frontiers in Human Neuroscience

MINI REVIEW
published: 13 August 2019
doi: 10.3389/fnhum.2019.00271

Neurobehavioral Mechanisms Supporting Trust and Reciprocity

*Dominic S. Fareri**

Gordon F. Derner School of Psychology, Adelphi University, Garden City, NY, United States

Trust and reciprocity are cornerstones of human nature, both at the levels of close interpersonal relationships and economic/societal structures. Being able to both place trust in others and decide whether to reciprocate trust placed in us is rooted in implicit and explicit processes that guide expectations of others, help reduce social uncertainty, and build relationships. This review will highlight neurobehavioral mechanisms supporting trust and reciprocity, through the lens of implicit and explicit social appraisal and learning processes. Significant consideration will be given to the neural underpinnings of these implicit and explicit processes, and special focus will center on the underlying neurocomputational mechanisms facilitating the integration of implicit and explicit signals supporting trust and reciprocity. Finally, this review will conclude with a discussion of how we can leverage findings regarding the neurobehavioral mechanisms supporting trust and reciprocity to better inform our understanding of mental health disorders characterized by social dysfunction.

Keywords: trust, social learning, computational modeling, reciprocity, social decision-making

OPEN ACCESS

Edited by:
Frank Krueger,
George Mason University,
United States

Reviewed by:
Hongbo Yu,
Yale University, United States
Brent L. Hughes,
University of California, Riverside,
United States

***Correspondence:**
Dominic S. Fareri
dfareri@adelphi.edu

Received: *01 March 2019*
Accepted: *22 July 2019*
Published: *13 August 2019*

Citation:
Fareri DS (2019) Neurobehavioral Mechanisms Supporting Trust and Reciprocity. Front. Hum. Neurosci. 13:271. doi: 10.3389/fnhum.2019.00271

INTRODUCTION

Virtually all aspects of human social life are based on trust and reciprocity. Trust is a multifaceted, socially risky construct (Krueger and Meyer-Lindenberg, 2019) that not only underlies the success of our business and economic structures but is also a pillar on which close relationships and social networks are built. We can conceptualize trust as a process based on *social expectations* (Cox, 2004), whereby we assume mutual risk with another person—e.g., business colleague, close friend—in collaboration toward a shared goal (Simpson, 2007); reciprocity (or betrayal) of trust, then, provides feedback that guides social learning (Fareri et al., in press). Continued reciprocity from others can be socially rewarding, and can fulfill basic social needs of support and belongingness (Baumeister and Leary, 1995), without which we are likely to experience poor physical, emotional, and mental health outcomes (Cacioppo et al., 2015; Eisenberger et al., 2017).

Decisions to place trust in others and reciprocate trust placed in us are rooted in both *implicit* social appraisals—i.e., rapid evaluations of others based on minimal information—and *explicit* social learning processes resulting from direct interpersonal experience or presentation of social information. Both types of processes (summarized in **Table 1**) facilitate learning about others and reduction of social uncertainty (FeldmanHall and Chang, 2018; Fareri et al., in press). This article will review evidence regarding how decisions

TABLE 1 | Implicit and explicit signals guiding trust and reciprocity.

	Trust	Reciprocity
Implicit	Facial characteristics	Prosocial orientation
	Race bias	Personality traits
	In/out-group bias	Pupil dilation
	Prosocial orientation	
Explicit	Experienced reciprocity	Bestowed trust
	Experienced trust violation	Risk associated with trust
	Moral character	Prior experience of reciprocity
	Reputational priors	Threat of punishment

to trust and reciprocate are informed by implicit and explicit processes, how updating social expectations may be influenced by social priors, and how social and reward-related neural circuits support trust and reciprocity. This article also discusses the utility of computational accounts in characterizing implicit and explicit influences on trust and reciprocity, which may have important implications for mental health conditions characterized by dysregulated social function.

DECISIONS TO TRUST OTHERS

Deciding to place trust in someone represents one component of prosocial behavior. Recent theories suggest that prosociality is an automatic, intuitive process (Zaki and Mitchell, 2013) driven by the likelihood of success it brings in day-to-day life (Rand et al., 2016). Thus on its own, placing trust in another might reflect an adaptive strategy to ensure positive social outcomes by broadcasting to other people that we have a reputation for being willing and reliable interaction partners (Berg et al., 1995; Jordan et al., 2016).

Implicit Influences on Trust Decisions

Prosocial intuitions to trust others can be shaped by a number of implicit processes. Drawing on basic evolutionary threat detection mechanisms (Delgado et al., 2006; Lindström and Olsson, 2015), we are capable of estimating the trustworthiness of a stranger almost automatically (i.e., on the order of milliseconds), from relatively minimal perceptual information such as the precise configuration of an individual's facial features (Todorov, 2008), as well as from assumed social group knowledge. These rapid social evaluations implicitly guide initial *social approach/avoidance* decisions (i.e., should I engage with this new person) in that they occur outside of conscious awareness (Willis and Todorov, 2006). Evidence from fMRI studies implicate the amygdala (Engell et al., 2007; Todorov et al., 2008), as well as the medial prefrontal cortex (mPFC) and the precuneus (Todorov et al., 2008), all of which have been implicated in representing different forms of social information (Amodio and Frith, 2006; Van Overwalle and Baetens, 2009; Stanley, 2016), in rapid evaluations of facial trustworthiness. These regions differentially encode trustworthiness information: separate amygdala subregions show both quadratic and negative linear associations with facial trustworthiness (Todorov et al., 2008; Rule et al., 2013; Freeman et al., 2014), while the mPFC and precuneus demonstrate heightened responses to moderately trustworthy faces (Todorov et al., 2008).

Social approach and avoidance signals are themselves subjects to implicit influence. Social heuristics (e.g., race/gender stereotypes) can shape judgments of others and decisions to trust them outside of conscious awareness. Racial bias can be particularly pervasive: an individual's implicit bias (IAT; Greenwald et al., 1998) towards white partners (and away from black partners) positively correlates with the degree to which one invests with a white (vs. black) partner (Stanley et al., 2011). Interestingly, this pattern correlates with striatal activation, implicating this region in the implicit encoding of group-level reputation, while the degree to which people invest with black vs. white partners correlates positively with amygdala activation (Stanley et al., 2012), possibly suggesting an implicit representation of perceived threat or social risk.

Implicit influences on trust decisions can more generally be shaped by in-group vs. out-group biases that may be rooted in political affiliations (Rigney et al., 2018) or support for rival sports teams (Cikara et al., 2011), for instance. People tend to trust individuals who attend their university or are from the same country relative to those from rival universities or other countries (Hughes et al., 2017a,b). Interestingly, as with race, the difference in striatal activation when trusting in-group relative to out-group members significantly correlates with the degree of one's in-group bias (Hughes et al., 2017a). This effect may be mitigated by the amount of time one has to process trusting an outgroup member, suggesting that overcoming outgroup biases may require more deliberative thought and neural mechanisms of cognitive control (Hughes et al., 2017a). Taken together, implicit signals can significantly and quickly shape choices to place trust in others.

Explicit Influences on Trust Decisions

Our choices to trust others can also be guided by particularly diagnostic information about another person (Bhanji and Beer, 2013; Mende-Siedlecki et al., 2013), which we can acquire explicitly through direct experiences with others or from another source (e.g., rumor spread by another person). Initial choices to place trust in others are subsequently met with either reciprocity or a violation of trust. This direct experience of a social outcome (i.e., positive or negative) can then inform our representation of a partner's reputation and whether we want to trust them in the future. As is the case with implicit representation of group-level reputation, the striatum encodes the reputation of individuals stronger responses in the striatum are elicited by experienced reciprocity (vs. violations) of trust during repeated interactions, and these neural signals shift temporally backward as we learn to anticipate that a partner will act in a trustworthy manner (King-Casas et al., 2005). Similar patterns of activation have been reported in the mPFC, with enhanced BOLD activation observed when trusting others during initial stages of partnership building that decreases once reputation has been learned (Krueger et al., 2007). Thus, direct experience of reciprocity serves as both a socially rewarding commodity (Phan et al., 2010) that fluctuates based on patterns of cooperation (Rilling et al., 2002; Phan et al., 2010) and an explicit social *learning* signal that can guide trust decisions.

Information about moral character—which can be inferred from how likely one would be to consider another's welfare (i.e., deontological) relative to the bottom line outcome (i.e., consequentialist)—may be particularly diagnostic when deciding whether to trust a partner (Everett et al., 2016). Individuals endorsing deontological choices (i.e., would not endorse killing one person to save five) are consistently seen as more moral and trustworthy and are trusted more often in one-shot trust games (Everett et al., 2016, 2018). Information about moral character can create a persistent and outsized influence on our ability to encode explicit signals of reciprocation and defection of trust from another person. Learning that someone performed a selfless deed (e.g., risking their life for another person) can facilitate impressions (i.e., prior) of that person as highly moral and trustworthy, even if they happen to violate our trust at a later point, a process resembling confirmation bias (Doll et al., 2009). Strong moral impressions also modulate striatal function during repeated social interactions based on trust such that they eliminate the canonical social reward response in the striatum to reciprocity relative to defection, inhibiting learning *via* explicit experience (Delgado et al., 2005). Reputational priors may be encoded within the mPFC, as this region demonstrates increased activation when faced with a choice to trust a partner about whom priors exist (relative to those about whom we have no knowledge; Fouragnan et al., 2013). Thus, explicitly acquired social information can both incrementally shape our ability to learn to trust others while also inhibiting our ability to adapt to social interactions.

Neurocomputational Mechanisms Supporting Trust Decisions

Implicit and explicit signals thus both play a role in decisions to trust. Increases in the use of computational modeling approaches (Kishida and Montague, 2012; Cheong et al., 2017) may shed light on the interaction between these different types of signals. One hypothesis suggests that a choice to place trust in another is not static, but rather dynamically evolves over time as we update initial implicit appraisals of others with explicitly experienced patterns of reciprocity (Chang et al., 2010) using associative mechanisms that enable learning the value of a partner on a trial-by-trial basis (Rescorla and Wagner, 1972). Importantly, this *dynamic belief* model of trustworthiness allows for: (1) differential weighting of reciprocity (or defection) as a function of the strength and valence of an initial impression of a partner; and (2) for initial impressions and experienced outcomes to influence each other in order to learn about a partner. In other words, whether a partner reciprocates or violates trust informs the updating of an impression, which then feeds forward to differentially weight subsequent instances of reciprocity/violation of trust (Chang et al., 2010).

Computational modeling also highlights different ways in which priors shape trust decisions. Learning to trust partners can be better explained by a model assuming that we learn differently about others (relative to a model assuming no social biases) based on whether we have priors about their tendency to cooperate (Fouragnan et al., 2013). Further, striatal representation of prediction error signals (i.e., increased activation when expectations do not match outcomes) is *absent* for those partners about whom instructed priors exist (Fouragnan et al., 2013), suggesting that priors shape behavior *via* top-down neural mechanisms. Other work demonstrates that people tend to weight and use reciprocity and violations of trust (as indexed by different learning rates) in ways that are consistent with prior impressions of others as "good" or "bad": reciprocity from someone initially perceived as trustworthy contributes more heavily to subsequent choices to a violation of trust from that same partner, and *vice versa* (Fareri et al., 2012). Computational approaches have also tested competing hypotheses about whether trust decisions are motivated differently as a function of whether we have an existing relationship with a partner (Fareri et al., 2015). Modeling analyses revealed that choices to trust friends relative to strangers are driven not by stronger priors associated with a friend, but rather by differential weighting of experienced reciprocity as a function of social closeness with a partner (i.e., greater social value of reciprocation from a friend than from a stranger); this social weighting is encoded within the ventral striatum and mPFC (Fareri et al., 2015). Computational approaches thus demonstrate that implicit information (i.e., facial characteristics) and explicit information (i.e., social priors, relationship closeness) shape the way in which we use experienced reciprocity to inform choices to trust. It is worth noting that these computational processes allow for the possibility of explicitly acquired social information (i.e., moral information) to act implicitly in guiding neural and behavioral responses.

DECISIONS TO RECIPROCATE TRUST

Whereas a choice to place trust in another person involves approach/avoidance mechanisms and a desire to signal a willingness to be cooperative, reciprocity involves higher-level considerations in conjunction with implicit appraisal processes. Reciprocity is necessarily more informed by individual differences in other-regarding preferences, more explicit person-level information—i.e., did this person place their trust in me?—and consideration of how others will react to our actions.

Implicit Influences on Reciprocity

Reciprocity can be driven in part by inferences based on physiological signals from those that bestow trust upon us and from trait-level individual differences (Thielmann and Hilbig, 2015). For example, much like information from the eyes can inform us about potential threats in the environment (Kim et al., 2016), we can also use signals from the eyes to guide our choices to reciprocate trust. People tend to reciprocate trust from others who display more dilated pupils (Kret and De Dreu, 2019); increases in pupil dilation may reflect a desire for affiliation (though see also Fehr and Schneider, 2010). People may also be guided to reciprocate based on their *own tendencies* toward being prosocial (vs. self-interested), which may be quantified as the trade-off in value between outcomes for the self vs. outcomes for others (i.e., social value orientation; McClintock and Allison, 1989; Van Lange, 1999). Prosocial orientation positively correlates

with amygdala activation when evaluating distributions of reward outcomes between self and other (Haruno and Fridth, 2009), suggesting this to be an implicit, internal process that may guide future choices in social interactions. People who tend to be more prosocial demonstrate greater levels of reciprocity overall and are more sensitive to a partner's pattern of cooperation (Van Lange, 1999; van den Bos et al., 2009). These trait level differences tend to be reflected in the recruitment of neural circuits supporting social processes and cognitive control, such that acting contradictory to one's typical orientation engages activation in the right temporoparietal junction (rTPJ), precuneus and dorsal anterior cingulate cortex (dACC; van den Bos et al., 2009).

Explicit Influences on Reciprocity

Decisions to reciprocate trust necessarily involve evaluation of explicit social signals. The degree to which an interaction partner places trust in us informs whether we should feel inclined to reciprocate and whether we can expect that they will act similarly in the future. Bestowed trust may be perceived as a social reward signal that indicates something about reputation and shapes our own propensity to reciprocate. The striatum is implicated in encoding this type of explicit signal, and the degree to which these instances of trust are diagnostic of another person's reputation is associated with temporal shifts in the striatal response to anticipating a partner's decisions' (King-Casas et al., 2005). Importantly, these choices to reciprocate are subject to contextual information. Knowledge of the risk another person may be taking by placing trust in us can shape our choices to engage in reciprocity. People tend to reciprocate more often when someone is taking a large chance of incurring a loss by trusting us; reciprocity in such high-risk situations is associated with increases in rTPJ recruitment (van den Bos et al., 2009). Choices to reciprocate are also positively correlated with the degree to which people have experienced reciprocity from others in the past (Cáceda et al., 2017). Furthermore, if explicitly threatened with a penalty (i.e., sanction) for not reciprocating trust, people tend to reciprocate to a lesser degree (associated with anticipatory activation in the vmPFC) suggesting an aversive effect of explicit threats in reciprocity and relationship building (Li et al., 2009).

Neurocomputational Mechanisms Supporting Reciprocity Decisions

One hypothesis emerging from the idea that sensitivity to others' outcomes can drive reciprocity is that people are inequity averse, and try to remedy inequitable outcomes or distributions of resources (Fehr and Schmidt, 1999; Tricomi et al., 2010). A related, but competing hypothesis regarding reciprocity posits that we often act to minimize feelings associated with disappointing another person by not meeting their expectations of us. This phenomenon, known as guilt aversion (Dufwenberg and Gneezy, 2000; Battigalli and Dufwenberg, 2007), requires considering not just our own intentions and interests, but also the assumed expectations that a partner has about our own behavior (i.e., second-order beliefs: our estimation of the likelihood someone thinks that we will reciprocate their trust). Computational approaches have proven to be quite useful in parameterizing motivations for reciprocity such as inequity and guilt aversion. As many studies examining these processes are structured through the lens of economic interactions (i.e., trust game), guilt on the part of a trustee has been conceptualized as the *difference* between: (1) the monetary amount that a trustee thinks that an investor would expect them to send back (i.e., second-order belief); and (2) the amount that the trustee actually sends back. This difference is weighted by a guilt aversion parameter which indexes sensitivity to feeling guilt (Chang et al., 2011). People tend to use their second-order beliefs to guide reciprocity decisions, with greater guilt sensitivity associated with increased recruitment of regions supporting social, emotional, and cognitive control processes (i.e., TPJ, insula, dACC, dorsolateral PFC; Chang et al., 2011). Inequity, on the other hand, can be parameterized as a more general sensitivity to unequal distributions of resources, regardless of whom is affected (i.e., absolute difference between outcomes for self vs. other). Reciprocity decisions motivated by inequity aversion seem to implicate value-based regions such as the amygdala and ventral striatum (Nihonsugi et al., 2015). Interestingly, people may opportunistically alternate between reciprocation based on inequity aversion or guilt aversion by also accounting for one's own self-interest outcome. This moral strategy model accounts for such a possibility by including a social preference parameter, which indicates that people are acting out of self-interest when it is equivalent to 0, but are motivated by guilt aversion or inequity aversion when this term is negative or positive, respectively (van Baar et al., 2019). Thus, the application of computational models has helped to dissociate different mechanisms underlying reciprocity decisions.

CLINICAL IMPLICATIONS AND FUTURE DIRECTIONS

Interpersonal difficulties centered on trust and reciprocity are at the heart of a host of mental health conditions (Montague et al., 2012; Kishida and Montague, 2013; Stanley and Adolphs, 2013). It is thus useful to consider how breakdowns of implicit and explicit processes supporting such decisions may contribute to difficulties in social function. Borderline personality disorder (BPD) represents a condition characterized by unstable relationships and dysregulated affect (Stanley and Siever, 2010). Individuals with BPD tend to be biased towards suspicion of others, showing a greater likelihood than healthy controls to view others as untrustworthy, and reduced neural responses to untrustworthy faces in the insula and lateral PFC (Fertuck et al., 2019). Individuals with BPD are also unable to maintain reciprocity in repeated interactions over time (King-Casas et al., 2008), possibly driven by difficulty representing others' intentions (Stanley and Siever, 2010; Meyer-Lindenberg et al., 2011). Future work may incorporate computational approaches to probe whether failures to maintain reciprocity in BPD are related to altered computations of guilt aversion, or to the integration of beliefs about others' intentions as well as the

changes (i.e., volatility) of others' behavior (Behrens et al., 2008; Diaconescu et al., 2014; Siegel et al., 2018).

Difficulty with appraising trustworthiness and placing trust in others are hallmarks of post-traumatic stress disorder (PTSD), which may underlie the propensity for re-victimization in this population (Fertuck et al., 2016). For example, individuals with PTSD are more likely to appraise unfamiliar faces as trustworthy compared to healthy controls, suggesting an impaired ability to integrate trust-related facial signals (Todorov et al., 2008). Further, women with PTSD associated with prior sexual assault show significantly lower levels of investment over time with partners and a decreased ability to learn partner reputation (i.e., lower learning rates in a computational model) in an economic trust game compared with healthy controls (Cisler et al., 2015). The application of Bayesian learning approaches, which can capture the ability to flexibly and dynamically update representations of moral character (Siegel et al., 2018), may be useful in further elucidating aberrant social learning processes in PTSD.

While the literature reviewed here highlights the utility of computational approaches to understanding mechanisms (i.e., guilt aversion, moral opportunism, relationship value) involved in trust and reciprocity, our understanding of neural and behavioral mechanisms of trust and reciprocity will be bolstered by integration with advanced neuroimaging approaches. Recent efforts linking individual differences in computational strategies supporting reciprocity (i.e., guilt aversion vs. moral opportunism) with differential patterns of brain activity assessed by inter-subject representational similarity analysis (van Baar et al., 2019) provide one template for combining computational and advanced imaging techniques. Other neuroimaging advances involve the use of network-level resting-state and task-based connectivity approaches to characterizing neural dynamics (Smith et al., 2009, 2014; Utevsky et al., 2017). Combining network connectivity approaches with computational modeling of behavior can help characterize how communication within and between neural networks supporting social processes and decision-making (e.g., default mode network, executive control network) are involved in the computation of implicit and explicit signals supporting trust and reciprocity. Such work may provide deeper and differential insight into neurocomputational breakdowns in trust across mental health conditions.

AUTHOR CONTRIBUTIONS

DF wrote the manuscript.

FUNDING

This work was supported by an Adelphi University Faculty Development Grant to DF and internal funds.

ACKNOWLEDGMENTS

The author would like to thank Joanne Stasiak for helpful comments.

REFERENCES

Amodio, D. M., and Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 7, 268–277. doi: 10.1038/nrn1884

Battigalli, P., and Dufwenberg, M. (2007). Guilt in games. *Am. Econ. Rev.* 97, 170–176. doi: 10.1257/aer.97.2.170

Baumeister, R., and Leary, M. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* 117, 497–529. doi: 10.1037/0033-2909.117.3.497

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., and Rushworth, M. F. S. (2008). Associative learning of social value. *Nature* 456, 245–249. doi: 10.1038/nature07538

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav.* 10, 122–142. doi: 10.1006/game.1995.1027

Bhanji, J. P., and Beer, J. S. (2013). Dissociable neural modulation underlying lasting first impressions, changing your mind for the better and changing it for the worse. *J. Neurosci.* 33, 9337–9344. doi: 10.1523/JNEUROSCI.5634-12.2013

Cáceda, R., Prendes-Alvarez, S., Hsu, J.-J., Tripathi, S. P., Kilts, C. D., and James, G. A. (2017). The neural correlates of reciprocity are sensitive to prior experience of reciprocity. *Behav. Brain Res.* 332, 136–144. doi: 10.1016/j.bbr.2017.05.030

Cacioppo, J. T., Cacioppo, S., Capitanio, J. P., and Cole, S. W. (2015). The neuroendocrinology of social isolation. *Annu. Rev. Psychol.* 66, 733–767. doi: 10.1146/annurev-psych-010814-015240

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., and Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn. Psychol.* 61, 87–105. doi: 10.1016/j.cogpsych.2010.03.001

Chang, L. J., Smith, A., Dufwenberg, M., and Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* 70, 560–572. doi: 10.1016/j.neuron.2011.02.056

Cheong, J. H., Jolly, E., Sul, S., and Chang, L. J. (2017). "Computational models in social neuroscience," in *Computational Models of Brain and Behavior*, ed. A. A. Moustafa (Chichester: John Wiley & Sons, Ltd.), 229–244.

Cikara, M., Botvinick, M. M., and Fiske, S. T. (2011). Us versus them: social identity shapes neural responses to intergroup competition and harm. *Psychol. Sci.* 22, 306–313. doi: 10.1177/0956797610397667

Cisler, J. M., Bush, K., Steele, J. S., Lenow, J. K., Smitherman, S., and Kilts, C. D. (2015). Brain and behavioral evidence for altered social learning mechanisms among women with assault-related posttraumatic stress disorder. *J. Psychiatr. Res.* 63, 75–83. doi: 10.1016/j.jpsychires.2015.02.014

Cox, J. (2004). How to identify trust and reciprocity. *Games Econ. Behav.* 46, 260–281. doi: 10.1016/s0899-8256(03)00119-2

Delgado, M. R., Frank, R., and Phelps, E. A. (2005). Perceptions of moral character modulate the neural systmes of reward during the trust game. *Nat. Neurosci.* 8, 1611–1618. doi: 10.1038/nn1575

Delgado, M. R., Olsson, A., and Phelps, E. A. (2006). Extending animal models of fear conditioning to humans. *Biol. Psychol.* 73, 39–48. doi: 10.1016/j.biopsycho.2006.01.006

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., et al. (2014). Inferring on the intentions of others by hierarchical bayesian learning. *PLoS Comput. Biol.* 10:e1003810. doi: 10.1371/journal.pcbi.1003810

Doll, B. B., Jacobs, W. J., Sanfey, A. G., and Frank, M. J. (2009). Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Res.* 1299, 74–94. doi: 10.1016/j.brainres.2009.07.007

Dufwenberg, M., and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games Econ. Behav.* 30, 163–182. doi: 10.1006/game.1999.0715

Eisenberger, N. I., Moieni, M., Inagaki, T. K., Muscatell, K. A., and Irwin, M. R. (2017). In sickness and in health: the co-regulation of inflammation and social behavior. *Neuropsychopharmacology* 42, 242–253. doi: 10.1038/npp.2016.141

Engell, A. D., Haxby, J. V., and Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* 19, 1508–1519. doi: 10.1162/jocn.2007.19.9.1508

Everett, J. A. C., Faber, N. S., Savulescu, J., and Crockett, M. J. (2018). The costs of being consequentialist: social inference from instrumental harm and impartial beneficence. *J. Exp. Soc. Psychol.* 79, 200–216. doi: 10.1016/j.jesp.2018.07.004

Everett, J. A. C., Pizarro, D. A., and Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* 145, 772–787. doi: 10.1037/xge0000165

Fareri, D. S., Chang, L. J., and Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Front. Neurosci.* 6:148. doi: 10.3389/fnins.2012.00148

Fareri, D. S., Chang, L. J., and Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *J. Neurosci.* 35, 8170–8180. doi: 10.1523/JNEUROSCI.4775-14.2015

Fareri, D. S., Chang, L. J., and Delgado, M. R. (in press). "Neural mechanisms of social learning," in *The Cognitive Neurosciences*, eds M. S. Gazzaniga, G. R. Mangun and D. Poeppel (Cambridge, MA: MIT Press).

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868. doi: 10.1162/003355399556151

Fehr, E., and Schneider, F. (2010). Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity? *Proc. Biol. Sci.* 277, 1315–1323. doi: 10.1098/rspb.2009.1900

FeldmanHall, O., and Chang, L. J. (2018). "Social learning: emotions aid in optimizing goal-directed social behavior," in *Understanding Goal-Directed Decision-Making Computations and Circuits*, eds A. M. Bornstein and A. Shenhav. (London: Academic Press), 309–330.

Fertuck, E. A., Grinband, J., Mann, J. J., Hirsch, J., Ochsner, K., Pilkonis, P., et al. (2019). Trustworthiness appraisal deficits in borderline personality disorder are associated with prefrontal cortex, not amygdala, impairment. *Neuroimage Clin.* 21:101616. doi: 10.1016/j.nicl.2018.101616

Fertuck, E. A., Tsoi, F., Grinband, J., Ruglass, L., Melara, R., and Hien, D. A. (2016). Facial trustworthiness perception bias elevated in individuals with PTSD compared to trauma exposed controls. *Psychiatry Res.* 237, 43–48. doi: 10.1016/j.psychres.2016.01.056

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., and Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *J. Neurosci.* 33, 3602–3611. doi: 10.1523/JNEUROSCI.3086-12.2013

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., and Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. *J. Neurosci.* 34, 10573–10581. doi: 10.1523/JNEUROSCI.5063-13.2014

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74, 1464–1480. doi: 10.1037/0022-3514.74.6.1464

Haruno, M., and Fridth, C. D. (2009). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nat. Neurosci.* 13, 160–161. doi: 10.1038/nn.2468

Hughes, B. L., Ambady, N., and Zaki, J. (2017a). Trusting outgroup, but not ingroup members, requires control: neural and behavioral evidence. *Soc. Cogn. Affect. Neurosci.* 12, 372–381. doi: 10.1093/scan/nsw139

Hughes, B. L., Zaki, J., and Ambady, N. (2017b). Motivation alters impression formation and related neural systems. *Soc. Cogn. Affect. Neurosci.* 12, 49–60. doi: 10.1093/scan/nsw147

Jordan, J. J., Hoffman, M., Nowak, M. A., and Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proc. Natl. Acad. Sci. U S A* 113, 8658–8663. doi: 10.1073/pnas.1601280113

Kim, M. J., Solomon, K. M., Neta, M., Davis, C. F., Oler, J. A., Mazzula, E. C., et al. (2016). A face versus non-face context influences amygdala responses to masked fearful eye whites. *Soc. Cogn. Affect. Neurosci.* 11, 1933–1941. doi: 10.1093/scan/nsw110

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., and Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science* 321, 806–810. doi: 10.1126/science.1156902

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83. doi: 10.1126/science.1108062

Kishida, K. T., and Montague, P. R. (2012). Imaging models of valuation during social interaction in humans. *Biol. Psychiatry* 72, 93–100. doi: 10.1016/j.biopsych.2012.02.037

Kishida, K. T., and Montague, P. R. (2013). Economic probes of mental function and the extraction of computational phenotypes. *J. Econ. Behav. Organiz.* 94, 234–241. doi: 10.1016/j.jebo.2013.07.009

Kret, M. E., and De Dreu, C. K. W. (2019). The power of pupil size in establishing trust and reciprocity. *J. Exp. Psychol. Gen.* doi: 10.1037/xge0000508 [Epub ahead of print].

Krueger, F., and Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, mand economics. *Trends Neurosci.* 42, 92–101. doi: 10.1016/j.tins.2018.10.004

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., et al. (2007). Neural correlates of trust. *Proc. Natl. Acad. Sci. U S A* 104, 20084–20089. doi: 10.1073/pnas.0710103104

Li, J., Xiao, E., Houser, D., and Montague, P. R. (2009). Neural responses to sanction threats in two-party economic exchange. *Proc. Natl. Acad. Sci. U S A* 106, 16835–16840. doi: 10.1073/pnas.0908855106

Lindström, B., and Olsson, A. (2015). Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans. *J. Exp. Psychol. Gen.* 144, 688–703. doi: 10.1037/xge0000071

McClintock, C. G., and Allison, S. T. (1989). Social value orientation and helping behavior1. *J. Appl. Soc. Psychol.* 19, 353–362. doi: 10.1111/j.1559-1816.1989.tb00060.x

Mende-Siedlecki, P., Baron, S. G., and Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *J. Neurosci.* 33, 19406–19415. doi: 10.1523/JNEUROSCI.2334-13.2013

Meyer-Lindenberg, A., Domes, G., Kirsch, P., and Heinrichs, M. (2011). Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. *Nat. Rev. Neurosci.* 12, 524–538. doi: 10.1038/nrn3044

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018

Nihonsugi, T., Ihara, A., and Haruno, M. (2015). Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *J. Neurosci.* 35, 3412–3419. doi: 10.1523/JNEUROSCI.3885-14.2015

Phan, K. L., Sripada, C. S., Angstadt, M., and McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proc. Natl. Acad. Sci. U S A* 107, 13099–13104. doi: 10.1073/pnas.1008137107

Rand, D. G., Brescoll, V. L., Everett, J. A. C., Capraro, V., and Barcelo, H. (2016). Social heuristics and social roles: intuition favors altruism for women but not for men. *J. Exp. Psychol. Gen.* 145, 389–396. doi: 10.1037/xge0000154

Rescorla, R., and Wagner, A. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcment and non reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. Black and W. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Rigney, A. E., Koski, J. E., and Beer, J. S. (2018). The functional role of ventral anterior cingulate cortex in social evaluation: disentangling valence from subjectively rewarding opportunities. *Soc. Cogn. Affect. Neurosci.* 13, 14–21. doi: 10.1093/scan/nsx132

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. (2002). A neural basis for social cooperation. *Neuron* 35, 395–405. doi: 10.1016/s0896-6273(02)00755-9

Rule, N. O., Krendl, A. C., Ivcevic, Z., and Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: behavioral and neural correlates. *J. Pers. Soc. Psychol.* 104, 409–426. doi: 10.1037/a0031050

Siegel, J. Z., Mathys, C., Rutledge, R. B., and Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nat. Hum. Behav.* 2, 750–756. doi: 10.1038/s41562-018-0425-1

Simpson, J. A. (2007). Psychological foundations of trust. *Curr. Dir. Psychol. Sci.* 16, 264–268. doi: 10.1111/j.1467-8721.2007.00517.x

Smith, D. V., Utevsky, A. V., Bland, A. R., Clement, N., Clithero, J. A., Harsch, A. E. W., et al. (2014). Characterizing individual differences in functional connectivity using dual-regression and seed-based

approaches. *Neuroimage* 95, 1–12. doi: 10.1016/j.neuroimage.2014.03.042

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U S A* 106, 13040–13045. doi: 10.1073/pnas.0905267106

Stanley, B., and Siever, L. J. (2010). The interpersonal dimension of borderline personality disorder: toward a neuropeptide model. *Am. J. Psychiatry* 167, 24–39. doi: 10.1176/appi.ajp.2009.09050744

Stanley, D. A. (2016). Getting to know you: general and specific neural computations for learning about people. *Soc. Cogn. Affect. Neurosci.* 11, 525–536. doi: 10.1093/scan/nsv145

Stanley, D. A., and Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron* 80, 816–826. doi: 10.1016/j.neuron.2013.10.038

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., and Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl. Acad. Sci. U S A* 108, 7710–7715. doi: 10.1073/pnas.1014345108

Stanley, D. A., Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., et al. (2012). Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 744–753. doi: 10.1098/rstb.2011.0300

Thielmann, I., and Hilbig, B. (2015). The traits one can trust. *Pers. Soc. Psychol. Bull.* 41, 1523–1536. doi: 10.1177/0146167215600530

Todorov, A. (2008). Evaluating faces on trustworthiness: an extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Ann. N Y Acad. Sci.* 1124, 208–224. doi: 10.1196/annals.1440.012

Todorov, A., Baron, S. G., and Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Soc. Cogn. Affect. Neurosci.* 3, 119–127. doi: 10.1093/scan/nsn009

Tricomi, E., Rangel, A., Camerer, C. F., and O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature* 463, 1089–1091. doi: 10.1038/nature08785

Utevsky, A. V., Smith, D. V., Young, J. S., and Huettel, S. A. (2017). Large-scale network coupling with the fusiform cortex facilitates future social motivation. *eNeuro* 4:ENEURO.0084–17.2017–43. doi: 10.1523/eneuro.0084-17.2017

van Baar, J. M., Chang, L. J., and Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nat. Commun.* 10:1483. doi: 10.1038/s41467-019-09161-6

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A. R. B., and Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Soc. Cogn. Affect. Neurosci.* 4, 294–304. doi: 10.1093/scan/nsp009

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: an integrative model of social value orientation. *J. Pers. Soc. Psychol.* 77, 337–349. doi: 10.1037/0022-3514.77.2.337

Van Overwalle, F., and Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage* 48, 564–584. doi: 10.1016/j.neuroimage.2009.06.009

Willis, J., and Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* 17, 592–598. doi: 10.1111/j.1467-9280.2006.01750.x

Zaki, J., and Mitchell, J. P. (2013). Intuitive prosociality. *Curr. Dir. Psychol. Sci.* 22, 466–470. doi: 10.1177/0963721413492764

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.