Joyce, D. W., Shergill, S. S., & Krabbendam, L. (2013). Trust versus paranoia: Abnormal response to social reward in psychotic illness. *Brain*, 136(Pt 6), 1968–1975. http://dx.doi.org/10.1093/brain/awt076

Gromann, P. M., Shergill, S. S., de Haan, L., et al. (2014). Reduced brain reward response during cooperation in first-degree relatives of patients with psychosis: An fMRI study. *Psychological Medicine*, 44(16), 3445–3454. http://dx.doi.org/10.1017/S0033291714000737

Hahn, T., Notebaert, K., Anderl, C., Reicherts, P., et al. (2015). Reliance on functional resting-state network for stable task control predicts behavioral tendency for cooperation. *NeuroImage*, 118, 231–236. http://dx.doi.org/10.1016/j.neuroimage.2015.05.093

Hahn, T., Notebaert, K., Anderl, C., Teckentrup, V., Kassecker, A., & Windmann, S. (2015). How to trust a perfect stranger: Predicting initial trust behavior from resting-state brain-electrical connectivity. *Social Cognitive and Affective Neuroscience*, 10(6), 809–813. http://dx.doi.org/10.1093/scan/nsu122

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(18), 6741–6746. http://dx.doi.org/10.1073/pnas.0711099105

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, 8(1), Article 2191. http://dx.doi.org/10.1038/s41467–017-02314-5

Hillebrandt, H., Sebastian, C., & Blakemore, S. J. (2011). Experimentally induced social inclusion influences behavior on trust games. *Cognitive Neuroscience*, 2(1), 27–33. http://dx.doi.org/10.1080/17588928.2010.515020

Ho, T. H., & Weigelt, K. (2005). Trust building among strangers. *Management Science*, 51(4), 519–530. http://dx.doi.org/10.1287/mnsc.1040.0350

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1(4), 304–309. http://dx.doi.org/10.1038/1124

Howard, J. D., & Kahnt, T. (2018). Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex. *Nature Communications*, 9(1), Article 1611. http://dx.doi.org/10.1038/s41467–018-04055-5

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte Carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, 11(6), Article e1004254. http://dx.doi.org/10.1371/journal.pcbi.1004254

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLoS Computational Biology*, 14(2), Article e1005935. http://dx.doi.org/10.1371/journal.pcbi.1005935

Ingvar, D. H. (1974). *Patterns of brain activity revealed by measurements of regional cerebral blood flow*. Paper presented at the Alfred Benzon Symposium VIII, Copenhagen.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *NPJ Science of Learning*, 2, Article 8. http://dx.doi.org/10.1038/s41539–017-0009-2

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.

Khalvati, K., Park, S. A., Mirbagheri, S., et al. (2019). Modeling other minds: Bayesian inference explains human choices in group decision making. *Science Advances*, 5(11), Article eaax8783. http://dx.doi.org/10.1126/sciadv.aax8783

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. http://dx.doi.org/10.1126/science.1156902

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. http://dx.doi.org/10.1126/science.1108062

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, 161, 9–18. http://dx.doi.org/10.1016/j.neuroimage.2017.08.026

Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1511), 3859–3874. http://dx.doi.org/10.1098/rstb.2008.0165

Krueger, F., McCabe, K., Moll, J., et al. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of*

America, 104(50), 20084–20089. http://dx.doi.org/10.1073/pnas.0710103104

Kunisato, Y., Okamoto, Y., Okada, G., et al. (2011). Personality traits and the amplitude of spontaneous low-frequency oscillations during resting state. *Neuroscience Letters*, 492(2), 109–113. http://dx.doi.org/10.1016/j.neulet.2011.01.067

Lamba, A., Frank, M. J., & FeldmanHall, O. (2020). Anxiety impedes adaptive social learning under uncertainty. *Psychological Science*, 31(5), 592–603. http://dx.doi.org/10.1177/0956797620910993

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 39(2), 329–358. http://dx.doi.org/10.1207/s15327906mbr3902_8

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, 63(4), 967–985. http://dx.doi.org/10.2307/2578601

Li, Y., Vanni-Mercier, G., Isnard, J., Mauguiere, F., & Dreher, J. C. (2016). The neural dynamics of reward value and risk coding in the human orbitofrontal cortex. *Brain*, 139(Pt 4), 1295–1309. http://dx.doi.org/10.1093/brain/awv409

Ligneul, R., Obeso, I., Ruff, C. C., & Dreher, J. C. (2016). Dynamical representation of dominance relationships in the human rostromedial prefrontal cortex. *Current Biology*, 26(23), 3107–3115. http://dx.doi.org/10.1016/j.cub.2016.09.015

Lis, S., Baer, N., Franzen, N., et al. (2016). Social interaction behavior in ADHD in adults in a virtual trust game. *Journal of Attention Disorders*, 20(4), 335–345. http://dx.doi.org/10.1177/1087054713482581

Lockwood, P. L., Apps, M. A., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America*, 113(35), 9763–9768. http://dx.doi.org/10.1073/pnas.1603198113

Lu, X., Li, T., Xia, Z., et al. (2019). Connectome-based model predicts individual differences in propensity to trust. *Human Brain Mapping*, 40(6), 1942–1954. http://dx.doi.org/10.1002/hbm.24503

Luhmann, N. (1979). Trust: A mechanism for the reduction of social complexity. In N. Luhmann (Ed.), *Trust and power* (pp. 4–103). Wiley.

Ma, I., Sanfey, A. G., & Ma, W. J. (2020). The social cost of gathering information for trust decisions. *Scientific Reports*, 10(1), Article 14073. http://dx.doi.org/10.1038/s41598–020-69766-6

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature Communications*, 9(1), Article 2474. http://dx.doi.org/10.1038/s41467–018-04925-y

Malhotra, D. (2004). Trust and reciprocity decisions: The differing perspectives of trustors and trusted parties. *Organizational Behavior and Human Decision Processes,* 94(2), 61–73. http://dx.doi.org/10.1016/j.obhdp.2004.03.001

Malhotra, D., & Murnighan, J. K. (2002). The effects of contracts on interpersonal trust. *Administrative Science Quarterly*, 47(3), 534–559. http://dx.doi.org/10.2307/3094850

Maurer, C., Chambon, V., Bourgeois-Gironde, S., Leboyer, M., & Zalla, T. (2018). The influence of prior reputation and reciprocity on dynamic trust-building in adults with and without autism spectrum disorder. *Cognition*, 172, 1–10. http://dx.doi.org/10.1016/j.cognition.2017.11.007

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, 20(3), 709–734.

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, 52 (2), 267–275. http://dx.doi.org/10.1016/s0167–2681(03)00003-9

McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin*, 6(3), 467–472. http://dx.doi.org/10.1177/014616728063023

Mellers, B., Ungar, L., Baron, J., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106–1115. http://dx.doi.org/10.1177/0956797614524255

Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PLoS One*, 7(11), Article e49748. http://dx.doi.org/10.1371/journal.pone.0049748

Metereau, E., & Dreher, J. C. (2015). The medial orbitofrontal cortex encodes a general unsigned value signal during anticipation of both appetitive and aversive events. *Cortex*, 63, 42–54. http://dx.doi.org/10.1016/j.cortex.2014.08.012

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, 16(5), 1936–1947. http://dx.doi.org/10.1523/jneurosci.16-05-01936.1996

Na, S., Chung, D., Hula, A., et al. (2019). Humans use forward thinking to exert social control. *bioRxiv*. http://dx.doi.org/10.1101/737353

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669), 452–454. http://dx.doi.org/10.1126/science.1094285

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia*, 41(2), 147–155. https://doi.org/10.1016/S0028–3932(02)00145-8

Park, S. A., Sestito, M., Boorman, E. D., & Dreher, J. C. (2019). Neural computations underlying strategic social decision making in groups. *Nature Communications*, 10(1), Article 5287. http://dx.doi.org/10.1038/s41467–019-12937-5

Pegors, T. K., Kable, J. W., Chatterjee, A., & Epstein, R. A. (2015). Common and unique representations in pFC for face and place attractiveness. *Journal of Cognitive Neuroscience*, 27(5), 959–973. http://dx.doi.org/10.1162/jocn_a_00777

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 of the United States of America*, 107(29), 13099–13104. http://dx.doi.org/10.1073/pnas.1008137107

Pillutla, M. M., Malhotra, D., & Keith Murnighan, J. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology*, 39(5), 448–455. http://dx.doi.org/10.1016/s0022–1031(03)00015-5

Platt, M. L., & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under uncertainty. *Nature Neuroscience*, 11(4), 398–403. http://dx.doi.org/10.1038/nn2062

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, 51(3), 381–390. http://dx.doi.org/10.1016/j.neuron.2006.06.024

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2), 816–847. http://dx.doi.org/10.1016/j.neuroimage.2012.04.062

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2), 676–682. http://dx.doi.org/10.1073/pnas.98.2.676

Reichelt, J., Sievert, J., & Jacob, F. (2013). How credibility affects eWOM reading: The influences of expertise, trustworthiness, and similarity on utilitarian and social functions. *Journal of Marketing Communications*, 20 (1–2), 65–81. http://dx.doi.org/10.1080/13527266.2013.797758

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current Research and Theory*, 64–99, Appleton-Century-Crofts.

Rosenberg, M. D., Finn, E. S., Scheinost, D., et al. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, 19(1), 165–171. http://dx.doi.org/10.1038/nn.4179

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, 50(1), 131–142. http://dx.doi.org/10.1037/0022-3514.50.1.131

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*, 23(3), 393–404. http://dx.doi.org/10.5465/amr.1998.926617

Rudebeck, P. H., & Murray, E. A. (2014). The orbitofrontal oracle: Cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. *Neuron*, 84(6), 1143–1156. http://dx.doi.org/10.1016/j.neuron.2014.10.049

Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S., & Murray, E. A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation

and value updating. *Nature Neuroscience*, 16(8), 1140–1145. http://dx.doi.org/10.1038/nn.3440

Sánchez-Pagés, S., & Vorsatz, M. (2007). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, 61(1), 86–112. http://dx.doi.org/10.1016/j.geb.2006.10.014

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, 19 (4), 1835–1842. http://dx.doi.org/10.1016/s1053–8119(03)00230-1

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, 1(3), 199–207. http://dx.doi.org/10.1038/35044563

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. http://dx.doi.org/10.1126/science.275.5306.1593

Siegel, J. Z., Estrada, S., Crockett, M. J., & Baskin-Sommers, A. (2019). Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature Communications*, 10(1), Article 1942. http://dx.doi.org/10.1038/s41467–019-09962-9

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, 2(10), 750–756. http://dx.doi.org/10.1038/s41562–018-0425-1

Sjöberg, L. (2009). Are all crowds equally wise? A comparison of political election forecasts by experts and the public. *Journal of Forecasting*, 28(1), 1–18. http://dx.doi.org/10.1002/for.1083

Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis*, 13(6), 675–682. http://dx.doi.org/10.1111/j.1539-6924.1993.tb01329.x

Snijders, C., & Keren, G. (2001). Do you trust? Whom do you trust? When do you trust? *Advances in Group Processes*, 18, 129–160. http://dx.doi.org/10.1016/S0882–6145(01)18006-9

Sperber, D. A. N., Clément, F., Heintz, C., et al. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393. http://dx.doi.org/10.1111/j.1468-0017.2010.01394.x

Sripada, C. S., Angstadt, M., Banks, S., Nathan, P. J., Liberzon, I., & Phan, K. L. (2009). Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *Neuroreport*, 20(11), 984–989. http://dx.doi.org/10.1097/WNR.0b013e32832d0a67

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, 21(3), 349–354. http://dx.doi.org/10.1177/0956797610362647

Strickland, L. H. (1958). Surveillance and trust. *Journal of Personality*, 26(2), 200–215. http://dx.doi.org/10.1111/j.1467-6494.1958.tb01580.x

Thielmann, I., & Hilbig, B. E. (2015). The traits one can trust: Dissecting reciprocity and kindness as determinants of trustworthy behavior. *Personality and Social Psychology Bulletin*, 41(11), 1523–1536. http://dx.doi.org/10.1177/0146167215600530

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, 27(6), 813–833. http://dx.doi.org/10.1521/soco.2009.27.6.813

Toelch, U., Bach, D. R., & Dolan, R. J. (2014). The neural underpinnings of an optimal exploitation of social information under uncertainty. *Social Cognitive and Affective Neuroscience*, 9(11), 1746–1753. http://dx.doi.org/10.1093/scan/nst173

Tsuchida, A., Doll, B. B., & Fellows, L. K. (2010). Beyond reversal: A critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. *Journal of Neuroscience*, 30(50), 16868–16875. http://dx.doi.org/10.1523/JNEUROSCI.1958-10.2010

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323. http://dx.doi.org/10.1007/Bf00122574

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3), 829–858. http://dx.doi.org/10.1002/hbm.20547

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, 48(3), 564–584. http://dx.doi.org/10.1016/j.neuroimage.2009.06.009

Vanni-Mercier, G., Mauguiere, F., Isnard, J., & Dreher, J. C. (2009). The hippocampus codes the uncertainty of cue-outcome associations: An intracranial electrophysiological study in humans. *Journal of Neuroscience*, 29(16), 5287–5294. http://dx.doi.org/10.1523/JNEUROSCI.5298-08.2009

Wardle, M. C., Fitzgerald, D. A., Angstadt, M., Sripada, C. S., McCabe, K., & Phan, K. L. (2013). The caudate signals bad reputation during trust decisions. *PLoS One*, 8(6), Article e68884. http://dx.doi.org/10.1371/journal.pone.0068884

Weiner, J. L., & Mowen, J. C. (1986). Source credibility: On the independent effects of trust and expertise. *Advances in Consumer Research*, 13, 306–310.

Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, 59(2), 189–202. http://dx.doi.org/10.1177/106591290605900202

Winston, J. S., O'Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems for assessing facial attractiveness. *Neuropsychologia*, 45(1), 195–206. http://dx.doi.org/10.1016/j.neuropsychologia.2006.05.009

Wu, H., Liu, X., Hagan, C. C., & Mobbs, D. (2020). Mentalizing during social interaction: A four component model. *Cortex*, 126, 242–252. http://dx.doi.org/10.1016/j.cortex.2019.12.031

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology*, 8(12), Article e1002841. http://dx.doi.org/10.1371/journal.pcbi.1002841

Yaniv, I. (2006). The benefit of additional opinions. *Current Directions in Psychological Science*, 13(2), 75–78. http://dx.doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260–281. http://dx.doi.org/10.1006/obhd.2000.2909

Zatorre, R. J. (2013). Predispositions and plasticity in music and speech learning: Neural correlates and implications. *Science*, 342(6158), 585–589. http://dx.doi.org/10.1126/science.1238414