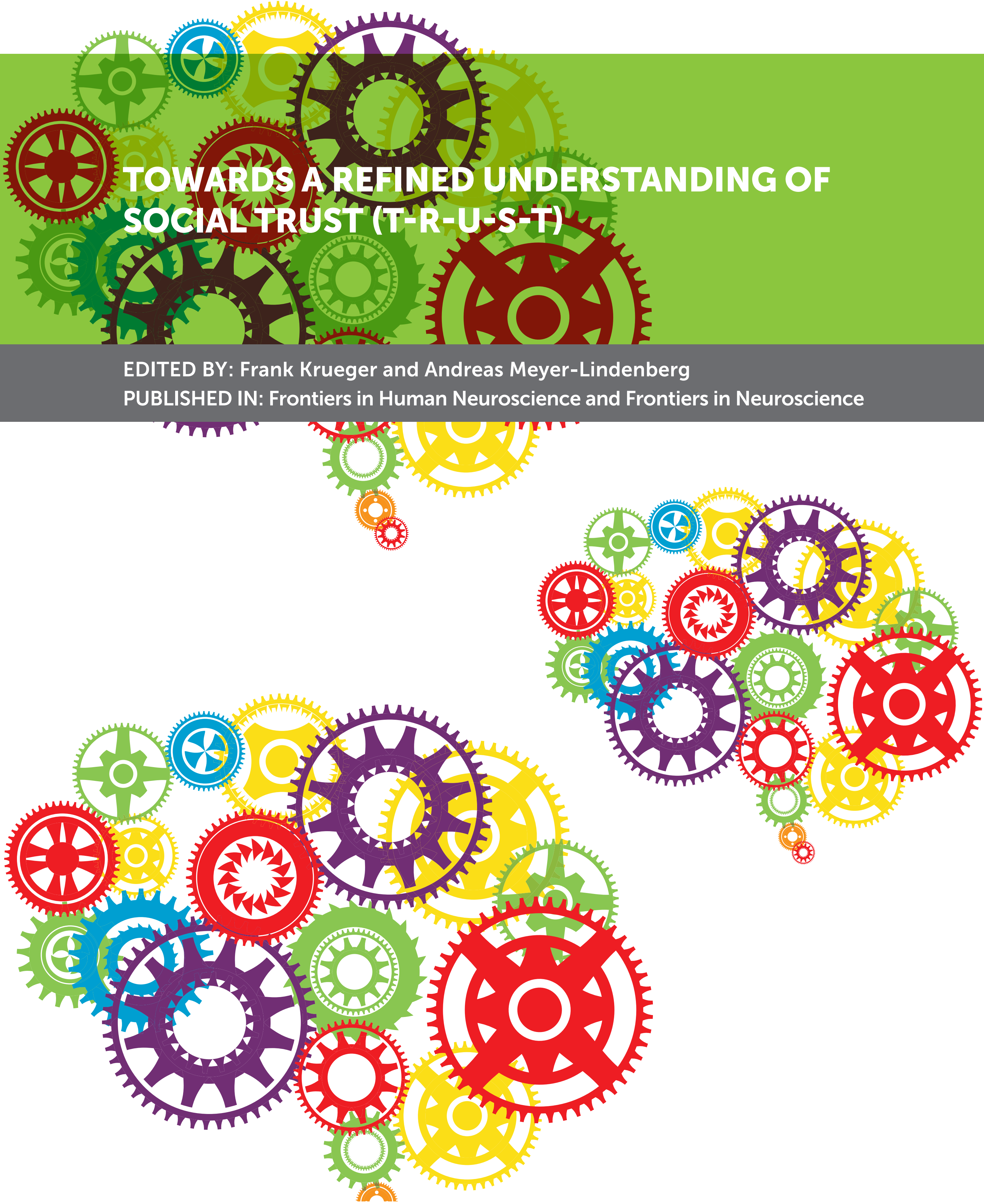
TOWARDS A REFINED UNDERSTANDING OF SOCIAL TRUST (T-R-U-S-T)

EDITED BY: Frank Krueger and Andreas Meyer-Lindenberg

PUBLISHED IN: Frontiers in Human Neuroscience and Frontiers in Neuroscience



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-075-9
DOI 10.3389/978-2-88966-075-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

TOWARDS A REFINED UNDERSTANDING OF SOCIAL TRUST (T-R-U-S-T)

Topic Editors:
Frank Krueger, George Mason University, United States
Andreas Meyer-Lindenberg, University of Heidelberg, Germany

Citation: Krueger, F., Meyer-Lindenberg, A., eds. (2020). Towards a Refined Understanding of Social Trust (T-R-U-S-T). Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-075-9

Table of Contents

EDITORIAL
published: 19 August 2020
doi: 10.3389/fnhum.2020.00305

Editorial: Towards a Refined Understanding of Social Trust (T-R-U-S-T)

Frank Krueger[1]* *and Andreas Meyer-Lindenberg*[2]

[1] School of Systems Biology, George Mason University, Fairfax, VA, United States, [2] Medical Faculty Mannheim, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany

Keywords: oxytocin, AVPR1A gene, gender, age, trust game

Editorial on the Research Topic

Towards a Refined Understanding of Social Trust (T-R-U-S-T)

Social trust is an essential ingredient for nearly every aspect of our daily lives. A plethora of investigations have started to gain a deeper understanding of trust; however, a coherent conceptual framework that integrates separate findings into a *psychoneurobiological model of trust* is still lacking. As a joined effort, psychologists, economists, and neuroscientists submitted for this Research Topic *empirical* and *theoretical work* in the form of *original research, review,* and *opinion papers* to shed light on the *behavioral*, *psychological,* and *neural levels* of trust:

At the *behavioral level,* the research community commonly relies on the trust game (TG) paradigm as an incentivized measure of individual variability for both trust and trustworthiness behavior. Alos-Ferrer and Farolfi reviewed not only the strength and limitations inherent in this popular paradigm but also explored the relations to alternative instruments for future investigations.

At the *neuropsychological level*, experimental paradigms allow evaluating the impact of contextual, idiosyncratic, and demographic factors on the psychological components of trust (motivation, affect, and cognition) and the underlying neural mechanisms—for example, through functional magnetic resonance imaging. Fareri highlighted in his review the neurobehavioral mechanisms of trust and reciprocity through the lens of implicit and explicit social appraisal and learning processes—stressing to focus more on its underlying *neurocomputational mechanisms* in future studies.

Fairley et al. examined people's own, naturally occurring beliefs to explore the subsequent outcome of their choices—implementing a TG for social and a lottery for non-social contexts. Only trust decisions as investment amount in TG parametrically modulated anticipatory reward and outcome evaluation in the ventral striatum—demonstrating a novel approach for using people's inherent sets of beliefs for studying reward processing.

Although economic decision-making is commonly characterized as a rational phenomenon, real-world decisions are clearly influenced by affect. Eimontaite et al. investigated cooperation as a precursor of trust while participants played a Prisoner's Dilemma game under partner-directed sympathy, anger, and neutral emotion conditions. Left amygdala activity was indicative of emotion enhancement and increment of cooperative behavior, whereas the left putamen suppressed emotion to overcome anger and engage in cooperation under the influence of partner directed emotion.

People may change their behavior, sometimes against their personal preferences, according to the opinions of their peers. Wei et al. studied the effect of social influence on trust behavior. Participants conformed to others' opinions and behaviors in the TG—activating ventromedial prefrontal cortex (PFC) and ventral striatum—indicating that they felt rewarded confirming to other's opinions.

Edited and reviewed by:
Lutz Jäncke,
University of Zurich, Switzerland

****Correspondence:***
Frank Krueger
fkrueger@gmu.edu

Specialty section:
This article was submitted to Cognitive Neuroscience, a section of the journal Frontiers in Human Neuroscience

Received: *29 June 2020*
Accepted: *08 July 2020*
Published: *19 August 2020*

Citation:
Krueger F and Meyer-Lindenberg A (2020) Editorial: Towards a Refined Understanding of Social Trust (T-R-U-S-T). Front. Hum. Neurosci. 14:305. doi: 10.3389/fnhum.2020.00305

Parental investment and social role theories predict that men trust more to maximize resources, whereas women trust less due to a higher sensitivity to social risk. Wu et al. examined gender differences in trust by simultaneously scanning male and female same-gender, fixed dyads, who played a multi-round TG with varying levels of payoff as an indicator of risk. Men trusted more than women, and the payoff level moderated the effect of gender on trust behavior. Men demonstrated equivalent activation in the subgenual anterior cingulate cortex across the payoff level, whereas women showed a decreased activation with increasing payoff level—explaining women's higher risk to social risk.

Gender differences in trust and trustworthiness during adolescence is a key period of change in social behavior. Lemmers-Jansen, Fett, Shergill et al. studied age-related gender differences in trust and trustworthiness in adolescence, implementing multi-round TGs simulating a pre-programmed cooperative and an unfair partner. For repeated cooperative interactions, no gender differences were found but younger compared to older adolescents showed a slightly steeper increase of investments, whereas younger males reacted with a stronger decrease of investments than older males for unfair interactions. Those gender-by-age interactions on trusting revealed activity in temporoparietal junction and caudate—showing a stronger influence of age in males than in females during cooperative and the reverse in unfair interactions.

At the *neurochemical level*, exogenous administration of neuropeptide hormones (e.g., oxytocin, OXT) helps to reveal the neural signaling pathway mechanisms underlying trust behavior. Original landmark studies claiming a crucial role of OXT in enhancing trust have been questioned by subsequent meta-analytic approaches, large scale non-replications, or failure to reproduce findings in different contexts. Xu et al. argue in their review that OXT may play a key role in conforming to and learning from trusted individuals who are either in-group members and/or perceived experts instead of facilitating trust *per se.* Therefore, future studies should establish how motivational, affective, and cognitive aspects of trust interact with the effects of OXT on social learning and conformity.

At the *neurogenetic level*, the impact of single-nucleotide polymorphisms such as the OXT receptor (OXTR) gene on trust behavior have been studied. Nishina et al. examined whether the association between a common repeat length polymorphism in an intron of the arginine-vasopressin receptor 1A (AVPR1a) gene is associated with TG and attitudinal trust measures. Compared to their previous OXTR gene findings, this polymorphism of AVPR1a also revealed sex differences: men with a short form of AVPR1a not only trusted but also reciprocated more in the TG, but no associations with attitudinal trust were found. As a result, future studies should examine the underlying brain functions and structures mediating the association between AVPR1a and trust behavior.

Identifying the psychoneurobiological patterns of trust in healthy people can potentially shed light on trust impairment, as present in some *psychiatric disorders*. A prime candidate helping to build trust as the glue to positive social interactions could be social mindfulness—the ability and willingness to see and consider another person's needs and wishes during social decision-making. Lemmers-Jansen, Fett, Van Doesum et al. investigated whether first-episode psychosis patients (FEP) and patients at clinical high-risk (CHR) show reduced social mindfulness applying a social mindfulness task. Relative to healthy controls and CHR, spontaneous social mindfulness was reduced in FEP—mirrored by reduced activity in caudate (sensitivity to the rewarding aspects of social mindfulness) and medial PFC (consideration for the other player)—but could be improved when explicitly told to act in another person's best interest.

The comprehensive collection of this T-R-U-S-T Research Topic will not only facilitate, broaden, and improve the current state of the psychoneurobiological signatures of social trust but also bring us a step closer to integrating research findings into a common conceptual framework of reciprocity behavior (including trust and trustworthiness behaviors). Krueger et al. presented an opinion about a neuropsychological framework that explains trust and trustworthiness in the context of reciprocity behaviors—determined by the evaluation of the kindness of a partner's normative action based on the intention as the underlying motivation and the outcome as the consequence of the action, highlining the role of the right anterior insula as a common currency of aversion for determining positive (i.e., norm compliance) and negative (i.e., norm enforcement) reciprocity.

AUTHOR CONTRIBUTIONS

FK drafted the editorial. FK and AM-L revised the final submitted version. All authors contributed to the article and approved the submitted version.

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

frontiers
in Neuroscience

REVIEW
published: 06 February 2019
doi: 10.3389/fnins.2019.00056

Oxytocin Facilitates Social Learning by Promoting Conformity to Trusted Individuals

Lei Xu, Benjamin Becker and Keith M. Kendrick*

The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Laboratory for NeuroInformation, University of Electronic Science and Technology of China, Chengdu, China

There is considerable interest in the role of the neuropeptide oxytocin in promoting social cohesion both in terms of promoting specific social bonds and also more generally for increasing our willingness to trust others and/or to conform to their opinions. These latter findings may also be important in the context of a modulatory role for oxytocin in improving the efficacy of behavioral therapy in psychiatric disorders. However, the original landmark studies claiming an important role for oxytocin in enhancing trust in others, primarily using economic game strategies, have been questioned by subsequent meta-analytic approaches or failure to reproduce findings in different contexts. On the other hand, a growing number of studies have consistently reported that oxytocin promotes conformity to the views of groups of in-group individuals. Most recently we have found that oxytocin can increase acceptance of social advice given by individual experts without influencing their perceived trustworthiness *per se*, but that increased conformity in this context is associated with how much an expert is initially trusted and liked. Oxytocin can also enhance the impact of information given by experts by facilitating expectancy and placebo effects. Here we therefore propose that a key role for oxytocin is not in facilitating social trust *per se* but in conforming to, and learning from, trusted individuals who are either in-group members and/or perceived experts. The implications of this for social learning and use of oxytocin as an adjunct to behavioral therapy in psychiatric disorders are discussed.

Keywords: oxytocin – therapeutic use, social conformity, interpersonal trust, expert, social learning

OPEN ACCESS

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

Reviewed by:
Jan B. Engelmann,
University of Amsterdam, Netherlands
Daniel Saul Levine,
The University of Texas at Arlington,
United States

****Correspondence:***
Keith M. Kendrick
k.kendrick.uestc@gmail.com

Specialty section:
This article was submitted to
Decision Neuroscience,
a section of the journal
Frontiers in Neuroscience

Received: *18 September 2018*
Accepted: *21 January 2019*
Published: *06 February 2019*

Citation:
Xu L, Becker B and Kendrick KM (2019) Oxytocin Facilitates Social Learning by Promoting Conformity to Trusted Individuals. Front. Neurosci. 13:56. doi: 10.3389/fnins.2019.00056

INTRODUCTION

Interpersonal trust within social groups is of key importance for social interactions, bonds, cooperation and learning and trust between different groups can also help ensure a stable and peaceful co-existence as well as mutually beneficial co-operation and trade. Trust is generally considered to be critical for co-operation and reciprocity in social and economic interactions but importantly trust also involves risk of potential injury if misplaced or broken and we have a natural aversion to taking such risks (Hardin, 2002; Ostrom and Walker, 2003). Indeed, an important factor influencing our trust behavior is that we are strongly motivated to avoid others betraying our trust (Bohnet and Zeckhauser, 2004; Bohnet et al., 2008). Trust can potentially be influenced by our assessment of the level of risk that trusting others might have and also by increased sensitivity to physical and/or other cues for detecting trustworthiness. It is therefore of great importance to identify both behavioral and physiological factors which can act to enhance trust, particularly in situations where individuals have impaired trust and therefore find it hard to interact socially with others and learn from them and/or to benefit optimally from cognitive and behavioral therapeutic strategies.

OXYTOCIN AND INTERPERSONAL TRUST IN THE CONTEXT OF ECONOMIC GAMES

For more than a decade the potential role of the hypothalamic neuropeptide oxytocin (OXT) in enhancing interpersonal trust in humans has received considerable attention. In an initial landmark study, Kosfeld et al. (2005) first reported that intranasal OXT administration could increase trust toward others in terms of being willing to make higher risk investments. This suggested that OXT might make individuals less risk-averse to trusting others. Following on from this Baumgartner et al. (2008) reported that intranasal OXT influenced the neural circuitry involved in trust and adaptation in response to it, notably in the amygdala and striatum critically engaged in fear and adaptive learning as well as the modulation of these functions by social contexts.

Subsequently it was reported that claimed trust-promoting effects of OXT in the context of economic games were sensitive to previous experience in that they did not occur following trust betrayal (Mikolajczak et al., 2010a), although this was contrary to the findings of Baumgartner et al. (2008) where experience of trust betrayal was found to have no effect on trust behavior after intranasal OXT. Another study reported that following experience of unfair treatment in the trust game women, but not men, were subsequently less forgiving of unfair treatment after intranasal OXT (Yao et al., 2014), which also suggested that it was not promoting "blind-trust." However, although an initial meta-analysis of OXT effects on trust reported a modest effect size (Van IJzendoorn and Bakermans-Kranenburg, 2012) a more recent review and meta-analysis restricted to studies involving economic games failed to demonstrate overall significant effects of OXT on interpersonal trust *per se*, thereby casting doubt on its role (Nave et al., 2015). Recently, another small-scale within subject study has actually reported some evidence for reduced trust in the context of a multi-round trust game following intranasal OXT (Ide et al., 2018).

Genetic association studies have additionally provided some support for a role of OXT and its receptor (OXTR) in trust during economic game paradigms (Israel et al., 2009; Krueger et al., 2012), although another study (Apicella et al., 2010) and a subsequent meta-analysis (see Bakermans-Kranenburg and van IJzendoorn, 2014) failed to confirm overall significance of these findings. While some studies have reported associations between trust/trustworthiness and blood concentrations of OXT (Zak et al., 2005; Zhong et al., 2012) others have not (Christensen et al., 2014) and some positive findings need to be treated with caution where unextracted assay protocols were employed which may not be reliable (see Leng and Ludwig, 2016).

There has been some other indirect support for a potential role of OXT in influencing interpersonal trust in economic games. Thus, testosterone treatment which would normally interact negatively with OXT (Crespi, 2016) was found to decrease trust in economic games (Bos et al., 2010). On the other hand, enhancing serotonin function using treatment with tryptophan, which would be expected to indirectly enhance OXT signaling (Dölen et al., 2013), increased trust in the same context (Colzato et al., 2013). Overall, however, the case for proposing that OXT can generally enhance trust in the context of economic games is difficult to support and it should also be noted that they measurement of trust is generally operationalized as willingness to transfer money to another individual who may or may not reciprocate. While studies have indicated that willingness to transfer money in these economic paradigms is significantly associated with levels of interpersonal trust (Van't Wout and Sanfey, 2008) they are nevertheless independent measures and factors other than altered trust alone could be contributing to OXT influencing investment decision making.

OXYTOCIN AND INTERPERSONAL TRUST IN OTHER CONTEXTS

The effects of intranasal OXT on increasing interpersonal trust have also been investigated in contexts where trust is not measured simply in terms of whether subjects are prepared to give specific individuals more or less money while playing economic games. Mikolajczak et al. (2010b) initially reported that intranasal OXT enhanced trust using an "Envelope Task" paradigm where subjects indicated their level of trust in an experimenter's promise that their recorded intimate personal details would be kept confidential by whether they wanted to seal the envelope containing their revelations or leave it open. However, this experiment was performed single rather than double blind and failed to be replicated under double-blind conditions (Lane et al., 2015). A few other unpublished studies found no OXT effects on self-reported measures of trust (see Nave et al., 2015) or ones that are dependent upon person-specific characteristics. Thus, one study showed that OXT enhanced levels of trust following social exclusion in the Cyberball game, but only in individuals who reported a negative emotional response to being excluded (Cardoso et al., 2013), and another showed that interpersonal trust was only increased in Democrats with low initial personal trust (Merolla et al., 2013). Finally, one study has reported that intranasal OXT enhanced trust/compliance with reliable, but not unreliable, human-like automatons (De Visser et al., 2017).

The effects of OXT on perception of implicit trustworthiness have also been investigated in a number of other contexts, although again with variable findings. An early study for example reported that intranasal OXT increased trustworthiness ratings of neutral expression faces although this was in combination with attractiveness (Theodoridou et al., 2009) but many studies have subsequently failed to find any effects on facial trustworthiness ratings *per se* (Guastella et al., 2008; Lambert et al., 2014; Quintana et al., 2015; Luo et al., 2017; Woolley et al., 2017), including in either young or old subjects (Grainger et al., 2018). This apparent lack of influence of OXT on ratings of implicit trustworthiness from faces is important given that there is a strong association between such ratings and subjects' willingness to transfer money to specific individuals in economic games (Van't Wout and Sanfey, 2008). This would therefore imply that OXT could somehow influence trustworthiness in terms of willingness to invest in another individual although without necessarily making them more implicitly trustworthy. On the other hand, OXT may improve judgments of trustworthiness from faces by enhancing detection of untrustworthy individuals (see Lambert et al., 2014).

Brain imaging studies have shown that perception of trust in faces is negatively associated with activation in limbic regions engaged in fear processing, particularly the amygdala (Engell et al., 2007) and positively associated with activation in core nodes of the reward processing circuitry such as the orbitofrontal cortex and striatum (Mende-Siedlecki et al., 2012). Brain lesion studies have provided additional support for a critical engagement of the amygdala (Adolphs et al., 1998; Koscik and Tranel, 2011; Van Honk et al., 2013), insula (Belfi et al., 2015), and medial prefrontal cortex (Moretto et al., 2013) in trust behavior. The activity and functional connectivity of all of these brain regions have repeatedly been shown to mediate behavioral effects of OXT administration (see Kendrick et al., 2017), although this does not necessarily directly imply that OXT is acting on these same regions to promote trust since they are also implicated more widely in many different aspects of social cognition.

OXYTOCIN AND INTERPERSONAL TRUST IN CLINICAL POPULATIONS

In a clinical context trust is of crucial importance in therapist-patient relationships, particularly in therapeutic counseling and behavioral therapy. Our ability to trust others can be impaired as a result of negative social experiences and insecure attachment (Corriveau et al., 2009) and in a number of mental disorders such as schizophrenia (Keri et al., 2009), and some personality disorders such as borderline personality disorder (Fonagy et al., 2015). On the other hand, individuals with disorders, such as Autism Spectrum Disorder may have problems in accurately perceiving trust cues (Yi et al., 2013; Ewing et al., 2015) and therefore find it harder to detect when they are being deceived. However, in the context of clinical studies, intranasal OXT has been found to actually reduce trust in subjects with borderline personality disorder (Bartz et al., 2011a), particularly in individuals with experience of childhood trauma (Ebert et al., 2013). On the other hand, OXT may enhance trust in Prader-Willi syndrome (Tauber et al., 2011), a genetic disorder characterized by disruptive behaviors and marked interpersonal problems. A study on individuals with Autism Spectrum Disorder using a modified Cyberball game paradigm also found that OXT increased co-operation more with individuals who reciprocated throwing the ball back to the subject and they also reported having increased trust in them (Andari et al., 2010). However, use of OXT as an adjunct to behavioral therapy has so far met with limited success, which again could be considered as indirect evidence for it failing to increase trust in the therapist (Guastella et al., 2009; MacDonald et al., 2013).

OXYTOCIN AND TRUST: SUMMARY

Overall therefore, despite the attractiveness and hypothetical support for OXT playing a key functional role in directly influencing interpersonal trust, accumulating empirical evidence makes this view hard to maintain. Indeed, it would seem that as with many behavioral effects of this neuropeptide, there may at the very least be complex contributions of both context and previous personal experience to what precise treatment outcomes on trust perception or behavior are observed (Bartz et al., 2011b; Shamay-Tsoory and Abu-Akel, 2016). In addition, if OXT really does play a fundamental role in promoting interpersonal trust then there would be an expectation that it would not promote behaviors which may serve to damage or weaken trust in some way. In this context recent findings showing that OXT can facilitate envy (Shamay-Tsoory et al., 2009), lying and deception (Shalvi and De Dreu, 2014; Aydogan et al., 2017), including for self-benefit (Scheele et al., 2014; Sindermann et al., 2018) and aggression (Ne'eman et al., 2016) suggest that it can indeed potentially promote trust-damaging behaviors.

OXYTOCIN AND SOCIAL CONFORMITY TO MEMBERS OF IN-GROUPS

We are not only more likely to accept the opinions and advice of others, co-operate with them more and learn from them simply as a result of trusting them more, this can also occur as a result of forming closer social ties with them (Feng and MacGeorge, 2006). It is well established that we will often change our views and preferences to match those expressed by others in our social group in order to fit in. This is referred to as the "social conformity" effect and while it often represents a transient (<3 days) change in our publically expressed views in response either to explicit or implicit social influence (see Huang et al., 2014), it can also result in more enduring changes in our privately held ones. It has been argued that such social conformity reflects learning which is reinforced by the positive reward value of adhering to social norms, together with fear of punishment when we fail to do so (Cialdini and Goldstein, 2004). We are also more likely to accept the advice and opinions of inherently trusted individuals who are part of our social in-group, most notably partners, friends and relatives (Brewer, 2008). We generally trust members of our social in-group more than others and this forms the basis of our increased willingness to co-operate with and protect and learn from them.

Given the evolutionary key role of OXT in promoting affiliative bonds (Kendrick, 2000; Striepens et al., 2011) it may be this aspect of its functioning which increases conformity and willingness to accept and co-operate with and learn from in-group members rather than by enhancing interpersonal trust *per se*. That OXT could function to influence our trust-associated behaviors indirectly by affecting the strength of our affiliation with others at either a group or individual level is firstly supported by another landmark paper in the field. This paper reported that OXT can increase both "trust-in" and "love-for" in-group but not out-group members in the context of monetary investment behavior exhibited during an economic trust game (De Dreu et al., 2010). Further studies have also established that OXT enhances liking for and co-operation with in-group members, irrespective of whether they co-operate with us or not (De Dreu et al., 2011; Ma et al., 2014) and can even increase deceptive behavior for the benefit of in-group members (Shalvi and De Dreu, 2014). A meta-analysis also suggested intranasal OXT elevates the level of in-group but not out-group trust (Van IJzendoorn and Bakermans-Kranenburg, 2012).

Our level of trust in members of our in-group such as friends, family and partners is naturally higher than for members of out-groups and strangers, as is our liking for them, and likability and trustworthiness are correlated to some extent. However, liking and trust are dissociable, with trust for example being associated with levels of perceived self-control in others whereas liking is not (Righetti and Finkenauer, 2011). Liking, attraction and trustworthy judgments from face features are also dissociable and the correlation between liking and trust can be weakened by age (Todorov et al., 2015). Importantly, in the context of establishing the nature of OXT's relative functional effects on these two social dimensions, studies have more consistently shown that it can enhance liking for the faces of individuals either presented alone (Striepens et al., 2014), or in combination with specific information about an individual's behavior or expertise (Chen et al., 2015; Gao et al., 2016; Zhao et al., 2018b) rather than trust. Unfortunately, to date only one study has measured both liking and trust following OXT administration and found that it specifically enhanced liking/attraction ratings for individuals and not those of trustworthiness (Xu L. et al., 2018). Here, male and female subjects were required to rate likeability and trustworthiness of potential romantic partners associated with a previous history of fidelity or infidelity and OXT only influenced likeability/attraction ratings and not those of trustworthiness.

In further support of the above proposal, and in marked contrast to the inconsistent findings from studies investigating the effects of OXT on interpersonal trust *per se*, there is substantial and consistent evidence demonstrating that it facilitates conformity to the opinions expressed by groups of in-group members (De Dreu and Kret, 2016). This effect occurs both within culturally long-term established in-groups (Huang et al., 2015) and those formed arbitrarily in the short-term context of a competitive environment (Stallen et al., 2012; Edelson et al., 2015). Furthermore, OXT can also facilitate norm-based compliance and can even counteract ethnocentrism which it normally promotes. This was demonstrated elegantly in the context of individuals who exhibited xenophobic tendencies in terms of charitable donations becoming more likely to donate to immigrants when OXT was administered in association with reinforcement of norm-based altruism (Marsh et al., 2017).

OXYTOCIN AND CONFORMITY TO PERCEIVED "EXPERTS"

Our acceptance that someone is an "expert" of some kind (e.g., elders, teachers, doctors, or professionals in other areas, etc.) implies that we are more likely to consider their opinions and advice in relation to the specific field of their expertise as trustworthy (Bonaccio and Dalal, 2006). As such, we will also learn from and potentially co-operate with them, although this does not necessarily imply that we will consider such individuals as generally more trustworthy than others outside of their area of expertise. In contrast to the situation with in-groups where the effect of OXT on increased conformity and co-operation may be partly contributed to by increased affiliation, that involving similar behavior in relation to perceived experts may be different.

Recently we have demonstrated that OXT-enhancement of conformity can extend to the context of acceptance of social advice from individual experts in psychological counseling (Luo et al., 2017). In this study, male participants were first invited to provide solutions to a number of everyday social problems and then following treatment were given alternative advice (either better or worse) by male or female experts or non-experts (landscape gardeners). Participants were not familiar with the advisors and were simply shown pictures of them together with information about their expertise. All the advisors were older than the subjects to further enhance their perceived experience and potential reliability (see Lourenco et al., 2015). Oxytocin treatment significantly increased the proportion of advice accepted from female experts, irrespective of whether the solutions offered by them were better or worse than those originally chosen by the subjects themselves (see **Figure 1A**). The use of a counterbalanced design ensured that this effect of OXT was independent of the appearance of specific advisors and therefore influenced only by their attributed expertise. Importantly, OXT did not influence the perceived trustworthiness or likeability of either the experts or the non-experts, but its effects on acceptance of advice were positively associated with both (see **Figure 1B**). This resulted in a greater degree of acceptance of advice from female experts who were generally rated as more trustworthy and likeable than both female non-experts and the equivalent male experts. Indeed, overall acceptance of advice across all advisors was positively correlated with their perceived trustworthiness and OXT tended to increase this correlation (see **Figure 1B**). Thus, the study provides the first direct evidence for an interaction between perceived trustworthiness and likeability and the ability of OXT to increase conformity to advice given by individual experts. While both this study and one from another group (Edelson et al., 2015) showed that the effect of OXT on increasing conformity was transient, this is perhaps not that surprising given the rather controlled contexts and that advice is given only once. Indeed, we don't tend to give up self-held beliefs and judgments very easily and sometimes will deliberately disobey expert advice (Engelmann et al., 2009; Suen et al., 2014). Further experiments are required to investigate whether OXT can influence long-term privately held views following repetition of advice in more natural circumstances and the extent to which it can alter the behavior of individuals who tend to disobey expert advice.

While studies have yet to establish the neural substrates where OXT may act to facilitate taking advice from experts the amygdala may be of importance in this respect. The amygdala shows increased activation during the positive evaluation of advisor competence (Schilbach et al., 2013) and following advice (Biele

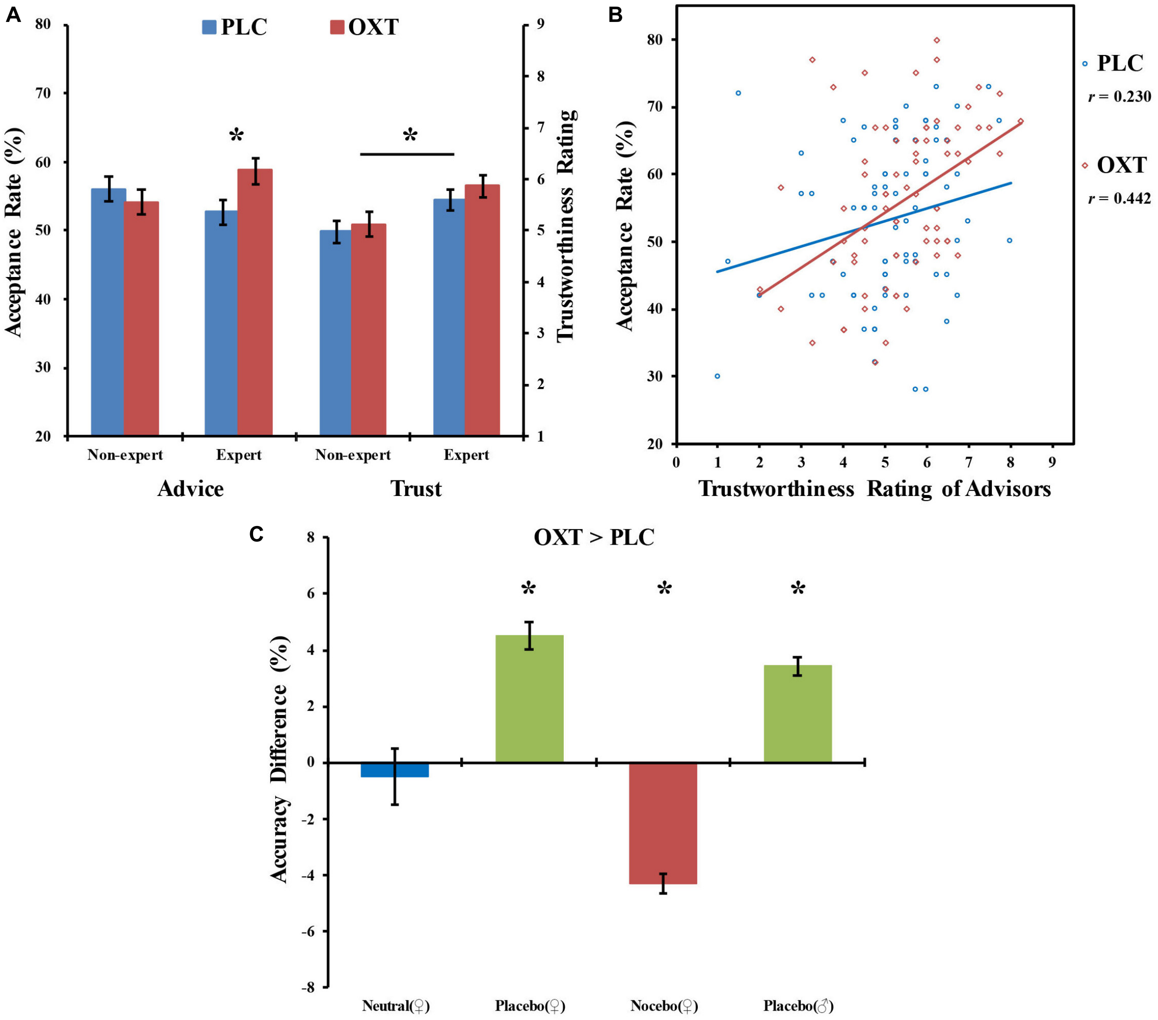


FIGURE 1 | (A) Mean ± SEM % acceptance of advice by male subjects on solutions to social problems given by the same female advisors who were either designated as non-experts (landscape gardeners) or experts (psychological counselors) in giving social advice. Before the paradigm participants were randomly assigned to either intranasal oxytocin (OXT – 40 IU) or placebo (PLC) treatment. OXT significantly increased acceptance of advice from the advisor when she was designated as an expert but not a non-expert. While the advisor as an expert was trusted significantly more than as a non-expert, OXT administration *per se* did not influence ratings of trustworthiness. *$p < 0.05$ for OXT vs. PLC or trust ratings in expert vs. non-expert advisors, respectively. **(B)** Regression graph showing a positive correlation between acceptance of advice from different male and female expert and no-expert advisors and their trustworthiness ratings in subjects receiving PLC or OXT treatment. Subjects receiving OXT generally showed a stronger positive correlation between advice acceptance and trustworthiness ratings [OXT: $r = 0.442$, $p < 0.001$; PLC: $r = 0.230$, $p = 0.047$; Fisher z-score = 1.43, $p = 0.076$ (one tailed)]. Data for **(A,B)** are taken from Luo et al. (2017). **(C)** Histograms show the effects of intranasal OXT (24 IU) vs. PLC nasal spray alone (blue bar) or in combination with either advice from a female or male expert in a white coat telling subjects that their working memory performance will be improved (placebo effect, green bar) or impaired (nocebo effect, red bar). Results are combined data from verbal, spatial, and social n-back tasks (1-back and 2-back combined) taken from Zhao et al. (2018a). *$p < 0.05$ OXT vs. PLC.

et al., 2011) and the amygdala is also one of the primary regions where OXT has been found to produce functional effects on social cognition (Kendrick et al., 2017).

That OXT may act to enhance conformity to, and co-operation with, individuals who are highly trusted is consistent with the repeated observations that it facilitates these behaviors in members of an in-group, who are perceived as more trustworthy and likeable than out-group members. Within the context of in-groups, however, OXT may also enhance the perceived expertise of some individuals compared to others without necessarily influencing their trustworthiness. Thus in pairs of subjects working together to solve a visual search task intranasal OXT treatment made the less competent partner more likely to conform to the opinion of the more competent one, and had the opposite effect on the competent partner (Hertz et al., 2016). The perceived expertise factor can also help explain why OXT increases conformity with out-group members in some contexts. For instance, when male Chinese subjects were asked to judge the attractiveness of Asian female faces, and then informed about ratings given by male peers from an in-group (Chinese) or an out-group (Japanese), OXT increased conformity to the opinions of both (Huang et al., 2015). In contrast, OXT increased the likeability of Chinese people, monuments and commercial products but did not have any effect on liking/disliking of comparable Japanese stimuli, suggesting that it reliably induces an in-group preference within this context (Ma et al., 2014). Arguably, male peers from both Chinese and Japanese cultures would be considered to have similar expertise in judging the facial attractiveness of Asian female faces and this perceived similarity of expertise may have resulted in OXT enhancing the impact of the opinions of both due to equivalent levels of trust in the expertise of in-group and out-group members in this specific context.

Other contexts in which OXT appears to function to enhance acceptance of information or skills provided by trusted experts is in relation to its facilitation of placebo effects and also susceptibility to hypnosis. Several studies have now demonstrated that intranasal OXT can enhance or even generate placebo effects. Thus, in the context of analgesia OXT has been reported to enhance the placebo effect on pain perception (Kessner et al., 2013). In a recent experiment we also showed that OXT given in conjunction with a male or female experimenter wearing a white coat informing them that the treatment would enhance their working memory performance exhibited an impressive 5% increased improvement in accuracy in verbal, spatial, and social domains (Zhao et al., 2018a). In the absence of the adjunct OXT treatment there was no placebo effect and no effect of OXT given alone. Importantly, OXT could also generate an equivalent magnitude nocebo effect where subjects informed that the treatment would make them perform worse rather than better did indeed show an equivalent 5% significant performance deficit (Zhao et al., 2018a; see **Figure 1C**). Furthermore, intranasal OXT can increase the hypnotizability of individuals normally showing low pre-treatment susceptibility to hypnosis, although without influencing their perceived trust in the hypnotist (Bryant et al., 2012). Together these findings further demonstrate that OXT can facilitate the impact of advice/information received from individual experts without necessarily making such individuals either more likeable or trustworthy. Indeed, we have recently reported that OXT can promote increased co-operation with individuals in the Cyberball game who they rate as both less trustworthy and likeable due to their exclusion of other players in order to gain a higher monetary reward (Xu X. et al., 2018). In this case subjects receiving OXT played more with these specific players since doing so would be likely to increase their own financial gain, i.e., there would be a greater expectation that such excluder players would reciprocate with them for mutual benefit despite them being considered generally less trustworthy and likeable.

OXYTOCIN AS A FACILITATOR OF SOCIAL LEARNING

Overall therefore, it may be more relevant to consider OXT as functioning to facilitate social learning both as a result of enhancing affiliative bonds with trusted in-group members and also from the information/advice/skills transmitted by trusted experts who are not necessarily within an individual's immediate social group. Trust in these two contexts may potentially be more "affective" in the in-group context and more "cognitive" in the expert one. While distinctions between the relative importance of cognitive and affective trust are made routinely in economic and business contexts (see Dowell et al., 2015) they are not usually distinguished in interpersonal social ones, and this may be important when considering effects of OXT in light of it facilitating emotional empathy but not cognitive empathy in some tasks (Hurlemann et al., 2010; Geng et al., 2018).

Social learning from trusted individuals and groups plays a fundamental role throughout our lives in both promoting social cohesion as well as providing us with the information and strategies to cope with and adapt to the challenges we face every day. While social learning has often been considered as distinct from other forms of learning it has been shown to involve the same associative processes as simple reward-based learning (Behrens et al., 2008). There is increasing evidence that OXT may be playing a key role in promoting social learning from the most appropriate individuals and several studies have also demonstrated this in the context of enhanced probabilistic learning of arbitrary information following social but not non-social reinforcement (Hurlemann et al., 2010; Hu et al., 2015). In these two latter studies subjects were required to learn which of a group of random 3-digit numbers was arbitrarily associated with two different categories following receipt of either social (smiling vs. angry faces) or non-social (red vs. green traffic lights) feedback. OXT selectively enhanced learning with social, but not non-social, feedback in both Caucasian (Hurlemann et al., 2010) and Asian (Hu et al., 2015) subjects and this was associated with increased activation in the amygdala and striatal regions and their functional connectivity (Hu et al., 2015). These studies were unable to distinguish whether OXT differentially enhanced the effects of positive and/or negative social feedback, however, another group demonstrated that OXT enhances activity in the ventral tegmental area in response to both positive (smiling faces) and negative (angry faces) feedback (Groppe et al., 2013). Thus, and in line with social conformity being reinforced by both social reward and fear of social punishment, it seems likely that OXT is promoting social learning via not only increasing the impact of positive social reward cues but also those of social punishment via modulation of amygdalo-frontal-striatal reward networks.

In general reward-based learning involves two different decision control systems, a cognitive "model-based" system and a simpler "model-free" system based on habit (Dolan and Dayan, 2013). Both model-based and model free learning engage striatal circuitry (Daw et al., 2011) and therefore OXT could potentially influence both, although to date only simple probabilistic model-free learning paradigms have been used which guide action and do not involve any rule learning but can easily be utilized to compare the relative effects of social compared with non-social feedback.

CLINICAL IMPLICATIONS OF OXYTOCIN AS A FACILITATOR OF SOCIAL LEARNING

In support of an important role for OXT in promoting social learning in a clinical context a recent study has reported that it facilitates probabilistic learning with social feedback and enhanced striatal activation in individuals with Autism Spectrum Disorder (Kruppa et al., 2018). This study used the same learning paradigm where OXT was found to facilitate learning with social feedback in healthy subjects although only positive and neutral social feedback were included. In terms of the potential therapeutic use of intranasal OXT to improve the efficacy of cognitive or other therapist-based interventions for mental disorders there are several implications of the findings and interpretations detailed in the current review. Firstly, there is no convincing evidence that OXT treatment *per se* will make individuals more trustworthy, although it may do so indirectly by strengthening affiliative ties with in-group members. While trust in a specific therapist might effectively increase over time as they become equivalent to an in-group member for individual patients, clearly under the majority of circumstances it would be greater importance if patients have a high level of trust in the ability of the therapist as an expert if adjunct OXT treatment is likely to have any beneficial effect. If patients are suspicious of, or have low levels of trust in, a therapist then either OXT is unlikely to have any beneficial effect or quite possibly it might end up having a negative impact by further reducing trust levels, as for example observed in borderline personality disorder (Bartz et al., 2011a; Ebert et al., 2013). Thus, an important consideration for whether OXT might be beneficial as an adjunct to any kind of therapist-based behavioral intervention may be a patient's general levels of interpersonal trust and trust in their specific therapist. This underlines the well-established importance of initial trust-building between patient and therapist termed the "working-alliance," the strength of which has a strong bearing on treatment outcome and patient satisfaction (Fuertes et al., 2007).

CONCLUSION AND FUTURE DIRECTIONS

In summary, we have argued in this review that OXT administration rather than enhancing either implicit or explicit trust in others is instead primarily promoting social cohesion by facilitating increased conformity to others who we trust either as in-group members and/or as perceived experts. As such, OXT can be viewed as facilitating socially reinforced learning, particularly from trusted individuals, via amygdalo-frontal-striatal circuitry to increase the motivation to receive and respond to either social reward or punishment. It will be important in future studies to establish how cognitive and emotional aspects of trust interact with the effects of OXT on social learning and conformity. To date the effects of OXT have also only been investigated in the context of simple reinforcement paradigms involving model-free learning and it will be important to investigate whether they can extend to more cognitive model-based learning. From a therapeutic standpoint it will also be important in future studies to determine the extent to which levels of patient trust in the expertise of the therapist influence the effectiveness of OXT as an adjunct to behavior therapy.

AUTHOR CONTRIBUTIONS

All authors contributed to the information and ideas presented in the review and writing of the manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (NSFC) (31530032 and 91632117).

REFERENCES

Adolphs, R., Tranel, D., and Damasio, A. R. (1998). The human amygdala in social judgment. *Nature* 393, 470–474. doi: 10.1038/30982

Andari, E., Duhamel, J.-R., Zalla, T., Herbrecht, E., Leboyer, M., and Sirigu, A. (2010). Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4389–4394. doi: 10.1073/pnas.0910249107

Apicella, C. L., Cesarini, D., Johannesson, M., Dawes, C. T., Lichtenstein, P., Wallace, B., et al. (2010). No association between oxytocin receptor (OXTR) gene polymorphisms and experimentally elicited social preferences. *PLoS One* 5:e11153. doi: 10.1371/journal.pone.0011153

Aydogan, G., Jobst, A., D'Ardenne, K., Müller, N., and Kocher, M. G. (2017). The detrimental effects of oxytocin-induced conformity on dishonesty in competition. *Psychol. Sci.* 28, 751–759. doi: 10.1177/0956797617695100

Bakermans-Kranenburg, M. J., and van IJzendoorn, M. H. (2014). A sociability gene? Meta-analysis of oxytocin receptor genotype effects in humans. *Psychiatr. Genet.* 24, 45–51. doi: 10.1097/YPG.0b013e3283643684

Bartz, J. A., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., et al. (2011a). Oxytocin can hinder trust and cooperation in borderline personality disorder. *Soc. Cogn. Affect. Neurosci.* 6, 556–563. doi: 10.1093/scan/nsq085

Bartz, J. A., Zaki, J., Bolger, N., and Ochsner, K. N. (2011b). Social effects of oxytocin in humans: context and person matter. *Trends Cogn. Sci.* 15, 301–309. doi: 10.1016/j.tics.2011.05.002

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., and Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58, 639–650. doi: 10.1016/j.neuron.2008.04.009

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

Belfi, A. M., Koscik, T. R., and Tranel, D. (2015). Damage to the insula is associated with abnormal interpersonal trust. *Neuropsychologia* 71, 165–172. doi: 10.1016/j.neuropsychologia.2015.04.003

Biele, G., Rieskamp, J., Krugel, L. K., and Heekeren, H. R. (2011). The neural basis of following advice. *PLoS Biol.* 9:e1001089. doi: 10.1371/journal.pbio.1001089

Bohnet, I., Greig, F., Herrmann, B., and Zeckhauser, R. (2008). Betrayal aversion: evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *Am. Econ. Rev.* 100, 628–633. doi: 10.1257/aer.98.1.294

Bohnet, I., and Zeckhauser, R. (2004). Trust, risk and betrayal. *J. Econ. Behav. Organ.* 55, 467–484. doi: 10.1016/j.jebo.2003.11.004

Bonaccio, S., and Dalal, R. S. (2006). Advice taking and decision-making: an integrative literature review, and implications for the organizational sciences.

Organ. Behav. Hum. Decis. Process. 101, 127–151. doi: 10.1016/j.obhdp.2006.07.001

Bos, P. A., Terburg, D., and Van Honk, J. (2010). Testosterone decreases trust in socially naive humans. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9991–9995. doi: 10.1073/pnas.0911700107

Brewer, M. B. (2008). "Depersonalized trust and ingroup cooperation," in *Rationality and Social Responsibility: Essays in Honor of Robyn Mason Dawes. Modern pioneers in Psychological Science: An APS-Psychology Press series*, ed. J. I. Krueger (New York, NY: Psychology Press), 215–232.

Bryant, R. A., Hung, L., Guastella, A. J., and Mitchell, P. B. (2012). Oxytocin as a moderator of hypnotizability. *Psychoneuroendocrinology* 37, 162–166. doi: 10.1016/j.psyneuen.2011.05.010

Cardoso, C., Ellenbogen, M. A., Serravalle, L., and Linnen, A.-M. (2013). Stress-induced negative mood moderates the relation between oxytocin administration and trust: evidence for the tend-and-befriend response to stress? *Psychoneuroendocrinology* 38, 2800–2804. doi: 10.1016/j.psyneuen.2013.05.006

Chen, F. S., Mayer, J., Mussweiler, T., and Heinrichs, M. (2015). Oxytocin increases the likeability of physically formidable men. *Soc. Cogn. Affect. Neurosci.* 10, 797–800. doi: 10.1093/scan/nsu116

Christensen, J. C., Shiyanov, P. A., Estepp, J. R., and Schlager, J. J. (2014). Lack of association between human plasma oxytocin and interpersonal trust in a prisoner's dilemma paradigm. *PLoS One* 9:e116172. doi: 10.1371/journal.pone.0116172

Cialdini, R. B., and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annu. Rev. Psychol.* 55, 591–621. doi: 10.1146/annurev.psych.55.090902.142015

Colzato, L. S., Steenbergen, L., de Kwaadsteniet, E. W., Sellaro, R., Liepelt, R., and Hommel, B. (2013). Tryptophan promotes interpersonal trust. *Psychol. Sci.* 24, 2575–2577. doi: 10.1177/0956797613500795

Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., et al. (2009). Young children's trust in their mother's claims: Longitudinal links with attachment security in infancy. *Child Dev.* 80, 750–761. doi: 10.1111/j.1467-8624.2009.01295.x

Crespi, B. J. (2016). Oxytocin, testosterone, and human social cognition. *Biol. Rev.* 91, 390–408. doi: 10.1111/brv.12175

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215. doi: 10.1016/j.neuron.2011.02.027

De Dreu, C. K. W., Greer, L. L., Handgraaf, M. J. J., Shalvi, S., Van Kleef, G. A., Baas, M., et al. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* 328, 1408–1411. doi: 10.1126/science.1189047

De Dreu, C. K. W., Greer, L. L., Van Kleef, G. A., Shalvi, S., and Handgraaf, M. J. J. (2011). Oxytocin promotes human ethnocentrism. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1262–1266. doi: 10.1073/pnas.1015316108

De Dreu, C. K. W., and Kret, M. E. (2016). Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense. *Biol. Psychiatry* 79, 165–173. doi: 10.1016/j.biopsych.2015.03.020

De Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., et al. (2017). A little Anthropomorphism goes a long way: effects of oxytocin on trust, compliance, and team performance with automated agents. *Hum. Fact.* 59, 116–133. doi: 10.1177/0018720816687205

Dolan, R. J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron* 80, 312–325. doi: 10.1016/j.neuron.2013.09.007

Dölen, G., Darvishzadeh, A., Huang, K. W., and Malenka, R. C. (2013). Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. *Nature* 501, 179–184. doi: 10.1038/nature12518

Dowell, D., Morrison, M., and Heffernan, T. (2015). The changing importance of affective trust and cognitive trust across the relationship lifecycle: a study of business-to-business relationships. *Ind. Mark. Manage.* 44, 119–130. doi: 10.1016/j.indmarman.2014.10.016

Ebert, A., Kolb, M., Heller, J., Edel, M.-A., Roser, P., and Brüne, M. (2013). Modulation of interpersonal trust in borderline personality disorder by intranasal oxytocin and childhood trauma. *Soc. Neurosci.* 8, 305–313. doi: 10.1080/17470919.2013.807301

Edelson, M. G., Shemesh, M., Weizman, A., Yariv, S., Sharot, T., and Dudai, Y. (2015). Opposing effects of oxytocin on overt compliance and lasting changes to memory. *Neuropsychopharmacology* 40:966. doi: 0.1038/npp.2014.273

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

Engelmann, J. B., Capra, C. M., Noussair, C., and Berns, G. S. (2009). Expert financial advice neurobiologically "offloads" financial decision-making under risk. *PLoS One* 4:e4957. doi: 10.1371/journal.pone.0004957

Ewing, L., Caulfield, F., Read, A., and Rhodes, G. (2015). Appearance-based trust behaviour is reduced in children with autism spectrum disorder. *Autism* 19, 1002–1009. doi: 10.1177/1362361314559431

Feng, B., and MacGeorge, E. L. (2006). Predicting receptiveness to advice: characteristics of the problem, the advice-giver, and the recipient. *South. Commun. J.* 71, 67–85. doi: 10.1080/10417940500503548

Fonagy, P., Luyten, P., and Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: a new conceptualization of borderline personality disorder and its psychosocial treatment. *J. Pers. Disord.* 29, 575–609. doi: 10.1521/pedi.2015.29.5.575

Fuertes, J. N., Mislowack, A., Bennett, J., Paul, L., Gilbert, T. C., Fontan, G., et al. (2007). The physician–patient working alliance. *Patient Educ. Counsel.* 66, 29–36. doi: 10.1016/j.pec.2006.09.013

Gao, S., Becker, B., Luo, L., Geng, Y., Zhao, W., Yin, Y., et al. (2016). Oxytocin, the peptide that bonds the sexes also divides them. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7650–7654. doi: 10.1073/pnas.1602620113

Geng, Y., Zhao, W., Zhou, F., Ma, X., Yao, S., Hurlemann, R., et al. (2018). Oxytocin enhancement of emotional empathy: Generalization across cultures and effects on amygdala activity. *Front. Neurosci.* 12:512. doi: 10.3389/fnins.2018.00512

Grainger, S. A., Henry, J. D., Steinvik, H. R., and Vanman, E. J. (2018). Intranasal oxytocin does not alter initial perceptions of facial trustworthiness in younger or older adults. *J. Psychopharmacol.* doi: 10.1177/0269881118806303 [Epub ahead of print].

Groppe, S. E., Gossen, A., Rademacher, L., Hahn, A., Westphal, L., Gründer, G., et al. (2013). Oxytocin influences processing of socially relevant cues in the ventral tegmental area of the human brain. *Biol. Psychiatry* 74, 172–179. doi: 10.1016/j.biopsych.2012.12.023

Guastella, A. J., Howard, A. L., Dadds, M. R., Mitchell, P., and Carson, D. S. (2009). A randomized controlled trial of intranasal oxytocin as an adjunct to exposure therapy for social anxiety disorder. *Psychoneuroendocrinology* 34, 917–923. doi: 10.1016/j.psyneuen.2009.01.005

Guastella, A. J., Mitchell, P. B., and Mathews, F. (2008). Oxytocin enhances the encoding of positive social memories in humans. *Biol. Psychiatry* 64, 256–258. doi: 10.1016/j.biopsych.2008.02.008

Hardin, R. (2002). *Trust and Trustworthiness*. New York, NY: Russell Sage Foundation.

Hertz, U., Kelly, M., Rutledge, R. B., Winston, J., Wright, N., Dolan, R. J., et al. (2016). Oxytocin effect on collective decision making: a randomized placebo controlled study. *PLoS One* 11:e0153352. doi: 10.1371/journal.pone.0153352

Hu, J., Qi, S., Becker, B., Luo, L., Gao, S., Gong, Q., et al. (2015). Oxytocin selectively facilitates learning with social feedback and increases activity and functional connectivity in emotional memory and reward processing regions. *Hum. Brain Mapp.* 36, 2132–2146. doi: 10.1002/hbm.22760

Huang, Y., Kendrick, K. M., and Yu, R. (2014). Conformity to the opinions of other people lasts for no more than 3 days. *Psychol. Sci.* 25, 1388–1393. doi: 10.1177/0956797614532104

Huang, Y., Kendrick, K. M., Zheng, H., and Yu, R. (2015). Oxytocin enhances implicit social conformity to both in-group and out-group opinions. *Psychoneuroendocrinology* 60, 114–119. doi: 10.1016/j.psyneuen.2015.06.003

Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., et al. (2010). Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. *J. Neurosci.* 30, 4999–5007. doi: 10.1523/JNEUROSCI.5538-09.2010

Ide, J. S., Nedic, S., Wong, K. F., Strey, S. L., Lawson, E. A., Dickerson, B. C., et al. (2018). Oxytocin attenuates trust as a subset of more general reinforcement learning, with altered reward circuit functional connectivity in males. *Neuroimage* 174, 35–43. doi: 10.1016/j.neuroimage.2018.02.035

Israel, S., Lerer, E., Shalev, I., Uzefovsky, F., Riebold, M., Laiba, E., et al. (2009). The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLoS One* 4:e5535. doi: 10.1371/journal.pone.0005535

Kendrick, K. M. (2000). Oxytocin, motherhood and bonding. *Exp. Physiol.* 85, 111–124. doi: 10.1111/j.1469-445X.2000.tb00014.x

Kendrick, K. M., Guastella, A. J., and Becker, B. (2017). "Overview of human oxytocin research," in *Behavioral Pharmacology of Neuropeptides: Oxytocin*, eds R. Hurlemann and V. Grinevich (Cham: Springer), 321–348. doi: 10.1007/7854_2017_19

Keri, S., Kiss, I., and Kelemen, O. (2009). Sharing secrets: oxytocin and trust in schizophrenia. *Soc. Neurosci.* 4, 287–293. doi: 10.1080/17470910802319710

Kessner, S., Sprenger, C., Wrobel, N., Wiech, K., and Bingel, U. (2013). Effect of oxytocin on placebo analgesia: a randomized study. *JAMA* 310, 1733–1735. doi: 10.1001/jama.2013.277446

Koscik, T. R., and Tranel, D. (2011). The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia* 49, 602–611. doi: 10.1016/j.neuropsychologia.2010.09.023

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

Krueger, F., Parasuraman, R., Iyengar, V., Thornburg, M., Weel, J., Lin, M., et al. (2012). Oxytocin receptor genetic variation promotes human trust behavior. *Front. Hum. Neurosci.* 6:4. doi: 10.3389/fnhum.2012.00004

Kruppa, J. A., Gossen, A., Oberwelland Weiß, E., Kohls, G., Großheinrich, N., Cholemkery, H., et al. (2018). Neural modulation of social reinforcement learning by intranasal oxytocin in male adults with high-functioning autism spectrum disorder: a randomized trial. *Neuropsychopharmacology* [Epub ahead of print]. doi: 10.1038/s41386-018-0258-7

Lambert, B., Declerck, C. H., and Boone, C. (2014). Oxytocin does not make a face appear more trustworthy but improves the accuracy of trustworthiness judgments. *Psychoneuroendocrinology* 40, 60–68. doi: 10.1016/j.psyneuen.2013.10.015

Lane, A., Mikolajczak, M., Treinen, E., Samson, D., Corneille, O., de Timary, P., et al. (2015). Failed replication of oxytocin effects on trust: the envelope task case. *PLoS One* 10:e0137000. doi: 10.1371/journal.pone.0137000

Leng, G., and Ludwig, M. (2016). Intranasal oxytocin: myths and delusions. *Biol. Psychiatry* 79, 243–250. doi: 10.1016/j.biopsych.2015.05.003

Lourenco, F. S., Decker, J. H., Pedersen, G. A., Dellarco, D. V., Casey, B. J., and Hartley, C. A. (2015). Consider the source: adolescents and adults similarly follow older adult advice more than peer advice. *PLoS One* 10:e0128047. doi: 10.1371/journal.pone.0128047

Luo, R., Xu, L., Zhao, W., Ma, X., Xu, X., Kou, J., et al. (2017). Oxytocin facilitation of acceptance of social advice is dependent upon the perceived trustworthiness of individual advisors. *Psychoneuroendocrinology* 83, 1–8. doi: 10.1016/j.psyneuen.2017.05.020

Ma, X., Luo, L., Geng, Y., Zhao, W., Zhang, Q., and Kendrick, K. M. (2014). Oxytocin increases liking for a country's people and national flag but not for other cultural symbols or consumer products. *Front. Behav. Neurosci.* 8:266. doi: 10.3389/fnbeh.2014.00266

MacDonald, K., MacDonald, T. M., Brüne, M., Lamb, K., Wilson, M. P., Golshan, S., et al. (2013). Oxytocin and psychotherapy: a pilot study of its physiological, behavioral and subjective effects in males with depression. *Psychoneuroendocrinology* 38, 2831–2843. doi: 10.1016/j.psyneuen.2013.05.014

Marsh, N., Scheele, D., Feinstein, J. S., Gerhardt, H., Strang, S., Maier, W., et al. (2017). Oxytocin-enforced norm compliance reduces xenophobic outgroup rejection. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9314–9319. doi: 10.1073/pnas.1705853114

Mende-Siedlecki, P., Said, C. P., and Todorov, A. (2012). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Soc. Cogn. Affect. Neurosci.* 8, 285–299. doi: 10.1093/scan/nsr090

Merolla, J. L., Burnett, G., Pyle, K. V, Ahmadi, S., and Zak, P. J. (2013). Oxytocin and the biological basis for interpersonal and political trust. *Polit. Behav.* 35, 753–776. doi: 10.1007/s11109-012-9219-8

Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., and Luminet, O. (2010a). Oxytocin makes people trusting, not gullible. *Psychol. Sci.* 21, 1072–1074. doi: 10.1177/0956797610377343

Mikolajczak, M., Pinon, N., Lane, A., de Timary, P., and Luminet, O. (2010b). Oxytocin not only increases trust when money is at stake, but also when confidential information is in the balance. *Biol. Psychol.* 85, 182–184. doi: 10.1016/j.biopsycho.2010.05.010

Moretto, G., Sellitto, M., and di Pellegrino, G. (2013). Investment and repayment in a trust game after ventromedial prefrontal damage. *Front. Hum. Neurosci.* 7:593. doi: 10.3389/fnhum.2013.00593

Nave, G., Camerer, C., and McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspect. Psychol. Sci.* 10, 772–789. doi: 10.1177/1745691615600138

Ne'eman, R., Perach-Barzilay, N., Fischer-Shofty, M., Atias, A., and Shamay-Tsoory, S. G. (2016). Intranasal administration of oxytocin increases human aggressive behavior. *Horm. Behav.* 80, 125–131. doi: 10.1016/j.yhbeh.2016.01.015

Ostrom, E., and Walker, J. (2003). *Trust and Reciprocity: Interdisciplinary Lessons for Experimental Research*. New York, NY: Russell Sage Foundation.

Quintana, D. S., Westlye, L. T., Rustan, Ø. G., Tesli, N., Poppy, C. L., Smevik, H., et al. (2015). Low-dose oxytocin delivered intranasally with Breath Powered device affects social-cognitive behavior: a randomized four-way crossover trial with nasal cavity dimension assessment. *Transl. Psychiatry* 5:e602. doi: 10.1038/tp.2015.93

Righetti, F., and Finkenauer, C. (2011). If you are able to control yourself, I will trust you: the role of perceived self-control in interpersonal trust. *J. Pers. Soc. Psychol.* 100:874. doi: 10.1037/a0021827

Scheele, D., Striepens, N., Kendrick, K. M., Schwering, C., Noelle, J., Wille, A., et al. (2014). Opposing effects of oxytocin on moral judgment in males and females. *Hum. Brain Mapp.* 35, 6067–6076. doi: 10.1002/hbm.22605

Schilbach, L., Eickhoff, S. B., Schultze, T., Mojzisch, A., and Vogeley, K. (2013). To you I am listening: perceived competence of advisors influences judgment and decision-making via recruitment of the amygdala. *Soc. Neurosci.* 8, 189–202. doi: 10.1080/17470919.2013.775967

Shalvi, S., and De Dreu, C. K. W. (2014). Oxytocin promotes group-serving dishonesty. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5503–5507. doi: 10.1073/pnas.1400724111

Shamay-Tsoory, S. G., and Abu-Akel, A. (2016). The social salience hypothesis of oxytocin. *Biol. Psychiatry* 79, 194–202. doi: 10.1016/j.biopsych.2015.07.020

Shamay-Tsoory, S. G., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N., and Levkovitz, Y. (2009). Intranasal administration of oxytocin increases envy and schadenfreude (gloating). *Biol. Psychiatry* 66, 864–870. doi: 10.1016/j.biopsych.2009.06.009

Sindermann, C., Luo, R., Becker, B., Kendrick, K. M., and Montag, C. (2018). Oxytocin promotes lying for personal gain in a genotype-dependent manner. *bioRxiv* [Preprint]. doi: 10.1101/361212

Stallen, M., De Dreu, C. K. W., Shalvi, S., Smidts, A., and Sanfey, A. G. (2012). The herding hormone: oxytocin stimulates in-group conformity. *Psychol. Sci.* 23, 1288–1292. doi: 10.1177/0956797612446026

Striepens, N., Kendrick, K. M., Maier, W., and Hurlemann, R. (2011). Prosocial effects of oxytocin and clinical evidence for its therapeutic potential. *Front. Neuroendocrinol.* 32:426–450. doi: 10.1016/j.yfrne.2011.07.001

Striepens, N., Matusch, A., Kendrick, K. M., Mihov, Y., Elmenhorst, D., Becker, B., et al. (2014). Oxytocin enhances attractiveness of unfamiliar female faces independent of the dopamine reward system. *Psychoneuroendocrinology* 39, 74–87. doi: 10.1016/j.psyneuen.2013.09.026

Suen, V. Y. M., Brown, M. R. G., Morck, R. K., and Silverstone, P. H. (2014). Regional brain changes occurring during disobedience to "Experts" in financial decision-making. *PLoS One* 9:e87321. doi: 10.1371/journal.pone.0087321

Tauber, M., Mantoulan, C., Copet, P., Jauregui, J., Demeer, G., Diene, G., et al. (2011). Oxytocin may be useful to increase trust in others and decrease disruptive behaviours in patients with Prader-Willi syndrome: a randomised placebo-controlled trial in 24 patients. *Orphanet J. Rare Dis.* 6:47. doi: 10.1186/1750-1172-6-47

Theodoridou, A., Rowe, A. C., Penton-Voak, I. S., and Rogers, P. J. (2009). Oxytocin and social perception: oxytocin increases perceived facial trustworthiness and attractiveness. *Horm. Behav.* 56, 128–132. doi: 10.1016/j.yhbeh.2009.03.019

Todorov, A., Olivola, C. Y., Dotsch, R., and Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* 66, 519–545. doi: 10.1146/annurev-psych-113011-143831

Van Honk, J., Eisenegger, C., Terburg, D., Stein, D. J., and Morgan, B. (2013). Generous economic investments after basolateral amygdala damage. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2506–2510. doi: 10.1073/pnas.1217316110

Van IJzendoorn, M. H., and Bakermans-Kranenburg, M. J. (2012). A sniff of trust: meta-analysis of the effects of intranasal oxytocin administration on face recognition, trust to in-group, and trust to out-group. *Psychoneuroendocrinology* 37, 438–443. doi: 10.1016/j.psyneuen.2011.07.008

Van't Wout, M., and Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition* 108, 796–803. doi: 10.1016/j.cognition.2008.07.002

Woolley, J. D., Chuang, B., Fussell, C., Scherer, S., Biagianti, B., Fulford, D., et al. (2017). Intranasal oxytocin increases facial expressivity, but not ratings of trustworthiness, in patients with schizophrenia and healthy controls. *Psychol. Med.* 47, 1311–1322. doi: 10.1017/S0033291716003433

Xu, L., Becker, B., Luo, R., Zheng, X., Zhao, W., Zhang, Q., et al. (2018). Oxytocin amplifies sex differences in human mate choice. *bioRxiv* [Preprint]. doi: 10.1101/416198

Xu, X., Liu, C., Zhou, X., Chen, Y., Gao, Z., Zhou, F., et al. (2018). Oxytocin facilitates self-serving rather than altruistic tendencies in competitive social interactions via orbitofrontal cortex. *bioRxiv* [Preprint]. doi: 10.1101/501171

Yao, S., Zhao, W., Cheng, R., Geng, Y., Luo, L., and Kendrick, K. M. (2014). Oxytocin makes females, but not males, less forgiving following betrayal of trust. *Int. J. Neuropsychopharmacol.* 17, 1785–1792. doi: 10.1017/S146114571400090X

Yi, L., Pan, J., Fan, Y., Zou, X., Wang, X., and Lee, K. (2013). Children with autism spectrum disorder are more trusting than typically developing children. *J. Exp. Child Psychol.* 116, 755–761. doi: 10.1016/j.jecp.2013.05.005

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

Zhao, W., Becker, B., Yao, S., Ma, X., Kou, J., and Kendrick, K. M. (2018a). Oxytocin enhancement of the placebo effect may be a novel therapy for working memory impairments. *Psychother. Psychosom.* doi: 10.1159/000495260

Zhao, W., Ma, X., Le, J., Ling, A., Xin, F., Kou, J., et al. (2018b). Oxytocin biases men to be more or less tolerant of others' dislike dependent upon their relationship status. *Psychoneuroendocrinology* 88, 167–172. doi: 10.1016/j.psyneuen.2017.12.010

Zhong, S., Monakhov, M., Mok, H. P., Tong, T., San Lai, P., Chew, S. H., et al. (2012). U-shaped relation between plasma oxytocin levels and behavior in the trust game. *PLoS One* 7:e51095. doi: 10.1371/journal.pone.0051095

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 11 February 2019
doi: 10.3389/fnins.2019.00089

Following the Majority: Social Influence in Trusting Behavior

Zhenyu Wei[1,2], *Zhiying Zhao*[1] *and Yong Zheng*[2]*

[1] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, [2] Key Laboratory of Cognition and Personality (MOE), Faculty of Psychology, Southwest University, Chongqing, China

When making decisions, people may change their behavior, sometimes against their personal preference, according to the opinions of peers. However, the effect of social influence on trust is still unknown. In our study, we used the event-related functional magnetic resonance imaging to investigate brain activity in social influence during a trust game. The behavioral results revealed that people tend to conform to others' opinions and behaviors in a trust game. Decreased activations were observed in superior temporal gyrus during processing of social influences. Moreover, brain regions supporting value processing and reward learning were activated when subjects decided to follow the majority. These regions include the ventral medial prefrontal cortex, ventral striatum, and parahippocampal gyrus. Finally, our exploratory analysis revealed an increase in functional connectivity between the prefrontal cortex and the ventral striatum during conformity in trusting behavior. These findings indicate that the neural basis of social influence in trusting behavior are similar to the mechanisms implicated in reward learning. The brain regions involved in reward learning might reflect the reward value of agreeing with others in our study.

Keywords: social influence, trust game, superior temporal gyrus, ventral striatum, reward learning

OPEN ACCESS

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

Reviewed by:
Delin Sun,
Duke University, United States
Ruolei Gu,
Institute of Psychology (CAS), China

Correspondence:
Yong Zheng
zhengy@swu.edu.cn

Specialty section:
This article was submitted to Decision Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *09 October 2018*
Accepted: *25 January 2019*
Published: *11 February 2019*

Citation:
Wei Z, Zhao Z and Zheng Y (2019) Following the Majority: Social Influence in Trusting Behavior. Front. Neurosci. 13:89. doi: 10.3389/fnins.2019.00089

INTRODUCTION

Our opinions and behaviors are often affected by the majority (Asch, 1956; Turner, 1991). People tend to change their opinions and behaviors in order to follow with social norms, even if the majority decision is against their personal preference (Cialdini and Goldstein, 2004; Morgan and Laland, 2012; Haun et al., 2013). Psychologists defined this phenomenon as "social conformity." It refers to individuals' action of adopting the opinions, behaviors, and judgments of others (Turner, 1991). Asch (1951) used a simple line judgment task to investigate social conformity. Since then social psychologists began to explore the causes of social conformity. Based on previous study, there are three types of intrinsic motivations underlying social conformity: a desire to obtain social approval of others, a desire to make a correct choice, and a desire to keep a positive self-concept (Cialdini and Goldstein, 2004).

Recent studies have investigated the effect of conformity in many judgment tasks as well as the neural basis of conformity. By using mental rotation task and music rating task, Berns et al. (2005, 2010) found that the opinions of peers could change participants' initial judgments and affect neural activity within relatively low-level processing brain areas related to each task. In addition, previous literatures have reported that the brain regions associated with reward processing and behavioral adjustment were closely associated with social influence. Mason et al. (2009) exposed subjects to popular, unpopular and novel symbols and reported that the medial prefrontal cortex (mPFC) was involved in normative social influence by comparing socially and not socially marked symbols, while the striatum (the caudate) might be a possible index of informational social influence by comparing popular and unpopular symbols. Wei et al. (2013) also found that confliction with group norms during an ultimatum game activated the bilateral insula, bilateral middle frontal gyrus (MFG) and mPFC. Additionally, Klucharev et al. (2009) found that conflicting group opinions triggered a neuronal response in the nucleus accumbens and the rostral cingulate zone (RCZ). These brain regions are often associated with reward processing and behavioral adjustment, which is similar to prediction error signal (Berns et al., 2001; Holroyd and Coles, 2002; Ridderinkhof et al., 2004). Neural activity in these regions could predict participants' subsequent conforming behaviors (Klucharev et al., 2009). By using stock task and music choice task, Burke et al. (2010) and Campbell-Meiklejohn et al. (2010) found that neural activity in the ventral striatum was involved in social influence, suggesting that the opinions of others could modulate the basic value signals in known reinforcement learning neural circuitry (Campbell-Meiklejohn et al., 2010).

Conformity effect was also found in economic decisions, such as ultimatum game (Wei et al., 2013), dictator game (Wei et al., 2017), risk taking (Gardner and Steinberg, 2005), stock market participation (Hong et al., 2004), consuming decision and investment decision (Bursztyn et al., 2014). These results indicated that the opinion of majority could influence people's own preferences in economic decision context. Trust plays an important role in economic decision interactions (Cochard et al., 2004). Previous study suggested that, for the trusting behaviors, genetics only explain about 20% of the cross-sectional variation while environmental factors would explain 80% of the variation (Cesarini et al., 2008; Ahern et al., 2014). One potential environment factor is social conformity. Prior studies have found that individuals tended to change their rating of trustworthiness toward social norm in a trustworthiness judgment task (Campbell-Meiklejohn et al., 2012; Simonsen et al., 2014). In present study, we used trust game to explore whether peers' decision could change the choices of individuals. Trust game is widely used to measure trusting behavior. There are two players in the classic trust game: an investor and a trustee. Both players are endowed with $10. The investor decides whether give the money to the trustee. If the investor gives the money to the trustee, the endowment would be multiplied by experimenter then. In the end, the trustee decides whether to give any portion of the money she/he received back to the investor or just keep it. In our study, we developed a modified trust game. In this task, participants were able to see peer' choices when they made the trust decision.

Firstly, we hypothesized that the choices of the majority would affect subjects' trust preference. Subjects may invest the money to the trustee when they see that the majority of the group trusts the trustee. Conversely, participants may distrust the trustee if they see that the majority does not trust the trustee. Otherwise, subjects will insist on their own trust preferences if social influence has no effects on trust decision. Secondly, we predicted that participants may conform to the opinion of the majority with a relatively high level of decision confidence, since they may have high reward expectancy in the trust social influence condition. Finally, previous literatures had reported that social influence might affect participants' behaviors through the neural underpinnings of reward learning and behavioral adjustment, such as ventral medial prefrontal cortex (vmPFC) and anterior cingulate cortex (ACC), and also brain structures underlying social reward processing especially the striatum (Izuma et al., 2008; Mason et al., 2009; Klucharev et al., 2011; Wei et al., 2013). Therefore, we hypothesized that the activity in brain reward circuits such as the vmPFC and caudate may be associated with social influence. Recent brain imaging studies have suggested evidence that enhanced functional connectivity between the prefrontal cortex and ventral striatum during reward processing (Camara et al., 2008). Hence, we hypothesized that a psychophysiological interaction (PPI) analysis may confirm an enhanced functional connectivity between the prefrontal cortex and ventral striatum during conformity in the trust social influence condition.

MATERIALS AND METHODS

Participants

Twenty-seven healthy right-handed participants (mean age = 21.1, female = 16, male = 11) participated in the experiment. These participants were recruited from Southwest University through advertisements in the online student forums, none of them came from department of psychology or economics. All were native Mandarin speakers, with no neurological illness as confirmed by psychiatric clinical assessment or psychological disorders, and with (corrected to) normal vision. Written informed consent was obtained in accordance with the regulations of the Ethics Committee of Southwest University. This study was approved by the Ethics Committee of Southwest University.

Stimulus Materials

Peers' choices were presented in the form of a table to the participants. The number "1" refers to a choice to send the endowment to the stranger and the number "2" indicates a choice to keep the endowment. There were four conditions of social influence: trust influence (three or four group members decided to send the endowment to the stranger); moderate (two group members decided to trust the stranger while the other two decided keep their endowments); distrust influence (three or four group members decided to keep the endowment); and no information (the boxes corresponding to each group members' choices were replaced with "×"). There were 70 offers in total. The offer stimuli consisted of the number of the trustee (randomly from 1 to 70), the choices available, and the social information (peers' choices). The former was presented in the upper portion of the screen. The choices available were presented in the center of the screen and the latter in the lower part of the picture.

Experimental Procedures

Participants were told that they would play an on-line monetary game with four other participants, who would be in a separate behavioral laboratory. They would see the choices of the other peers on the computer screen during the decision phase of the experiment. Participants acted as an investor and play the game independently with 70 different strangers (trustees). These trustees were randomly selected from the university and played the game on the other floor. Participants and their group members did not know anything about these seventy trustees. At the beginning of each trial, both players (investor and trustee) were endowed with ￥10. The investor was asked to decide whether to send the endowment. The endowment would be tripled if the investor decided to invest. Then the trustee was asked to decide whether to send half of the money back (￥15). The investor would not know the outcome (i.e., trustees' choice) during the task. Subjects were told that they will receive ￥50 for participating in the experiment plus the additional money earned from ten of their trust decisions, chosen at random, in the trust game. Subjects earned on average about ￥60 for their participated in the experiment which was not based on investment outcome. We asked participants whether she/he believed the existence of trustees after they finished the task. All the participants reported that they believed the existence of trustees. After the data of all the participants were collected, participants received payment and were told that the peers and trustees did not exist.

Participants then received details about the procedure of the experiment. At the beginning of each trial, they saw a fixation point for a 2–4 s jittered duration that varied pseudo-randomly. Then, the decision screen was presented for 3 s. They used the index and middle fingers of their right hands to separately respond to the offer by pressing one of two buttons on an MRI-compatible button box ("1" to invest and "2" to keep the endowment). Peers' choices were placed in the lower part of the decision interface. Subsequently, confidence ratings were provided for 2 s. Finally, the word "next" displayed for 1 s, indicating that the next trial was about to begin. The sequence of events in a trial is illustrated in **Figure 1**.

There were seventy trials in present experiment. The duration of a trial is approximately 9 seconds. In 10 of the trials, participants were informed that two peers decided to send the money to the trustee while the other two decided to keep the endowments. These trials were used solely to maintain the believability of the interaction between the participant and the four peers. They were excluded in the final analysis. In one-third of the remaining trials (20 trials), participants could not see the group's choices (the no information, or baseline condition; we told participants that the decisions in these trials were not made by all the four peers). For the 20 trials of the trust influence condition, three or four peers' choices were to send the endowments to the trustee. For the 20 trials of the distrust influence condition, one or none of the group members decided to invest. Before performing the task in the scanner, all participants completed a training session. They were told that the computer for the pre-experiment training is not connected to the local network, therefore they could not receive anything information about the peers' choices.

We used a PC running E-Prime 2.0 to display the stimuli and acquire the responses of the participants, as well as the reaction times (RTs). In the scanner, there was a mirror placed on the top of the image acquisition coil. Participants saw the experiment task via this mirror that reflected the screen mounted at the back of the scanner.

Image Acquisition

Functional MRI data were acquired using a 3T Siemens Trio scanner. Each scan contains 355 functional volumes, using an echo-planar imaging (EPI) sequence with the following parameters: TR/TE = 2000/30 ms, flip angle = 90°, acquisition matrix = 64 × 64, FOV = 192 mm × 192 mm, axial slices = 32, slice thickness/gap = 3mm/1 mm, voxel size = 3 mm × 3 mm × 3 mm. The first three images were discarded for the saturation effect.

Data Analysis

Behavioral Data Analysis

We used statistical product and service solutions (SPSS) to analyze the behavioral data. We predicted that the choices of the majority may influence participants' decision. A repeated measure (social influence: baseline, trust influence, distrust

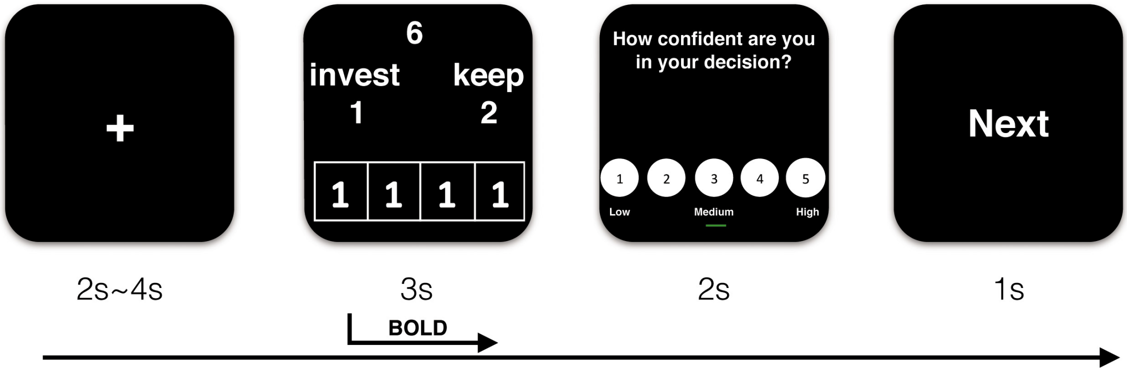


FIGURE 1 | Demonstration of sequence of events in a trial (take trust influence condition for example).

influence) ANOVA was used to analyze the RTs in the decision phase, as well as the rate of trust. Since we predicted that subjects may have high reward expectancy in the trust social influence condition, we conducted a 3 (social influence: baseline, trust influence, distrust influence) × 2 (choices: trust, distrust) ANOVA on the mean confidence rating.

fMRI Data Analysis

Image preprocessing was performed with statistical parametric mapping 8 (SPM8; Welcome Department of Imaging Neuroscience, University of London, United Kingdom). Functional images were first corrected for motion artifacts. Then images were interpolated to correct for slice timing, and spatially normalized into the Montreal Neurological Institute (MNI)-space using the SPM8 EPI template, and resampled into 3 mm × 3 mm × 3 mm voxels. Images were smoothed using an 8 mm^3 full-width-at-half-maximum (FWHM) Gaussian kernel. A 0.01 Hz–0.08 Hz band-pass filter, which was composed of a discrete cosine-basis function with a cutoff period of 128 s for the high-pass filter was applied to the time courses of all brain voxels.

We conducted analysis on functional magnetic resonance imaging data of the decision phase. General linear model analysis was performed with SPM8. Three regressors were entered based on social information (baseline, trust influence and distrust influence). These regressors were then convolved with the standard hemodynamic response function. In addition, the realignment parameters were included in the model to regress out potential movement artifacts. For a whole-brain analysis, the result was thresholded at $p < 0.05$ (FDR correction), cluster size > 10. The effect of social influence was estimated by contrasting the trust influence effect (*trust influence condition > no information*). For more detailed insights into the neural mechanisms underlying social conformity in trusting behavior, we did an exploratory analysis, analyzed the conforming behavior contrast (*conformity vs. non-conformity*) in trust influence condition (*trust influence condition – conformity > trust influence condition – non-conformity*). Activations in this analysis were thresholded at $p < 0.05$ (FDR correction), cluster size > 10.

Finally, an exploratory PPI analysis was performed in order to identify brain regions that showed significantly increased coordination (i.e., increased functional connectivity) with the ventral striatum activity related to conformity compared to non-conformity in the trust influence condition (Friston et al., 1997). Based on our fMRI results and previous literature, the region of interest (ROI) was defined as a sphere with 6-mm-radius centered at the peak voxel in the ventral striatum (MNI coordinates: [10, 18, -9]) (Campbell-Meiklejohn et al., 2010). The time series was extracted from each subject in the ventral striatum. And the PPI regressor was calculated as the element-by-element product of the mean-corrected activity of ROI and a vector coding for differential task effects of conformity-trust influence versus non-conformity-trust influence. The PPI regressors reflected the interaction between psychological variable (*trust influence condition - conformity > trust influence condition – non-conformity*) and the activation time course of the ventral striatum. Individual contrast images for conformity-trust influence versus non-conformity-trust influence were computed and entered into second-level one-sample *t*-tests. Brain regions surviving the cluster-extent based threshold $p < 0.05$ (FDR correction, with a primary voxel-level threshold of $p < 0.001$) were considered significant.

RESULTS

Behavioral Results

Data from twenty-seven subjects entered the behavioral analysis. We used a one-way repeated measures (social influence: baseline, trust influence, distrust influence) ANOVA to analyze the RTs in the decision phase. The effect of social influence was significant, $F(2,25) = 4.204$, $p < 0.05$. Participants responded faster in the trust influence condition ($M = 1222.46$ ms, $SD = 312.76$) than in the baseline condition ($M = 1328.58$ ms, $SD = 333.87$), $t_{(26)} = -2.845$, $p < 0.01$. The responses were also faster in the trust influence condition ($M = 1222.46$ ms, $SD = 312.76$) than in the distrust condition ($M = 1294.6$ ms, $SD = 294.42$), $t_{(26)} = -2.479$, $p < 0.05$.

Regarding the subjects' choices, a one-way repeated measures (social influence: baseline, trust influence, distrust influence) ANOVA was used to analyze the rate of trust in the decision phase. The effect of social influence was significant, $F(2,25) = 7.714$, $p < 0.01$. Subjects decided to trust the trustee at a significantly higher rate in the trust influence condition ($M = 0.72$, $SD = 0.2$) than in the baseline condition ($M = 0.53$, $SD = 0.22$), $t_{(26)} = 3.543$, $p < 0.01$. We also found this phenomenon in the contrast between trust influence condition ($M = 0.72$, $SD = 0.2$) and distrust influence condition ($M = 0.43$, $SD = 0.27$), $t_{(26)} = 3.926$, $p < 0.001$. Participants chose to trust the trustee at a significantly higher rate in the baseline condition ($M = 0.53$, $SD = 0.22$) than in the distrust influence condition ($M = 0.43$, $SD = 0.27$), $t_{(26)} = 2.074$, $p < 0.05$.

Because we predicted that subjects may have high reward expectancy in the trust social influence condition, we hypothesized that participants may conform to the opinion of the majority with a relatively high level of decision confidence. We conducted a 3 (social influence: baseline, trust influence, distrust influence) × 2 (choices: trust, distrust) ANOVA on the mean confidence rating. As predicted, the interaction between social influence and choices was significant, $F(2,25) = 9.202$, $p < 0.001$. The level of decision confidence is higher in the trust influence-trust condition ($M = 3.78$, $SD = 0.52$) than in the baseline-trust condition ($M = 3.44$, $SD = 0.83$), $t_{(26)} = 2.632$, $p < 0.05$, as well as in the distrust influence-trust condition ($M = 3.37$, $SD = 0.74$), $t_{(26)} = 3.227$, $p < 0.01$. Confidence ratings for the trust influence-trust condition ($M = 3.78$, $SD = 0.52$) seemed to be overall higher than ratings for the trust influence-distrust condition ($M = 3.37$, $SD = 0.61$), $t_{(26)} = 3.827$, $p < 0.001$.

fMRI Results

We compared the neural activity in trust influence condition with baseline condition and found significantly greater deactivation in superior temporal gyrus (STG) (for more details see **Table 1** and

TABLE 1 | Significant activation clusters for trust social influence.

Brain regions	HEM	x	y	z	No. of voxels	*t*-value
Trust influence > Baseline						
Activation						
No Cluster						
Deactivation						
STG	R	60	−45	9	28	5.97

Voxels were selected for p < 0.05, cluster size > 10, FDR correction. HEM, hemisphere; STG, superior temporal gyrus.

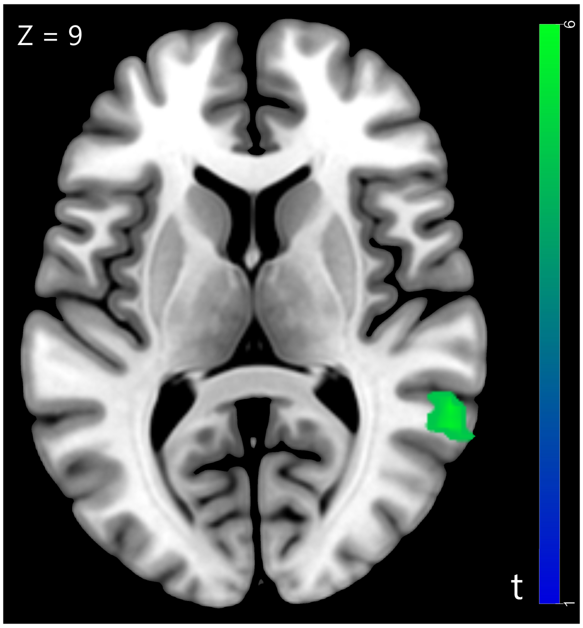


FIGURE 2 | The superior temporal gyrus was involved in trust influence condition (Trust influence > Baseline), $p < 0.05$, cluster size = 10, FDR correction.

Figure 2). The STG is a key brain region that involved in the cognitive capacity of perspective taking (Frith and Frith, 2003).

To capture the neural mechanisms underlying conformity effect in trusting behavior, exploratory analyses were performed. We compared the trust influence-conformity trials (mean number of trials 14) to trust influence-non-conformity (mean number of trials 6). Results shown that the trust influence which successfully induced conformity in trusting behavior activated the brain regions such as bilateral parahippocampal gyrus, vmPFC, RCZ, ACC/ caudate, middle occipital gyrus (MOG), MFG, middle temporal gyrus (MTG), postcentral gyrus and inferior parietal lobule (IPL) (see **Table 2** and **Figure 3** for more details). Comparison of activity in non-conformity trials with conformity trials did not show any significant activation.

Moreover, psychophysiological interaction (PPI) analysis showed that activity in the ventral striatum was accompanied by task-dependent (conformity > non-conformity) functional interaction with brain areas: STG, superior frontal gyrus (SFG), MTG and inferior temporal gyrus (ITG). The opposite contrast did not reveal any significant changes in functional connectivity (see **Table 3** and **Figure 4** for more details).

TABLE 2 | Significant activation clusters for conformity in trusting behavior.

Brain regions	HEM	x	y	z	No. of voxels	*t*-value
MFG	L	−36	42	42	16	3.8
MTG	R	60	−63	−9	29	4.9
MOG	L	−51	−81	3	29	3.98
RCZ	R	3	−3	39	23	3.77
ACC/Caudate	L	−9	27	−18	43	5.64
vmPFC	L	−6	51	−18	137	5.36
IPL	R	57	−30	30	32	3.92
Postcentral gyrus	R	60	−12	48	210	4.88
Parahippocampal gyrus	L	−9	−87	30	11	5.4
Parahippocampal gyrus	R	39	−6	−36	38	7.96

Voxels were selected for p < 0.05, cluster size > 10, FDR correction. HEM, hemisphere; MFG, middle frontal gyrus; MTG, middle temporal gyrus; MOG, middle occipital gyrus; RCZ, rostral cingulate zone; ACC, anterior cingulate cortex; vmPFC, ventral medial prefrontal cortex; IPL, inferior parietal lobule.

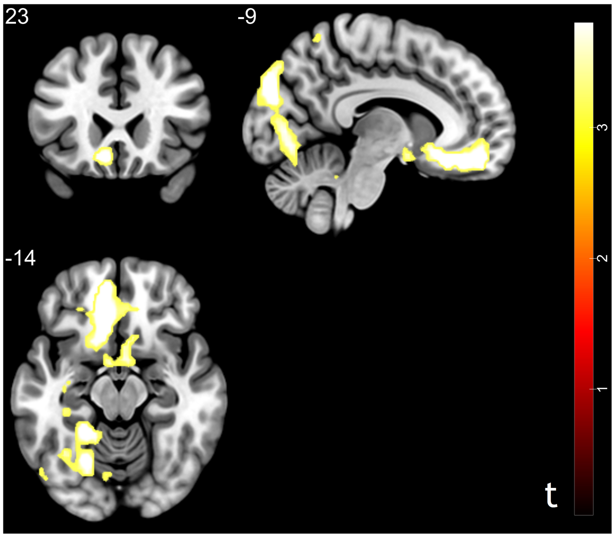


FIGURE 3 | Brain regions correlated with social influence in trusting behavior (trust influence – conformity > trust influence – non-conformity). Significant activations in middle frontal gyrus, middle temporal gyrus, middle occipital gyrus, rostral cingulate zone, anterior cingulate cortex, ventral medial prefrontal cortex, and inferior parietal lobule. $p < 0.05$, cluster size = 10, FDR correction.

DISCUSSION

In the present study, we used psychological and neuroscientific methods to investigate the impact of social influence on trust. We found that individuals are likely to conform to the opinions of their peers in a trust game. The rate of trust was higher when participants found that the majority of group members trusted the trustee compared to in the baseline condition. Conversely,

TABLE 3 | Results of psychophysiological interaction (PPI) analysis.

Brain regions	HEM	x	y	z	No. of voxels	*t*-value
STG	L	−51	−63	21	87	4.74
SFG	L	−18	48	51	48	4.83
MTG	R	57	−24	−9	51	5.16
ITG	L	−57	−18	−27	53	4.78

Voxels were selected for $p < 0.05$, FDR cluster-level correction with an initial peak-level threshold $p < 0.001$. HEM, hemisphere; STG, superior temporal gyrus; SFG, superior frontal gyrus; MTG, middle temporal gyrus; ITG, inferior temporal gyrus.

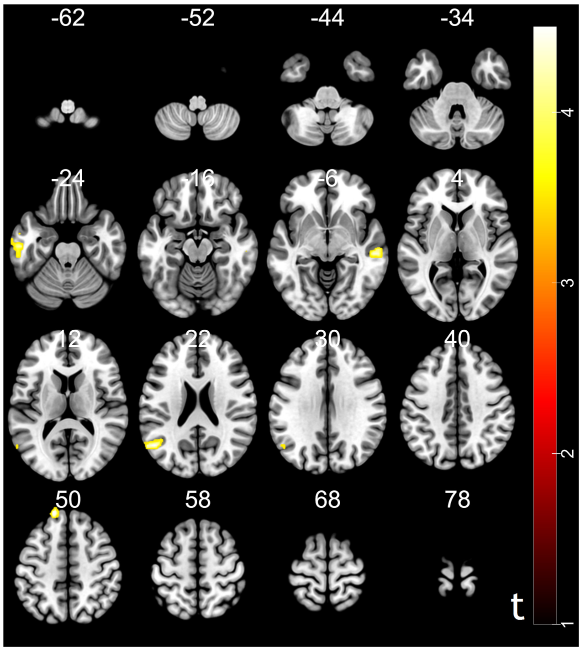


FIGURE 4 | Results of psychophysiological interaction (PPI) analysis. The region of interest was ventral striatum, MNI coordinates: [10, 18, −9]. Functional connectivity with the ventral striatum (conformity > non-conformity) in the trust influence condition. Voxels were selected for $p < 0.05$, FDR cluster-level correction with an initial peak-level threshold $p < 0.001$.

the rate of trust was lower when participants saw that most group members decided to keep the endowment (distrust) compared to in the baseline condition. In addition, participants conformed to the opinion of the majority with relatively high levels of decision confidence in the trust influence condition.

Functional imaging data suggested that the STG, a brain region involved in perspective-taking, was decreased when participants made decision in the trust influence condition comparing with the baseline condition. The activity of STG is associated with perspective taking, which can be termed as theory of mind (Frith and Frith, 2003). As the decision to trust is concerned with perspective-taking, it should activate brain regions involved in theory-of-mind tasks (Fehr and Camerer, 2007). Moreover, researchers found the STG was involved in the processing of gaze direction in a modified trust game (Sun et al., 2018). A previous study that focused on the neurobiological correlates of conformity during mental rotation task has reported that the presence of external information was associated with decreased activation in the mental rotation neural network (Berns et al., 2005). They inferred that the external information relieved the mental rotation processing load (Berns et al., 2005). Similarly, decreased activations were observed during trust game in STG when external information was presented in our study. This result might suggest that external trust information affected neural activity in brain regions associated with trust game, which relieved the perspective-taking process in the game.

In our study, we tried to capture the conformity effect in the imaging data and found that brain regions involved in reward learning such as the vmPFC, ACC, ventral striatum, parahippocampal gyrus, and RCZ were also related with social influence in trusting behavior. The vmPFC has been previously implicated in processing reward expectations and computing the subjective value of multiple reward types (Rushworth et al., 2009, 2011; Rangel and Hare, 2010; Grabenhorst and Rolls, 2011). The study of brain activity during decision-making suggested that fictive reward signals (rewards that could have been, but were not directly received) have been represented in the ACC (Hayden et al., 2009). The RCZ is engaged when the need for adjustments to achieve action goals becomes evident (Ridderinkhof et al., 2004). Previous studies have demonstrated that the caudate is involved in gain prediction in response to reward cues and implicated in reward processing, social learning, and reciprocate cooperation (Rilling et al., 2002, 2004; McCoy and Platt, 2005; Knutson and Wimmer, 2007). According to PPI results, we found possible enhanced functional connectivity between the ventral striatum and prefrontal cortex during conformity compared to non-conformity in trusting behavior. Notably, recent research demonstrated that increased functional connectivity between the ventral striatum and prefrontal cortex was related to reward processing (Frank and Claus, 2006; Camara et al., 2008, 2009; van den Bos et al., 2012). Taken together, these exploratory imaging results suggest that the underlying mechanisms of social influence in trusting behavior may be similar to those implicated in reward learning. Agreement with the other group members might predict future acceptance from peer, which can also activate the reward system (Izuma and Adolphs, 2013). These exploratory findings were consistent with the results of previous studies that reported that social influence effect affects participants' behaviors through the neural mechanisms involved in reward learning and behavioral adjustment (Izuma et al., 2008; Mason et al., 2009; Wei et al., 2013).

Several limitations of this study should be noted. Firstly, the present task is different from the Asch's experiment. In our study, subjects had no other information about trust decision except the group members' choices. This manipulation can potentially lead to conforming to the group member. Secondly, we did not use scale to quantitatively measure whether participants believed the experiment manipulation, which might also affect the result. Thirdly, the number of non-conformity trials that were included in exploratory analysis was less than 10 which limited the power of our GLM model. Despite that the results for these analyses survived correction, further studies could consider increasing the number of trials in order to more reliably evaluate these effects.

CONCLUSION

The present study provides evidence of the relationship between social influence and trust decisions. It complements previous research by assessing the neural basis of social influence and extends our understanding of the decision to trust. Our behavioral results revealed that individuals are likely to be influenced by others' opinions and conform to the opinions of peers in a trust game. Participants conformed to the opinion of the majority with a relatively high level of decision confidence as a result of the high reward expectancy in the trust social influence condition. Decreased activations were observed in STG when external information was presented and this result might suggest that external trust information affected neural activity in brain regions associated with trust game, which relieved the perspective-taking process in the trust game. The results of exploratory analysis indicated that the brain regions involved in value processing and reward learning, such as the vmPFC, ventral striatum, ACC, and parahippocampal gyrus, were activated when subjects decided to follow the majority in trusting behavior. The PPI analysis confirmed possible increased functional connectivity between the ventral striatum and the prefrontal cortex during conformity in trusting behavior. In conclusion, these findings suggest that the mechanisms underlying social influence in trusting behavior may be similar to those implicated in reward learning.

AUTHOR CONTRIBUTIONS

ZW and YZ conceived and designed the experiments. ZW and ZZ programed the task and analyzed the data. ZW performed the experiments. ZW, ZZ and YZ wrote the paper.

FUNDING

This project was funded by China Postdoctoral Science Foundation #2018M633337.

ACKNOWLEDGMENTS

The authors thank the reviewers for their constructive comments that improved the manuscript considerably.

REFERENCES

Ahern, K. R., Duchin, R., and Shumway, T. (2014). Peer effects in risk aversion and trust. *Rev. Financ. Stud.* 27, 3213–3240. doi: 10.1093/rfs/hhu042

Asch, S. E. (1951). *Effects of Group Pressure Upon the Modification and Distortion of Judgments*. Pittsburgh: Carnegie Press.

Asch, S. E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority. *Psychol. Monogr.*70, 1–70. doi: 10.1037/h0093718

Berns, G. S., Capra, C. M., Moore, S., and Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage* 49, 2687–2696. doi: 10.1016/j.neuroimage.2009.10.070

Berns, G. S., Chappelow, J., Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., and Richards, J. (2005). Neurobiological correlates of social conformity and independence during mental rotation. *Biol. Psychiatry* 58, 245–253. doi: 10.1016/j.biopsych.2005.04.012

Berns, G. S., McClure, S. M., Pagnoni, G., and Montague, P. R. (2001). Predictability modulates human brain response to reward. *J. Neurosci.* 21, 2793–2798. doi: 10.1523/JNEUROSCI.21-08-02793.2001

Burke, C. J., Tobler, P. N., Schultz, W., and Baddeley, M. (2010). Striatal BOLD response reflects the impact of herd information on financial decisions. *Front. Hum. Neurosci.* 4:48. doi: 10.3389/fnhum.2010.00048

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: evidence from a field experiment on financial decisions. *Econometrica* 82, 1273–1301. doi: 10.3982/ECTA11991

Camara, E., Rodriguez-Fornells, A., and Münte, T. F. (2008). Functional connectivity of reward processing in the brain. *Front. Hum. Neurosci.* 2:19. doi: 10.3389/neuro.09.019.2008

Camara, E., Rodriguez-Fornells, A., Ye, Z., and Münte, T. F. (2009). Reward networks in the brain as captured by connectivity measures. *Front. Neurosci.* 3:350. doi: 10.3389/neuro.01.034.2009

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., and Frith, C. D. (2010). How the opinion of others affects our valuation of objects? *Curr. Biol.* 20, 1165–1170. doi: 10.1016/j.cub.2010.04.055

Campbell-Meiklejohn, D. K., Simonsen, A., Jensen, M., Wohlert, V., Gjerløff, T., Scheel-Kruger, J., et al. (2012). Modulation of social influence by methylphenidate. *Neuropsychopharmacology* 37, 1517–1525. doi: 10.1038/npp.2011.337

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., and Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3721–3726. doi: 10.1073/pnas.0710069105

Cialdini, R. B., and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annu. Rev. Psychol.* 55, 591–621. doi: 10.1146/annurev.psych.55.090902.142015

Cochard, F., Nguyen Van, P., and Willinger, M. (2004). Trusting behavior in a repeated investment game. *J. Econ. Behav. Organ.* 55, 31–44. doi: 10.1016/j.jebo.2003.07.004

Fehr, E., and Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn. Sci.* 11, 419–427. doi: 10.1016/j.tics.2007.09.002

Frank, M. J., and Claus, E. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol. Sci.* 113, 300–326. doi: 10.1037/0033-295X.113.2.300

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., and Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* 6, 218–229. doi: 10.1006/nimg.1997.0291

Frith, U., and Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 358, 459–473. doi: 10.1098/rstb.2002.1218

Gardner, M., and Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study. *Dev. Psychol.* 41, 625–635. doi: 10.1037/0012-1649.41.4.625

Grabenhorst, F., and Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends Cogn. Sci.* 15, 56–67. doi: 10.1016/j.tics.2010.12.004

Haun, D., van Leeuwen, E. J., and Edelson, M. G. (2013). Majority influence in children and other animals. *Dev. Cogn. Neurosci.* 3, 61–71. doi: 10.1016/j.dcn.2012.09.003

Hayden, B. Y., Pearson, J. M., and Platt, M. L. (2009). Fictive reward signals in the anterior cingulate cortex. *Science* 324, 948–950. doi: 10.1126/science.1168488

Holroyd, C. B., and Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* 109:679. doi: 10.1037/0033-295X.109.4.679

Hong, H., Kubik, J. D., and Stein, J. C. (2004). Social interaction and stock-market participation. *J. Finance* 59, 137–163. doi: 10.1111/j.1540-6261.2004.00629.x

Izuma, K., and Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron* 78, 563–573. doi: 10.1016/j.neuron.2013.03.023

Izuma, K., Saito, D. N., and Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron* 58, 284–294. doi: 10.1016/j.neuron.2008.03.020

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* 61, 140–151. doi: 10.1016/j.neuron.2008.11.027

Klucharev, V., Munneke, M. A., Smidts, A., and Fernandez, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *J. Neurosci.* 31, 11934–11940. doi: 10.1523/JNEUROSCI.1869-11.2011

Knutson, B., and Wimmer, G. E. (2007). Splitting the difference. *Ann. N. Y. Acad. Sci.* 1104, 54–69. doi: 10.1196/annals.1390.020

Mason, M. F., Dyer, R., and Norton, M. I. (2009). Neural mechanisms of social influence. *Organ. Behav. Hum. Decis. Process.* 110, 152–159. doi: 10.1016/j.obhdp.2009.04.001

McCoy, A. N., and Platt, M. L. (2005). Expectations and outcomes: decision-making in the primate brain. *J. Comp. Physiol. A Neuroethol. Sens. Neural. Behav. Physiol.* 191, 201–211. doi: 10.1007/s00359-004-0565-9

Morgan, T. J. H., and Laland, K. N. (2012). The biological bases of conformity. *Front. Neurosci.* 6:87. doi: 10.3389/fnins.2012.00087

Rangel, A., and Hare, T. (2010). Neural computations associated with goal-directed choice. *Curr. Opin. Neurobiol.* 20, 262–270. doi: 10.1016/j.conb.2010.03.001

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science* 306, 443–447. doi: 10.1126/science.1100301

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., and Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron* 35, 395–405. doi: 10.1016/S0896-6273(02)00755-9

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* 15, 2539–2543. doi: 10.1097/00001756-200411150-00022

Rushworth, M. F., Mars, R. B., and Summerfield, C. (2009). General mechanisms for making decisions? *Curr. Opin. Neurobiol.* 19, 75–83. doi: 10.1016/j.conb.2009.02.005

Rushworth, M. F. S., Noonan, M. A. P., Boorman, E. D., Walton, M. E., and Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron* 70, 1054–1069. doi: 10.1016/j.neuron.2011.05.014

Simonsen, A., Scheel-Krüger, J., Jensen, M., Roepstorff, A., Møller, A., Frith, C. D., et al. (2014). Serotoninergic effects on judgments and social learning of trustworthiness. *Psychopharmacology* 231, 2759–2769. doi: 10.1007/s00213-014-3444-2

Sun, D. L., Shao, R. B., Wang, Z. X., and Lee, T. M. C. (2018). Perceived gaze direction modulates neural processing of prosocial decision making. *Front. Hum. Neurosci.* 12:52. doi: 10.3389/fnhum.2018.00052

Turner, J. C. (1991). *Social Influence*. Milton Keynes: Open University Press.

van den Bos, W., Cohen, M. X., Kahnt, T., and Crone, E. A. (2012). Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cereb. Cortex* 22, 1247–1255. doi: 10.1093/cercor/bhr198

Wei, Z., Zhao, Z., and Zheng, Y. (2013). Neural mechanisms underlying social conformity in an ultimatum game. *Front. Hum. Neurosci.* 7:896. doi: 10.3389/fnhum.2013.00896

Wei, Z., Zhao, Z., and Zheng, Y. (2017). The neural basis of social influence in a dictator decision. *Front. Psychol.* 8:2134. doi: 10.3389/fpsyg.2017.02134

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

frontiers
in Human Neuroscience

ORIGINAL RESEARCH
published: 13 February 2019
doi: 10.3389/fnhum.2019.00047

Social Mindfulness and Psychosis: Neural Response to Socially Mindful Behavior in First-Episode Psychosis and Patients at Clinical High-Risk

Imke L. J. Lemmers-Jansen[1,2]*, Anne-Kathrin J. Fett[3], Niels J. Van Doesum[4,5], Paul A. M. Van Lange[4], Dick J. Veltman[6] and Lydia Krabbendam[2]

[1] Section of Educational Neuroscience, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [2] Section Clinical, Neuro- and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [3] Department of Psychology, City, University of London, London, United Kingdom, [4] Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [5] Social and Organisational Psychology, Leiden University, Leiden, Netherlands, [6] Neuroscience Campus Amsterdam, Vrije Universiteit Amsterdam, VU Medical Center Amsterdam, Amsterdam, Netherlands

Background: Psychosis is characterized by problems in social functioning and trust, the assumed glue to positive social relations. But what helps building trust? A prime candidate could be social mindfulness: the ability and willingness to see and consider another person's needs and wishes during social decision making. We investigated whether first-episode psychosis patients (FEP) and patients at clinical high-risk (CHR) show reduced social mindfulness, and examined the underlying neural mechanisms.

Methods: Twenty FEP, 17 CHR and 46 healthy controls, aged 16–31, performed the social mindfulness task (SoMi) during fMRI scanning, spontaneously and after the instruction "to keep the other's best interest in mind." As first of two people, participants had to choose one out of four products, of which three were identical and one was unique, differing in a single aspect (e.g., color).

Results: FEP tended to choose the unique item (unmindful choice) more often than controls. After instruction, all groups significantly increased the number of mindful choices compared to the spontaneous condition. FEP showed reduced activation of the caudate and medial prefrontal cortex (mPFC) during mindful, and of the anterior cingulate cortex (ACC), mPFC, and left dorsolateral prefrontal cortex (dlPFC) during unmindful decisions. CHR showed reduced activation of the ACC compared to controls.

Discussion: FEP showed a trend toward more unmindful choices. A similar increase of mindful choices after instruction indicated the ability for social mindfulness when prompted. Results suggested reduced sensitivity to the rewarding aspects of social mindfulness in FEP, and reduced consideration for the other player. FEP (and CHR to a lesser extent) might perceive unmindful choices as less incongruent with the automatic mindful responses than controls. Reduced socially mindful behavior in FEP may hinder the building of trust and cooperative interactions.

Keywords: social mindfulness, trust, first-episode psychosis, clinical high-risk, fMRI, mentalizing, reward

OPEN ACCESS

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

Reviewed by:
Poppy L. A. Schoenberg,
Vanderbilt University Medical Center,
United States
Anne Caroline Wheeler,
RTI International, United States

****Correspondence:***
Imke L. J. Lemmers-Jansen
imke.jansen@vu.nl

Received: *16 October 2018*
Accepted: *29 January 2019*
Published: *13 February 2019*

Citation:
Lemmers-Jansen ILJ, Fett A-KJ, Van Doesum NJ, Van Lange PAM, Veltman DJ and Krabbendam L (2019) Social Mindfulness and Psychosis: Neural Response to Socially Mindful Behavior in First-Episode Psychosis and Patients at Clinical High-Risk. Front. Hum. Neurosci. 13:47. doi: 10.3389/fnhum.2019.00047

INTRODUCTION

Psychotic disorder is characterized by positive psychotic symptoms (e.g., delusions and hallucinations), negative symptoms (e.g., affective flattening and lack of motivation), and cognitive impairments (American Psychiatric Association, 2013). In addition, patients display problems in social functioning (Couture et al., 2006; Fett et al., 2012), which are already present before the onset of psychosis, and have also been reported in individuals at high-risk for psychosis (Yung et al., 2003; Ballon et al., 2007; Cornblatt et al., 2007; Corcoran et al., 2011; Velthorst et al., 2016a,b). One of these social impairments is reduced trust in unknown others, a common aspect of the psychosis spectrum, which is also found in individuals at genetic and clinical high-risk for psychosis. In chronic patients reduced trust seems to persist in the face of trustworthy behavior of others, possibly due to repeated negative experiences. In contrast to first episode patients and individuals at genetic and clinical high-risk, initially reduced trust can be overcome when others are trustworthy (Gromann et al., 2013; Fett et al., 2014a, 2015, 2016; Lemmers-Jansen et al., 2018a). Additionally, patients may sometimes misplace trust: patients with a first-episode psychosis did not decrease their levels of trust when confronted with an unfair partner to the same degree as healthy controls did (Fett et al., 2016). Although trust is often assumed to be the glue to positive social interactions, little is known about what it is that helps to build trust. A prime candidate could be social mindfulness. Social mindfulness is expressed as low-cost cooperative behavior, that involves the ability and willingness to see and consider another person's needs and wishes during social decision making (Van Lange and Van Doesum, 2015). In this paper social mindfulness is explored in first-episode psychosis patients and in patients at clinical high-risk for psychosis. We investigate whether first-episode and clinical high-risk patients show reduced spontaneous socially mindful behavior, and whether they show reduced neural activation in brain areas associated with social decision making compared to controls, similar to the trust literature in these patient groups (Gromann et al., 2013; Lemmers-Jansen et al., 2018a).

Social mindfulness (SoMi) is being thoughtful of others in the present moment, and considering their needs and wishes when making a decision (Van Doesum et al., 2013; Lemmers-Jansen et al., 2018b). Perceived socially mindful behavior will promote close relationships, facilitate cooperation, and increase trust in the other person (Declerck et al., 2013; Van Doesum et al., 2013; Van Lange and Van Doesum, 2015; Dou et al., 2018). On the contrary, displays of low socially mindful behavior may elicit reduced feelings of trust in the counterpart, who in turn will behave less trusting toward the initial actor. The ability and willingness to think about preferences of and benefits for others are two core requirements for SoMi, for trust, and for positive social interactions in general. The ability, the *skill*, reflects social cognitive processes, especially mentalizing, to recognize the needs and wishes of others, to judge the other's trustworthiness and intentions; the willingness, the *will*, reflects social motivation, the sensitivity to the intrinsic pleasurable effects of positive social interactions, to act socially mindful or to trust (Declerck et al., 2013; Lemmers-Jansen et al., 2018b). Apart from social cognition and reward, other mechanisms may also play a role, like self-representation and self-other distinction (Fonagy and Target, 2006; van Os et al., 2010). In the SoMi task participants are presented with four items, of which three are identical and one only differed in a single aspect (e.g., three green baseball caps and one yellow baseball cap). Choosing the unique item removes the option of choice for the second player. This is the socially unmindful choice. Choosing one of the three identical items still leaves the next player a choice, making it the socially mindful choice.

Previously Lemmers-Jansen et al. (2018b) have shown that making mindful decisions engaged the fronto-parietal network and when choosing unmindfully the default mode network was recruited. Mindful and unmindful choices showed an overlap of activated regions, especially in medial prefrontal cortex (mPFC) and the temporo-parietal junction (TPJ). Exclusion analysis revealed condition specific activation for mindful choices in parietal regions. Unmindful choices activated frontal regions (anterior cingulate cortex (ACC) and mPFC). The caudate was associated with mindful choices in prosocially oriented subjects, indicating a rewarding aspect of prosocial behavior. These regions are consistent with the reward, cognitive control, and social cognition systems, each of which is implicated in prosocial decision making (Declerck et al., 2013).

Patients with psychotic disorder show aberrant activation of these brain areas, which are often associated with mentalizing and reward processing (Juckel et al., 2006; Murray et al., 2008; Schilbach et al., 2016; Bartholomeusz et al., 2018). Both mechanisms have been linked to trust (Brüne, 2005; King-Casas et al., 2005; Baas et al., 2008; Marwick and Hall, 2008; Benedetti et al., 2009; Gromann et al., 2013; Billeke et al., 2015; Horat et al., 2017; Lemmers-Jansen et al., 2017). In patients at clinical high-risk for psychosis (CHR) and in unaffected siblings of patients similar social cognitive impairments are found, albeit to a lesser degree, suggesting milder impairments in high-risk populations, and a major decline with the first episode (Pinkham et al., 2007; Bora and Pantelis, 2013; Lavoie et al., 2013; McCleery et al., 2014). CHR are already in care for other psychopathology, reporting psychotic-like symptoms, but have not yet experienced (or never will) full-blown psychosis (Velthorst et al., 2009; Woods et al., 2009; van Os and Linscott, 2012; Wigman et al., 2012; van Os and Reininghaus, 2016). With the conversion to psychosis, impairments in social function increase, therefore it is important to understand the changes that occur during this transition. Investigating social interactions in patients with psychotic symptoms, first-episode psychosis patients (FEP) and CHR, who are unbiased with regard to long lasting stigma and institutionalized living can help identifying processes that decline at first onset. This may provide specific targets for intervention, to prevent or delay social decline, which is crucial for outcome prognosis and early intervention.

Isolated social cognitive skills have been successfully assessed with off-line tasks; however, they do not capture the wide range of mechanisms involved in social interactions. Real life social interactions are difficult to measure in a controlled environment, but neuro-economics provide paradigms, investigating sharing or trusting behavior in real interactions. They can capture social cognitive skills, as well as the neural processes underlying social behavior. When investigating impairments in social behavior in psychopathology, especially schizophrenia/psychosis, studying these paradigms with fMRI can advance the understanding of the neurobiology of social dysfunction (Kishida et al., 2010; Hasler, 2012; Cáceda et al., 2014; Riccardi et al., 2015). Studies have shown aberrant behavioral outcomes and neural mechanisms during trust processing in patients with psychosis (Fett et al., 2012, 2014a, 2015, 2016; Gromann et al., 2013; Lemmers-Jansen et al., 2018a). The SoMi paradigm resembles everyday interpersonal situations by involving very little costs (c.f. giving compliments or making nice gestures), and low-level cooperation, as reflected in a straightforward choice for an item, whereas trust can be seen as high-level cooperation, with more at stake, including risk, and building a model about the counterpart. Furthermore, unlike other neuro-economical paradigms, where the pay-offs for the player and the other person are usually very clear, in the SoMi task participants have to recognize or see what others want, and how their actions influence the outcomes for others. Thus, the situation has to be recognized as a social one, with all the associated demands and opportunities. This realization is an intricate part of the construct.

The current study sets out to investigate behavioral and neural mechanisms of spontaneous socially mindful decisions in FEP and CHR patients. Given that patients show impairments in reward processing and social cognitive skills, including taking the perspective of the other person, we hypothesized that (1) FEP will opt more often for individual gain (the unique item), and therefore spontaneously make more unmindful choices compared to controls. Given the straightforward nature of the task, we further hypothesized that (2) FEP, similar to controls, make more socially mindful choices after being asked to keep the other's best interest in mind. Given the evidence for altered brain activation during social decisions and impairments in reward processing and mentalizing in patients, we hypothesized that (3) FEP will show reduced activation of the caudate during spontaneous mindful choices, and generally less activation in mPFC and TPJ compared to controls. With regard to CHR, we hypothesized that they will show (4) an intermediate behavioral performance compared to FEP and controls (Giuliano et al., 2012; Thompson et al., 2012; Lemmers-Jansen et al., 2018a), and intermediate neural activation compared to FEP and controls. Additionally, associations of positive and negative symptoms, and paranoia with behavioral and neural outcomes are explored, based on the association between paranoia and reduced trust, and mixed outcomes in the trust game literature (Gromann et al., 2013; Fett et al., 2014b; Lemmers-Jansen et al., 2018a).

MATERIALS AND METHODS

Subjects

Twenty-nine young adolescents with a first psychotic episode (FEP), aged 16–22 were recruited in the Amsterdam area. Additionally, 18 patients at clinical high-risk for developing psychosis (CHR) and 52 controls, aged 16–31 were recruited in the Amsterdam and The Hague area. All patients were contacted through their treating clinicians at the academic medical center Amsterdam (AMC), the Amsterdam early intervention team psychosis ("Vroege Interventie Psychose" or VIP team), and PsyQ The Hague. FEP were diagnosed at the AMC, according to the DSM-IV criteria (American Psychiatric Association, 2000), and included within 18 months of the diagnosis (M = 5.6 months). Thirty percent was unmedicated, 55% was on atypical antipsychotic medication, and 15% on other psychotropic medication. FEP illness ranged from hospitalized to reentering work and society living, with symptoms ranging from mildly to markedly ill, and one severely ill patient (Leucht et al., 2005). CHR were help seeking individuals that were referred to PsyQ by their general practitioners or other mental health institutions. After an initial diagnosis based on their complaints, all new admissions (between age 14–35) were screened for an "at-risk mental state" (ARMS) with the Comprehensive Assessment of At-Risk Mental States [CAARMS; (Yung et al., 2005)], a semi-structured interview that assesses psychotic experiences in the last year before assessment. Additionally, patients had to display marked problems in socially useful activities (work and study), relationships, and self-care, indicated by a score below 55 on the Social and Occupational Functioning Assessment Scale [SOFAS; mean score 46.9; (Goldman et al., 1992; Morosini et al., 2000)], see also (Rietdijk et al., 2012). CHR were included within 1 year after CAARMS assessment (M = 4.8 months). Symptoms of depression and anxiety are often the primary presenting complaints of CHR patients, rather than (subclinical) psychotic symptoms (Modinos et al., 2014). Similar to other CHR samples (Woods et al., 2009; Kelleher et al., 2012; Morrison et al., 2012; Wigman et al., 2012; Fusar-Poli et al., 2014), the current CHR sample had comorbid diagnoses of anxiety (5), personality (3), eating (2) and mood (2) disorders, trauma (2), and ADHD (3). Exclusion criteria for both patient groups were primary diagnosis of mood disorders, comorbidity with autism spectrum disorder (ASD) and an IQ < 80, information provided by their primary clinicians, based on the initial assessment and diagnosis. And for the healthy control group this was a family history of psychiatric disorders, ASD and an IQ < 80, as was assessed with a questionnaire and by recruiting participants from regular educational institutes. All participants were fluent in Dutch. We excluded nine FEP, one CHR, and six controls from analyses due to invalid or missing data. The remaining sample consisted of 20 FEP, 17 CHR, and 46 controls. The first study on the neural mechanisms of social mindfulness was based on the same sample of healthy controls (Lemmers-Jansen et al., 2018b). This research was approved by the Ethics Committee of the VU Medical Center Amsterdam.

Measures

Social Mindfulness Paradigm (SoMi Task)

The SoMi task consisted of a dyadic game in which the participant and a fictitious other (someone "who you don't know and are not likely to meet in the near future") repeatedly choose what to take from a set of four similar products [identical task characteristics as in Lemmers-Jansen et al. (2018b)]. One of these products was unique in a single aspect, whereas the other three were identical, for example one red among three green apples (1:3 ratio; see **Figure 1**). Participants were instructed that they would always choose first, and that chosen items would not be replaced. Choosing an identical item would leave the next person a choice, and was scored as socially mindful; taking away the unique item would limit this other person's choice, and was scored as socially unmindful. Each of the experimental trials featured different products. All products were low in value, e.g., pens, water bottles, etc. We added control trials as a baseline measure to the analyses, which displayed the items in a 2:2 ratio in which the participant's choices would have no social consequences (see **Figure 1**).

The SoMi paradigm was administered twice. In the first round (spontaneous condition), participants only received the general information that someone else would choose after them. In the second round (instructed condition), participants received the additional instruction to "keep the best interest of the other person in mind" (cf. Van Doesum et al., 2013, Studies 1a–c). This round was added to check if lower scores on social mindfulness are the result of a lack of ability to understand how one's own behavior affects the other player. Each round consisted of 24 experimental trials, with one unique versus three identical items (e.g., one red and three green apples); 24 control trials, offering two pairs of identical items (e.g., two blue and two yellow baseball hats), and 12 low-level baseline trials, where participants passively watched a blank screen. Each trial had a duration of 5000 ms. A final score of social mindfulness was computed. This SoMi-index is the proportion of socially mindful answers, varying from 0 (only socially unmindful choices) to 1 (only socially mindful choices).

Positive and Negative Syndromes Scale (PANSS)

The well validated 30-item PANSS semi-structured interview was used for rating symptoms in the 2 weeks prior to testing. The PANSS distinguishes between positive, negative, and general symptoms (Kay et al., 1987). The item P6 was used as an indication for paranoia. Items are scaled on a 7-point Likert scale, ratings 3 and higher indicating clinical values. All FEP and 13 CHR completed the interview.

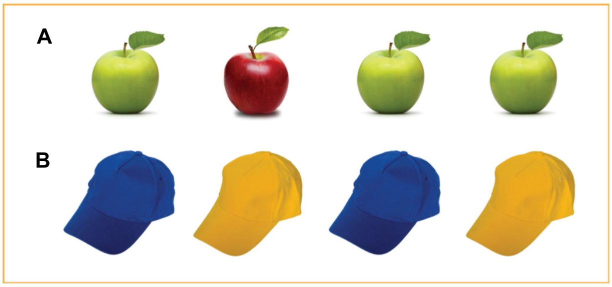


FIGURE 1 | Example trials of the Social Mindfulness task (SoMi), displaying **(A)** an experimental trial (3:1 ratio presentation) where the participant's choice can influence the choice options of the other player; and **(B)** a control trial (2:2 ratio presentation) where the choice has no social consequences. The stimulus was displayed for 5000 ms, followed by an inter-stimulus interval (0, 1000, or 2000 ms). Reproduced from (Lemmers-Jansen et al., 2018b).

Wechsler Adult Intelligence Scale (WAIS) Vocabulary

A subtest of the WAIS-III (Wechsler, 1997) was included as a proxy for intelligence. The vocabulary subscale, a measure of verbal comprehension, consisted of 33 words that had to be defined or described by the participants (e.g., winter, catastrophe, and reckless). Answers were either fully correct (2 points), partially correct (1) or wrong (0). After six consecutive 0 scores, the test was discontinued.

Procedure

All participants provided general informed consent; patients also signed a form that allowed the researchers to obtain additional patient data from their care giving institution. After signing the consent forms, participants completed several pen and paper questionnaires, followed by two computer-administered tasks. Both patient groups were assessed with the PANSS. Medication use was assessed with the pre-scanning questionnaire, a questionnaire pertaining to the safety procedure for scanning. Subsequently participants were scanned for about an hour. For patients, extra time was needed to guide them into the scanner, comfort them and to ensure they understood the tasks. Therefore, we planned 15 min extra for them. First, all participants performed an unrelated task [the Trust Game, see (Lemmers-Jansen et al., 2018a)]. Next the structural scan was made, during which participants could relax, while watching a movie if they wanted. The SoMi task was the second task participants performed in the scanner. Two rounds of the SoMi task were played as described above, each round lasting 6 min. Instructions for the task were given in the scanner, immediately prior to the task. Four practice trials were completed before the task started to ensure that instructions were clear. Instructions for the second round were given visually and orally, while scanning was paused. Scanning sessions ended with a resting state scan. After scanning participants received an image of their structural brain scan, 25€ for participation and travel costs were reimbursed.

fMRI Data Acquisition

fMRI data were obtained at the Spinoza center Amsterdam, using a 3.0 T Philips Achieva whole body scanner (Philips Healthcare, Best, Netherlands) equipped with a 32 channel head coil. A T2* EPI sequence (TR = 2, TE = 27.63, FA = 76.1°, FOV 240 mm, voxel size 3 × 3 × 3, 37 slices, 0.3 mm gap) was used, resulting in 185 images per condition. A T1-weighed anatomical scan was acquired for anatomical reference (TR = 8.2, TE = 3.8, FA = 8°, FOV 240 mm*188 mm, voxel size 1 × 1 × 1, 220 slices).

Data Analysis

Behavioral Data

Demographic and behavioral data were analyzed using Stata 13 (StataCorp, 2013) with regression analyses and chi-square tests. For behavioral outcomes, *t*-tests and regression analyses were used. Analyses included spontaneous choices and choices after instruction, and were controlled for age and gender as *a priori* confounders, and for WAIS Vocabulary, to avoid potential confounding effects of group differences. To examine whether the results were influenced by general cognitive impairment in patients, all analyses were repeated without WAIS Vocabulary.

Imaging Data

Imaging data were analyzed using Statistical Parametric Mapping (SPM8; Wellcome Trust Centre for Neuroimaging, London, United Kingdom). Functional images for each participant were preprocessed with the following steps: realign and unwarp, coregistration with individual structural images, segmented for normalization to an MNI template and smoothing with a 6 mm full width at half maximum (FWHM) Gaussian kernel. At fist-level, a general linear model (GLM) was used to construct individual time courses for the onset of the presentation of the trial, and individual reaction times for the spontaneous and instructed conditions. Decision making was defined as the interval between stimulus onset and button press. In the SoMi trials (3:1 ratio) a distinction was made between the socially mindful (choosing one of the three identical items) and unmindful responses (choosing the unique item). The choices made in the spontaneous and instructed rounds were contrasted with the corresponding control trials (2:2 ratio).

At second level, a three-group factorial design was used for the main effects and group comparisons. Participants were only included in the analysis of the SoMi trials if they had at least 1/3 of the 24 responses within a response category: Participants with 1–7 unmindful responses were included only in the mindful condition, with 8–16 unmindful choices were included in both mindful and unmindful conditions, and with 17–24 only in the unmindful condition. Due to this procedure, sample size varied per condition. Mindful and unmindful responses in the spontaneous condition and mindful responses after instruction were included in the neural analyses. The unmindful condition after instruction included too few participants for reliable analyses. Analyses were controlled for age, gender, and WAIS.

Whole brain main effects of social choice (all SoMi trials, including mindful and unmindful choices; FWE corrected) over groups were calculated, to define the coordinates for the regions of interest (ROI). Regions involved in social decision making, conflict processing, and self- and other-representation, were predefined on the basis of previous neuroimaging studies (Zhu et al., 2007; Rilling and Sanfey, 2011). When activated in the whole brain analysis, peak coordinates of the predefined regions were extracted and a 10 mm sphere was built around this peak. For the bilateral caudate a 5 mm sphere was used. Whole brain results did not show activation clusters for the ACC and right insula. Coordinates for the ACC were therefore manually defined from a larger prefrontal cluster, covering the ACC; right insula coordinates were mirrored from the contralateral region. This resulted in the following ROIs: mPFC (MNI coordinates: 0, 50, 34), precuneus (9, −52, 31), ACC (3, 47, 13), and bilateral insula (33, 20, −14 and −27, 20, −14), caudate (12, 8, 13 and −12, 5, 13), TPJ (51, −52, 46 and −51, −55, 43), and dlPFC (42, 14, 49 and −39, 20, 46). A priori ROI analyses compared group activation per condition. *P*-values were Bonferroni corrected for multiple comparisons and adjusted for internal correlations, by using the Simple Interactive Statistical Analysis Bonferroni tool[1], resulting in adjusted significance thresholds (Woudstra et al., 2013; Li et al., 2014; Lemmers-Jansen et al., 2018a). Additional whole-brain group comparisons were performed, to investigate activation outside the predefined ROIs.

RESULTS

Participant Characteristics

Participant characteristics are shown in **Table 1**. FEP, CHR and controls did not differ significantly from each other with respect to gender, handedness, and other measures (see **Table 1**). However, CHR were significantly older than FEP (β = 0.56, $p < 0.001$), and controls (β = 0.40, $p < 0.001$). Furthermore, FEP scored significantly lower than CHR (β = −0.30, p = 0.02), and controls (β = −0.34, p = 0.003) on the WAIS Vocabulary scale. Between the patient groups, no significant differences were found in number of medicated participants, nor in symptom severity.

Behavioral Results

Spontaneous Choices

Partly confirming our first hypothesis, FEP showed a trend toward spontaneously choosing the unique item more often than controls, (β = −0.22, f^2 = 0.15, p = 0.08; see **Table 2**), but not than CHR (β = 0.08, p = 0.59). CHR did not differ significantly from controls (β = −0.12, p = 0.33). The difference between spontaneous mindful and unmindful choices was significant in all groups (controls: t = −4.0, $p < 0.001$; CHR: t = −2.0, p = 0.05; FEP: t = 2.1, p = 0.04). Note that spontaneously FEP made more socially unmindful than socially mindful choices, whereas CHR and controls made more socially mindful choices, resulting in a SoMi index under 0.5 for FEP (i.e., 0.45), and above 0.5 for CHR and controls (0.54 and 0.56, respectively).

Choices After Instruction

After instruction FEP showed the same trend to choose the unique item more often than controls (β = −0.22, p = 0.08), but not than CHR (β = 0.12, p = 0.45), and CHR did not differ significantly from controls (β = −0.08, p = 0.51). After instruction all groups made significantly more socially mindful than socially unmindful choices (all t's < −4, all p's < 0.001). Additionally, all groups significantly increased the number of mindful choices compared to the spontaneous condition (all t's < −3.5, all p's $\leq$ 0.001), indicating that the manipulation was effective. The difference at trend level between FEP and controls in the number of socially mindful choices persisted after instruction, showing no significant group differences in the number of socially mindful choices after instruction similarly (β = −0.05, p = 0.67). The CHR group performed in between FEP and controls, resembling the control group most.

Additional analyses without WAIS as a covariate showed the same results, with similar significance levels, and comparable medium to large effect sizes. However, the trend result of FEP choosing more often the unique option than controls now

[1] http://www.quantitativeskills.com/sisa/calculations/bonfer.htm

TABLE 1 | Participant characteristics.

	FEP *N* = 20	CHR *N* = 17	Controls *N* = 46	Statistics
Gender (*n* male, %)	13 (65%)	7 (41%)	24 (52%)	$\chi^2 = 2.12$
Age (Mean/*SD*)	19.96 (1.56)	**23.78 (2.49)**	21.10 (2.72)	*F* = 11.85*
WAIS (Mean/*SD*)	**32.8 (11.02)**	41.71 (12.16)	42.11 (11.26)	*F* = 4.96*
Right handed *n* (%)	16 (80%)	17 (100%)	38 (83%)	$\chi^2 = 4.09$
Medicated *n* (%)	14 (70%)	8 (47%)		$\chi^2 = 0.16$
• Atypical antipsychotics (*n*)	11	–		
• Other psychotropics (*n*)	3	8		
PANSS – total (*SD*)	60.70 (15.32)	58.92 (11.84)		*F* = 0.13
• Mean severity (*SD*)	2.02 (.51)	1.96 (0.39)		
Positive – total (*SD*)	13.60 (6.0)	13.38 (2.69)		*F* = 0.02
• Mean (*SD*)	1.94 (0.86)	1.91 (0.38)		
Negative – total (*SD*)	16.80 (6.13)	13.69 (3.88)		*F* = 2.64
• Mean (*SD*)	2.40 (0.88)	1.96 (0.55)		
General – total (*SD*)	30.30 (7.73)	31.85 (6.31)		*F* = 0.36
• Mean	1.89 (0.48)	1.99 (0.39)		
P6 paranoia item (*SD*)	1.9 (1.6)	1.2 (0.4)		*F* = 2.64

**Significant group differences at $p < 0.05$, with the group in bold differing from the two other groups.*
FEP, first-episode psychosis; CHR, clinical high-risk. WAIS, Wechsler Adult Intelligence Scale; PANSS, Positive and Negative Syndrome Scale.

TABLE 2 | Number of choices and participants for fMRI analysis per condition by group.

Condition	FEP (*N* = 20)	CHR (*N* = 17)	Controls (*N* = 46)
Spontaneous			
Mindful, mean (*SD*)	10.80 (3.41)*	13.06 (3.17)	13.43 (3.65)
Unmindful, mean (*SD*)	13.05 (3.40)*	10.82 (3.23)	10.39 (3.64)
Instructed			
Mindful, mean (*SD*)	17.05 (7.19)*	20.12 (4.05)	20.76 (4.41)
Unmindful, mean (*SD*)	6.95 (7.19)*	3.76 (4.01)	3.24 (4.41)
SoMi-index	0.45 (0.14)*	0.54 (0.13)	0.56 (0.15)
Number of participants for fMRI analysis			
Social decision	20	17	46
Spontaneous mindful	18	16	43
Spontaneous unmindful	20	15	37
Mindful after instruction	18	17	45
Unmindful after instruction	8	3	8

**p = 0.08, FEP differing at trend level from controls. FEP, first-episode psychosis; CHR, clinical high-risk; SoMi index, proportion of socially mindful choices.*

reached significance, in both spontaneous choices ($\beta = -0.27$, $p = 0.02$) and choices after instruction ($\beta = -0.24$, $p = 0.02$).

Symptoms

Associations between the paranoia item, positive and negative symptoms and behavioral outcomes were investigated in FEP and CHR. Group-by-symptom interactions on spontaneous and instructed choices were non-significant (all $|\beta\text{'s}| < 1.4$, $p\text{'s} > 0.21$), as were the group-by-symptom interactions on increase of mindful choices after instruction ($\beta\text{'s} < 0.67$, $p\text{'s} > 0.59$). Removing the interactions from the model showed an inverse main effect at trend level of negative symptoms on increase of mindful choices after instruction ($\beta = -0.33$, $p = 0.08$), indicating that patients with higher levels of negative symptoms showed a smaller increase of mindful choices after instruction than patients with less negative symptoms.

fMRI Results

ROI Analyses

Analogous to our previous study (Lemmers-Jansen et al., 2018b), participants were only included in a condition when they had at least 1/3 of the decisions within that particular condition (see section "Imaging Data"). Due to this procedure, sample size varied per condition, see **Table 2**.

To determine the coordinates for the predefined ROI, whole brain analysis of social choice over all trials and all groups were conducted (see **Table 3**). Regions and coordinates used for ROI analyses are marked in bold font. ROI analyses were performed with 11 predefined ROIs. ROI analysis outcomes are presented in **Table 4**. During spontaneous mindful choices, the caudate was less activated in FEP than controls; and the mPFC was less activated in FEP than both CHR and controls. During spontaneous unmindful choices controls activated the

TABLE 3 | Whole brain main effects of social choices, including all SoMi trials, regardless of choice, over all groups.

Region	Hemisphere	MNI coordinates			Cluster size *k*	*z*
		X	***Y***	***Z***		
mPFC	L	−6	38	46	807	7.32
mPFC	**R**	**0**	**50**	**34**		7.28
mPFC	R	12	44	46		6.84
mPFC	R	6	68	7	6	5.42
mPFC	R	9	62	28	2	5.14
Inferior frontal gyrus	L	−51	17	7	21	5.65
dlPFC	**R**	**42**	**14**	**49**	89	6.94
dlPFC	**L**	**−39**	**20**	**46**	66	6.02
Inferior orbitofrontal gyrus	R	36	23	−11	144	6.88
Inferior orbitofrontal gyrus	R	48	35	−11		6.57
Middle orbitofrontal gyrus	R	39	56	−2		5.37
Middle orbitofrontal gyrus	L	−42	50	−2	1	4.74
Insula	**L**	**−27**	**20**	**−14**	30	6.16
Inferior orbitofrontal gyrus	L	−33	20	−23		5.40
Inferior orbitofrontal gyrus	L	−48	38	−8	3	5.39
Inferior frontal operculum	R	57	20	13	58	7.20
Superior frontal gyrus	L	−21	59	22	2	4.96
Middle temporal gyrus	R	63	−43	−5	29	5.94
Middle temporal gyrus	L	−54	−22	−11	42	5.45
Middle temporal gyrus	L	−63	−28	−5		5.40
Middle temporal gyrus	L	−48	−31	−5		5.32
Middle temporal gyrus	R	63	−13	−14	1	4.74
Superior temporal pole	L	−45	20	−14	14	5.78
Inferior temporal gyrus	L	−48	−1	−32	3	4.94
Angular gyrus	R	57	−61	34	416	>7.7
TPJ	**R**	**51**	**−52**	**46**		6.94
TPJ	**L**	**−51**	**−55**	**43**	342	7.68
Angular gyrus	L	−54	−64	25		6.91
Angular gyrus	L	−42	−67	46		6.69
Caudate	**R**	**12**	**8**	**13**	13	5.55
Caudate	**L**	**−12**	**5**	**13**	4	5.00
Mid cingulum	L	−3	−22	34	312	6.58
Precuneus	**R**	**9**	**−52**	**31**		6.07
Precuneus	R	3	−67	34		5.64

Regions displayed in bold font correspond with predefined regions of interest (ROI). MNI, Montreal Neurological Institute; mPFC, medial prefrontal cortex; dlPFC, dorsolateral prefrontal cortex; TPJ, temporo-parietal junction; L, left; R, right. Regions and coordinates for the ROI are displayed in bold font. Two additional ROI were defined: Anterior cingulate cortex (ACC): 3, 47, 13, based on the large prefrontal cluster, and right insula: 33, 20, −14, based on mirroring the left insula. Analyses were FWE corrected, at $p < 0.05$.

ACC significantly more than both CHR and FEP, and controls showed more activation in the mPFC and the left dlPFC than FEP. Summarizing, most activation was found in controls, with CHR performing in between FEP and controls. Mindful choices after instruction yielded no significant group differences. Replication of the analyses without WAIS Vocabulary as covariate yielded similar significance levels in the same ROIs as displayed in **Table 4**.

Exploratory Whole Brain Analyses

Additional whole brain analyses on group differences per SoMi condition revealed no group differences surviving the FWE cluster correction. To verify that all three groups showed similar brain activation, a global-null analysis was performed. Results are shown in **Supplementary Table S1**, and indicate similar networks as described in our previous paper with a partly overlapping sample (Lemmers-Jansen et al., 2018b). During spontaneous unmindful choices, however, this analysis also revealed additional activation in the ventrolateral prefrontal cortex, caudate, and insula.

Associations With Symptoms

Analyses showed no significant associations between contrast estimates and symptoms. Contrast estimates of the significant ROI were associated with positive and negative symptoms. In the mPFC during mindful choices (the only ROI with

TABLE 4 | Region of interest analysis outcome per condition in the SoMi paradigm.

	ROI	CHR > FEP		Con > FEP		Con > CHR	
		t	*p*	*t*	*p*	*t*	*p*
Spontaneous mindful*	mPFC	1.74	0.043ˆ	2.30	0.012		
	Right caudate			1.84	0.035		
Spontaneous unmindful**	ACC			1.93	0.029	1.85	0.34
	Left dlPFC			2.01	0.024		
	mPFC			1.86	0.034		

**Significance level of p = 0.039, Bonferroni corrected, adjusted for internal correlation. **Significance level of p = 0.042, Bonferroni corrected, adjusted for internal correlation. ˆBordering significance. ROI, region of interest; CHR, clinical high-risk; FEP, first-episode psychosis; Con, healthy controls; mPFC, medial prefrontal cortex; ACC, anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex.*

significant differences between the two patient groups), no significant group-by-symptom interactions were found (positive: $\beta = -0.69$, $p = 0.6$; negative: $\beta = 1.05$, $p = 0.34$; paranoia: $\beta = -0.13$, $p = 0.95$). After removing the interaction from the model, symptoms did not show a significant main effect on mPFC activation. In the ROIs where patient groups differed significantly from to controls, i.e., the right caudate during mindful choices, and the ACC, mPFC and left dlPFC during unmindful choices, the only significant association was in the dlPFC with paranoia, indicating increased activation with increasing paranoia ($\beta = 0.49$, $p = 0.029$).

DISCUSSION

The purpose of the present research was to examine the behavioral outcomes and neural substrates of socially mindful and unmindful choices, in a clinical high-risk (CHR) and first-episode psychosis (FEP) sample. The results showed a trend toward more spontaneously unmindful choices in FEP compared to the CHR and control group, but a similar increase of socially mindful choices after instruction across the three groups, indicating the ability for socially mindful behavior when prompted. At the neural level FEP showed decreased activation in the caudate compared to controls when making socially mindful choices, possibly suggesting reduced sensitivity to the rewarding aspects of social mindfulness. Additionally, reduced activation in the mPFC, ACC and dlPFC was found in FEP during unmindful choices, suggesting that FEP might perceive unmindful choices as less incongruent with the automatic mindful responses than controls. Scores for CHR were in between FEP and controls.

Behavioral Results

In partial support of our hypothesis, we found a marginal effect showing that FEP tended to make spontaneously more socially unmindful choices than controls. This result became significant when analyses were run without the covariate WAIS Vocabulary, a proxy for intelligence. Despite the visual nature of the task, social mindfulness seems to depend on cognitive ability. The small reduction of effect size, however, suggests only a minimal confounding effect. FEP opted more often for the unique than for the non-unique option, with a mean proportion of social mindfulness of 0.45, while the other groups chose more often the non-unique item (mean proportion CHR: 0.54; controls: 0.56). Other studies have shown that the mean proportion of social mindfulness toward strangers converges around 0.67 (Van Doesum et al., 2013; Van Lange and Van Doesum, 2015). Social mindfulness tends to be greater in prosocially orientated individuals; when the other player has a trustworthy face, is an in-group member, or is someone liked (Van Doesum et al., 2013, 2016) and when the second person is perceived as lower in social class than the participant (Van Doesum et al., 2017). When interacting with a friend, social mindfulness also increases (Van Lange and Van Doesum, 2015; Van Doesum et al., 2016). However, with a foe or an outgroup member, the proportion of socially mindful choices decreases to around 0.45, which could be labeled as social hostility (Van Doesum et al., 2016). FEP showed a similarly low proportion of social mindfulness, suggesting that they were spontaneously less inclined to consider the interest of the partner. This finding is of theoretical interest because it indicates that psychotic disorder is also linked to differences in spontaneous low-cost cooperation. As noted earlier, social mindfulness is causally linked to maintaining or enhancing trust: Greater social mindfulness yields greater trust in the recipient of socially mindful behavior. And especially, more social unmindfulness undermines trust (see Van Doesum et al., 2013; Dou et al., 2018), in that the negative consequences (ending up having no choice) tend to outweigh positive consequences in terms of attention, and of what people recall and reciprocate (Van Lange et al., 2002). Whether SoMi is sensitive to interventions remains to be determined in future research. We suggest that the SoMi task has some features, such as the emphasis on perspective taking and giving small favors to others, that might make it suitable for intervention purposes. However, there is a big differences between instructing social mindfulness and actually expressing it in a spontaneous manner in real life situations.

Contrary to our hypothesis, the mean SoMi score of CHR was not between FEP and controls, but CHR displayed a similar level of spontaneous socially mindful behavior as controls. Low level cooperation therefore seems to be still intact in CHR, contrary to the higher level trust processing, where CHR showed reduced levels of baseline trust, similar to FEP [cf. (Lemmers-Jansen et al., 2018a)]. Confirming our hypothesis, though, all groups showed a similar increase of socially mindful choices when instructed to keep the other's best interest in mind, indicating that low levels of social mindfulness in FEP did not reflect an inability to understand the impact of their behavior on the partner, but rather a reduced tendency to consider other's perspective spontaneously. These findings suggest an impact of the first psychotic episode on spontaneous socially mindful behavior. This tentatively suggests that reduced socially mindful behavior in FEP may affect social interactions with other people, which may fail to evolve according to the positive reciprocity that characterizes 'typical' patterns of interactions, if not made explicitly clear. However, similar to observations of initially reduced trust in FEP, our findings show that this pattern can be overcome through positive feedback (Lemmers-Jansen et al., 2018a).

Neural Results

The analyses of the brain activation corroborated the behavioral findings that FEP were able to act socially mindfully when prompted: No group differences in brain activation were found in the mindful condition after instruction.

As hypothesized, FEP showed reduced activation of the caudate compared to controls. Reduced caudate activity during socially mindful choices might reflect reduced feelings of reward when leaving the other the option, setting aside one's own preferences. Impairments in reward processing in psychosis have frequently been reported (Juckel et al., 2006; Waltz et al., 2010; Strauss et al., 2013). Neuro-economic research using the trust game in chronic patients similarly showed reduced caudate activity during positive social interactions (Gromann et al., 2013; Brown et al., 2014). The current findings suggest that reduced reward processing may extend to socially mindful behavior. When social interactions or doing good are not perceived as inherently rewarding (Higgins and Scholer, 2009), FEP will less likely engage in other regarding interactions. Furthermore, in line with our hypothesis, activation of the mPFC, one of the regions previously shown to be engaged in both mindful and unmindful choices (Lemmers-Jansen et al., 2018b) was reduced in FEP compared to controls (and CHR) in both choice types. The mPFC is involved in many aspects of social and general cognition, such as mentalizing, learning, memory, cognitive control, decision making, predicting valence and timing of expected outcomes of an action, reward anticipation and salience, and in processing emotions (Ridderinkhof et al., 2004; Frith and Frith, 2006; Van Overwalle and Baetens, 2009; Ziauddeen and Murray, 2010; Forster and Brown, 2011; Euston et al., 2012; Cáceda et al., 2014). Considering this range of functions in the context of the current paradigm, reduced mPFC activation might indicate that FEP consider the consequences of their decisions for the other player less than controls and CHR. It is important to consider that reduced mPFC activation in both decision types might also reflect general and not task related reduced activity of this region, inherent to psychosis patients (Sugranyes et al., 2011). Contrary to our predictions FEP did not display reduced TPJ activation in socially mindful, nor in socially unmindful choices. As hypothesized, reward and mentalizing mechanisms may play a role in social mindfulness. This is supported by the activation of mPFC and caudate during mindful decisions. No differences were found between groups in the ROIs that are typically related to self-perception and self-other representation (insula, precuneus, and TPJ), suggesting that these mechanisms are unlikely to play a role. However, the association between these mechanisms and social mindfulness warrants further investigation with additional measures.

When making socially unmindful decisions, FEP showed reduced activation of mPFC, ACC, and dlPFC, the latter being associated with the paranoia score. Reduced mPFC activation in both spontaneous choice options could indicate reduced anticipation of thoughts and feelings of others (Frith and Frith, 2006), although other process might also play a role in socially unmindful decisions. Alternatively, after instruction to mind the other's best interest, no differences in neural activation were present, suggesting that FEP only show impairments in spontaneously anticipating the feelings of others, but follow instructions similar to controls. The ACC and dlPFC are, among many cognitive processes, involved in cognitive control and conflict processing (MacDonald et al., 2000; Milham et al., 2003; Badre and Wagner, 2004; Ridderinkhof et al., 2004; Mitchell et al., 2009). Based on predominantly prefrontal activation during socially unmindful decisions, when contrasted with socially mindful decisions, we previously concluded that in healthy subjects socially unmindful decisions seemed to be more deliberate, requiring cognitive control, whereas socially mindful decisions were the more automatic response (Lemmers-Jansen et al., 2018b). Reduced ACC and dlPFC activation in FEP might therefore indicate that FEP perceive socially unmindful choices as less incongruent or deliberate, and less effortful. The association of dlPFC activation and paranoia warrants further investigation.

In contrast to FEP, CHR showed no impairments in reward processing areas, possibly explaining the intact spontaneous socially mindful behavior. No differences in mentalizing areas were found, suggesting normal functioning of this mechanism. CHR showed less reduction in activation than FEP, especially in prefrontal areas [see also (Morey et al., 2005; Broome et al., 2010; Schmidt et al., 2013)]. These results only partly confirm our hypothesis of intermediate neural activation compared to FEP and controls. Reduced ACC activation compared to controls during unmindful choices might indicate, similar to FEP, that CHR also perceive unmindful choices as less incongruent or effortful than controls. Differential neural activation in patients at-risk despite similar behavioral performance was previously found, although activation in CHR was often increased (Morey et al., 2005; Marjoram et al., 2006; Seiferth et al., 2008; Brüne et al., 2011; Derntl et al., 2015).

The frequency of spontaneous socially mindful behavior appeared to be independent of symptom severity, but reduced after a first psychotic episode. Future research could investigate this behavior in chronic illness, testing whether spontaneous socially mindful behavior further declines with illness duration. Interestingly, more negative symptoms were associated with less increase of mindful choices in both patient groups after instruction. Negative symptoms have been related to avolition, reduced social motivation, and poor social functioning and cognition in both FEP and CHR (Milev et al., 2005; Voges and Addington, 2005; Chan and Chen, 2011; Corcoran et al., 2011; Meyer et al., 2014). However, they are not related to reduced spontaneous socially mindful behavior, but to reduced changes in socially mindful behavior after being told to mind the other's best interest, possibly indicating reduced propensity to set aside their own preferences for the benefit of others.

Limitations and Future Directions

Several limitations should be considered. First, the size of the sample was modest, especially of the CHR group. Results should therefore be considered as a first step investigating socially mindful decision making in these patients, demanding replication and extension in future research. Larger samples would permit subtyping of FEP and comparing CHR that transitioned to psychosis with non-converters, yielding more information about social mindfulness and its underlying mechanisms in patient populations. Furthermore, only one CHR patient transitioned to psychosis, 1 year after participating in this study. This could raise questions about the representativeness of the sample. However, our sample was comparable to other samples in terms of comorbidities (Woods et al., 2009; Corcoran et al., 2011; Morrison et al., 2012; Fusar-Poli et al., 2014; Modinos et al., 2014; Ising et al., 2016), and participants were assessed with the CAARMS, and included when scoring below 55 on the SOFAS, following the procedure of previous CHR investigations (Shim et al., 2008; Phillips et al., 2009; Fusar-Poli et al., 2010; Wood et al., 2011; Rietdijk et al., 2012; Thompson et al., 2012; van der Gaag et al., 2012; McGorry and van Os, 2013; Valmaggia et al., 2013).

The CHR patients were not informed about their at-risk for psychosis status, to not unnecessarily alarm them, since most of them will not make the transition to psychosis. They were told they had 'extraordinary or unusual experiences', when discussing psychotic symptoms. These were regularly monitored by their treating clinicians. Regardless of transition rates, the presence of psychotic symptoms in these patients is associated with a poorer prognosis, showing that these patients are in need of special care (Ruhrmann et al., 2010; van Os and Linscott, 2012; McGorry and van Os, 2013; Valmaggia et al., 2013; van Os and Reininghaus, 2016). Further, FEP symptom severity was rather mild, possibly due to responsiveness to antipsychotic treatment. Similar symptom severity has been found in stable and medicated patients (Möller et al., 2005), but a wider range of symptoms might have revealed more associations with social mindfulness at the behavioral or neural level. Additionally, participants were scanned for about an hour, which could have caused fatigue, which may have affected neural outcomes, especially in patients with a psychotic disorder. Questions remain about the motivation for choosing socially (un)mindfully. For further research we recommend additional measures, such as a questionnaire after the task, to inquire after the motivation of participants' choices; measures of hostility toward other people; and tasks that could rule out the alternative explanation that FEP might encounter choosing the single option as the prepotent, automatic response [see also (Yamagishi et al., 2016)]. Despite controlling for WAIS vocabulary, questions about the association between social mindfulness and verbal and cognitive ability remain. This warrants further investigation.

CONCLUSION

This study is the first to examine social mindfulness in patients with problems in social cognition and functioning. Our results show that relative to the healthy control group, spontaneous social mindfulness seems reduced when patients have experienced a first full-blown psychosis. At the same time, social mindfulness was not lower for those at risk for psychosis (CHR). However, when explicitly told to act in the other person's best interest, FEP are just as capable to be socially mindful as anyone else. Neural outcomes suggest reduced feelings of reward during socially mindful decisions in FEP, and possibly a stronger, automatic inclination to focus on the unique options that seem most attractive for themselves in FEP and CHR. Left to themselves, FEP seem to have reduced appreciation for the more subtle social consequences of leaving or limiting choices. In all, the current research can be seen as a first step in showing reduced socially mindful behavior in psychosis. This aspect of social interactions may possibly underlie deficits in more complex cooperative interactions, such as trust, that patients might otherwise develop within their social environment. Alternatively, displays of low socially mindful behavior may elicit reduced feelings of trust in the counterpart, who in turn will behave less trusting. The next step is to investigate whether and how social unmindfulness serves as a cause underlying patients' low levels of trust.

AUTHOR CONTRIBUTIONS

IL-J collected and processed the data, and wrote the manuscript. LK and PVL designed the study. PVL and NVD designed the paradigm. NVD prepared it for fMRI. DV planned the fMRI analyses. IL-J, A-KF, LK, and DV interpreted the fMRI results. LK supervised the project. All authors discussed the results, contributed to the writing process, and approved the final manuscript.

FUNDING

This work was supported by funding of the Hersenstichting Nederland [KS2011(1)-75], a VIDI and VICI grant from the Netherlands Organization for Scientific Research (NWO) (452-07-007 and 453-11-005), and a ERC Consolidator grant (648082 SCANS) to LK. A-KF was supported by a Netherlands

Organization for Scientific Research (NWO) VENI grant (451-13-035) and a NARSAD Young Investigator Grant from the Brain & Behaviour research Foundation (24138).

ACKNOWLEDGMENTS

The authors would like to thank Dirk Heslenfeld for his help in making the SoMi paradigm suitable for fMRI, Esther Hanssen for her help during data collection and Julia Isacescu for preliminary analyses of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2019.00047/full#supplementary-material

REFERENCES

American Psychiatric Association (2000). *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders, Text Revision*. Washington, DC: American Psychiatric Association.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Washington, DC: American Psychiatric Pub. doi: 10.1176/appi.books.9780890425596

Baas, D., Aleman, A., Vink, M., Ramsey, N. F., De Haan, E. H., and Kahn, R. S. (2008). Evidence of altered cortical and amygdala activation during social decision-making in schizophrenia. *Neuroimage* 40, 719–727. doi: 10.1016/j.neuroimage.2007.12.039

Badre, D., and Wagner, A. D. (2004). Selection, integration, and conflict monitoring: assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron* 41, 473–487. doi: 10.1016/S0896-6273(03)00851-1

Ballon, J. S., Kaur, T., Marks, I. I., and Cadenhead, K. S. (2007). Social functioning in young people at risk for schizophrenia. *Psychiatry Res.* 151, 29–35. doi: 10.1016/j.psychres.2006.10.012

Bartholomeusz, C. F., Ganella, E. P., Whittle, S., Allott, K., Thompson, A., Abu-Akel, A., et al. (2018). An fMRI study of theory of mind in individuals with first episode psychosis. *Psychiatry Res.* 281, 1–11. doi: 10.1016/j.pscychresns.2018.08.011

Benedetti, F., Bernasconi, A., Bosia, M., Cavallaro, R., Dallaspezia, S., Falini, A., et al. (2009). Functional and structural brain correlates of theory of mind and empathy deficits in schizophrenia. *Schizophr. Res.* 114, 154–160. doi: 10.1016/j.schres.2009.06.021

Billeke, P., Armijo, A., Castillo, D., López, T., Zamorano, F., Cosmelli, D., et al. (2015). Paradoxical expectation: oscillatory brain activity reveals social interaction impairment in schizophrenia. *Biol. Psychiatry* 78, 421–431. doi: 10.1016/j.biopsych.2015.02.012

Bora, E., and Pantelis, C. (2013). Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. *Schizophr. Res.* 144, 31–36. doi: 10.1016/j.schres.2012.12.013

Broome, M. R., Fusar-Poli, P., Matthiasson, P., Woolley, J. B., Valmaggia, L., Johns, L. C., et al. (2010). Neural correlates of visuospatial working memory in the 'at-risk mental state'. *Psychol. Med.* 40, 1987–1999. doi: 10.1017/S0033291710000280

Brown, E. C., Tas, C., Gonzalez, C., and Brüne, M. (2014). Neurobiologic underpinnings of social cognition and metacognition in schizophrenia spectrum disorders. *Soc. Cogn. Metacogn. Schizophr.* 1, 1–27. doi: 10.1016/B978-0-12-405172-0.00001-6

Brüne, M. (2005). "Theory of mind" in schizophrenia: a review of the literature. *Schizophr. Bull.* 31, 21–42. doi: 10.1093/schbul/sbi002

Brüne, M., Özgürdal, S., Ansorge, N., Von Reventlow, H. G., Peters, S., Nicolas, V., et al. (2011). An fMRI study of "theory of mind" in at-risk states of psychosis: comparison with manifest schizophrenia and healthy controls. *Neuroimage* 55, 329–337. doi: 10.1016/j.neuroimage.2010.12.018

Cáceda, R., Nemeroff, C. B., and Harvey, P. D. (2014). Toward an understanding of decision making in severe mental illness. *J. Neuropsychiatry Clin. Neurosci.* 26, 196–213. doi: 10.1176/appi.neuropsych.12110268

Chan, K. K., and Chen, E. Y. (2011). Theory of mind and paranoia in schizophrenia: a game theoretical investigation framework. *Cogn. Neuropsychiatry* 16, 505–529. doi: 10.1080/13546805.2011.561576

Corcoran, C., Kimhy, D., Parrilla-Escobar, M., Cressman, V., Stanford, A., Thompson, J., et al. (2011). The relationship of social function to depressive and negative symptoms in individuals at clinical high risk for psychosis. *Psychol. Med.* 41, 251–261. doi: 10.1017/S0033291710000802

Cornblatt, B. A., Auther, A. M., Niendam, T., Smith, C. W., Zinberg, J., Bearden, C. E., et al. (2007). Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. *Schizophr. Bull.* 33, 688–702. doi: 10.1093/schbul/sbm029

Couture, S. M., Penn, D. L., and Roberts, D. L. (2006). The functional significance of social cognition in schizophrenia: a review. *Schizophr. Bull.* 32, S44–S63. doi: 10.1093/schbul/sbl029

Declerck, C. H., Boone, C., and Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain Cogn.* 81, 95–117. doi: 10.1016/j.bandc.2012.09.009

Derntl, B., Michel, T. M., Prempeh, P., Backes, V., Finkelmeyer, A., Schneider, F., et al. (2015). Empathy in individuals clinically at risk for psychosis: brain and behaviour. *Br. J. Psychiatry* 207, 407–413. doi: 10.1192/bjp.bp.114.159004

Dou, K., Wang, Y. J., Li, J. B., Li, J. J., and Nie, Y. G. (2018). Perceiving high social mindfulness during interpersonal interaction promotes cooperative behaviours. *Asian J. Soc. Psychol.* 21, 97–106. doi: 10.1111/ajsp.12210

Euston, D. R., Gruber, A. J., and Mcnaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron* 76, 1057–1070. doi: 10.1016/j.neuron.2012.12.002

Fett, A. J., Shergill, S., Gromann, P., and Krabbendam, L. (2015). Trust vs. Paranoia: the dynamics of social interaction in early and chronic psychosis. *Schizophr. Bull.* 41:S170.

Fett, A.-K., Gromann, P., Giampietro, V., Shergill, S., and Krabbendam, L. (2014a). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Soc. Cogn. Affect. Neurosci.* 9, 395–402. doi: 10.1093/scan/nss144

Fett, A.-K., Shergill, S., Gromann, P., Dumontheil, I., Blakemore, S.-J., Yakub, F., et al. (2014b). Trust and social reciprocity in adolescence–a matter of perspective-taking. *J. Adolesc.* 37, 175–184. doi: 10.1016/j.adolescence.2013.11.011

Fett, A.-K., Shergill, S., Joyce, D., Riedl, A., Strobel, M., Gromann, P., et al. (2012). To trust or not to trust: the dynamics of social interaction in psychosis. *Brain* 135, 976–984. doi: 10.1093/brain/awr359

Fett, A.-K., Shergill, S., Korver-Nieberg, N., Yakub, F., Gromann, P., and Krabbendam, L. (2016). Learning to trust: trust and attachment in early psychosis. *Psychol. Med.* 46, 1437–1447. doi: 10.1017/S0033291716000015

Fonagy, P., and Target, M. (2006). The mentalization-focused approach to self pathology. *J. Pers. Disord.* 20, 544–576. doi: 10.1521/pedi.2006.20.6.544

Forster, S. E., and Brown, J. W. (2011). Medial prefrontal cortex predicts and evaluates the timing of action outcomes. *Neuroimage* 55, 253–265. doi: 10.1016/j.neuroimage.2010.11.035

Frith, C. D., and Frith, U. (2006). The neural basis of mentalizing. *Neuron* 50, 531–534. doi: 10.1016/j.neuron.2006.05.001

Fusar-Poli, P., Broome, M. R., Matthiasson, P., Woolley, J. B., Johns, L., Tabraham, P., et al. (2010). Spatial working memory in individuals at high risk for psychosis: longitudinal fMRI study. *Schizophr. Res.* 123, 45–52. doi: 10.1016/j.schres.2010.06.008

Fusar-Poli, P., Nelson, B., Valmaggia, L., Yung, A. R., and Mcguire, P. K. (2014). Comorbid depressive and anxiety disorders in 509 individuals with an at-risk mental state: impact on psychopathology and transition to psychosis. *Schizophr. Bull.* 40, 120–131. doi: 10.1093/schbul/sbs136

Giuliano, A. J., Li, H., Mesholam-Gately, R. I., Sorenson, S. M., Woodberry, K. A., and Seidman, L. J. (2012). Neurocognition in the psychosis risk syndrome: a quantitative and qualitative review. *Curr. Pharm. Design* 18, 399–415. doi: 10.2174/138161212799316019

Goldman, H. H., Skodol, A. E., and Lave, T. R. (1992). Revising axis V for DSM-IV: a review of measures of social functioning. *Am. J. Psychiatry* 149, 1148–1156. doi: 10.1176/ajp.149.9.1148

Gromann, P., Heslenfeld, D., Fett, A.-K., Joyce, D., Shergill, S., and Krabbendam, L. (2013). Trust versus paranoia: abnormal response to social reward in psychotic illness. *Brain* 136(Pt 6), 1968–1975. doi: 10.1093/brain/awt076

Hasler, G. (2012). Can the neuroeconomics revolution revolutionize psychiatry? *Neurosci. Biobehav. Rev.* 36, 64–78. doi: 10.1016/j.neubiorev.2011.04.011

Higgins, E. T., and Scholer, A. A. (2009). Engaging the consumer: the science and art of the value creation process. *J. Consum. Psychol.* 19, 100–114. doi: 10.1016/j.jcps.2009.02.002

Horat, S. K., Favre, G., Prévot, A., Ventura, J., Herrmann, F. R., Gothuey, I., et al. (2017). Impaired social cognition in schizophrenia during the Ultimatum Game: an EEG study. *Schizophr. Res.* 192, 308–316. doi: 10.1016/j.schres.2017.05.037

Ising, H. K., Kraan, T. C., Rietdijk, J., Dragt, S., Klaassen, R. M., Boonstra, N., et al. (2016). Four-year follow-up of cognitive behavioral therapy in persons at ultra-high risk for developing psychosis: the Dutch early detection intervention evaluation (EDIE-NL) trial. *Schizophr. Bull.* 42, 1243–1252. doi: 10.1093/schbul/sbw018

Juckel, G., Schlagenhauf, F., Koslowski, M., Filonov, D., Wüstenberg, T., Villringer, A., et al. (2006). Dysfunction of ventral striatal reward prediction in schizophrenic patients treated with typical, not atypical, neuroleptics. *Psychopharmacology* 187, 222–228. doi: 10.1007/s00213-006-0405-4

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* 13, 261–276. doi: 10.1093/schbul/13.2.261

Kelleher, I., Keeley, H., Corcoran, P., Lynch, F., Fitzpatrick, C., Devlin, N., et al. (2012). Clinicopathological significance of psychotic experiences in non-psychotic young people: evidence from four population-based studies. *Br. J. Psychiatry* 201, 26–32. doi: 10.1192/bjp.bp.111.101543

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., King-Casas, B., and Montague, P. R. (2010). Neuroeconomic approaches to mental disorders. *Neuron* 67, 543–554. doi: 10.1016/j.neuron.2010.07.021

Lavoie, M.-A., Lacroix, J. B., Godmaire-Duhaime, F., Jackson, P. L., and Achim, A. M. (2013). Social cognition in first-degree relatives of people with schizophrenia: a meta-analysis. *Psychiatry Res.* 209, 129–135. doi: 10.1016/j.psychres.2012.11.037

Lemmers-Jansen, I. L., Fett, A.-K. J., Hanssen, E., Veltman, D. J., and Krabbendam, L. (2018a). Learning to trust: social feedback normalizes trust behavior in first episode psychosis and clinical high risk. *Psychol. Med.* doi: 10.1017/S003329171800140X. [Epub ahead of print].

Lemmers-Jansen, I. L., Krabbendam, L., Amodio, D. M., Van Doesum, N. J., Veltman, D. J., and Van Lange, P. A. (2018b). Giving others the option of choice: an fMRI study on low-cost cooperation. *Neuropsychologia* 109, 1–9. doi: 10.1016/j.neuropsychologia.2017.12.009

Lemmers-Jansen, I. L., Krabbendam, L., Veltman, D. J., and Fett, A.-K. J. (2017). Boys vs. girls: gender differences in the neural development of trust and reciprocity depend on social context. *Dev. Cogn. Neurosci.* 25, 235–245. doi: 10.1016/j.dcn.2017.02.001

Leucht, S., Kane, J. M., Kissling, W., Hamann, J., Etschel, E., and Engel, R. R. (2005). What does the PANSS mean? *Schizophr. Res.* 79, 231–238. doi: 10.1016/j.schres.2005.04.008

Li, W., Tol, M. J., Li, M., Miao, W., Jiao, Y., Heinze, H. J., et al. (2014). Regional specificity of sex effects on subcortical volumes across the lifespan in healthy aging. *Hum. Brain Mapp.* 35, 238–247. doi: 10.1002/hbm.22168

MacDonald, A. W., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288, 1835–1838. doi: 10.1126/science.288.5472.1835

Marjoram, D., Job, D. E., Whalley, H. C., Gountouna, V. E., Mcintosh, A. M., Simonotto, E., et al. (2006). A visual joke fMRI investigation into Theory of Mind and enhanced risk of schizophrenia. *Neuroimage* 31, 1850–1858. doi: 10.1016/j.neuroimage.2006.02.011

Marwick, K., and Hall, J. (2008). Social cognition in schizophrenia: a review of face processing. *Br. Med. Bull.* 88, 43–58. doi: 10.1093/bmb/ldn035

McCleery, A., Horan, W. P., and Green, M. F. (2014). "Social cognition during the early phase of schizophrenia," in *Social Cognition and Metacognition in Schizophrenia: Psychopathology and Treatment Approaches* eds P. H. Lysaker, G. Dimaggio, and M. Brune (Cambridge, MA: Academic Press), 49–67. doi: 10.1016/B978-0-12-405172-0.00003-X

McGorry, P., and van Os, J. (2013). Redeeming diagnosis in psychiatry: timing versus specificity. *Lancet* 381, 343–345. doi: 10.1016/S0140-6736(12)61268-9

Meyer, E. C., Carrión, R. E., Cornblatt, B. A., Addington, J., Cadenhead, K. S., Cannon, T. D., et al. (2014). The relationship of neurocognition and negative symptoms to social and role functioning over time in individuals at clinical high risk in the first phase of the North American Prodrome Longitudinal Study. *Schizophr. Bull.* 40, 1452–1461. doi: 10.1093/schbul/sbt235

Milev, P., Ho, B.-C., Arndt, S., and Andreasen, N. C. (2005). Predictive values of neurocognition and negative symptoms on functional outcome in schizophrenia: a longitudinal first-episode study with 7-year follow-up. *Am. J. Psychiatry* 162, 495–506. doi: 10.1176/appi.ajp.162.3.495

Milham, M., Banich, M., Claus, E., and Cohen, N. (2003). Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. *Neuroimage* 18, 483–493. doi: 10.1016/S1053-8119(02)00050-2

Mitchell, D. G., Luo, Q., Avny, S. B., Kasprzycki, T., Gupta, K., Chen, G., et al. (2009). Adapting to dynamic stimulus-response values: differential contributions of inferior frontal, dorsomedial, and dorsolateral regions of prefrontal cortex to decision making. *J. Neurosci.* 29, 10827–10834. doi: 10.1523/JNEUROSCI.0963-09.2009

Modinos, G., Allen, P., Frascarelli, M., Tognin, S., Valmaggia, L., Xenaki, L., et al. (2014). Are we really mapping psychosis risk? Neuroanatomical signature of affective disorders in subjects at ultra high risk. *Psychol. Med.* 44, 3491–3501. doi: 10.1017/S0033291714000865

Möller, H.-J., Llorca, P.-M., Sacchetti, E., Martin, S. D., Medori, R., Parellada, E., et al. (2005). Efficacy and safety of direct transition to risperidone long-acting injectable in patients treated with various antipsychotic therapies. *Int. Clin. Psychopharmacol.* 20, 121–130. doi: 10.1097/00004850-200505000-00001

Morey, R. A., Inan, S., Mitchell, T. V., Perkins, D. O., Lieberman, J. A., and Belger, A. (2005). Imaging frontostriatal function in ultra-high-risk, early, and chronic schizophrenia during executive processing. *Arch. Gen. Psychiatry* 62, 254–262. doi: 10.1001/archpsyc.62.3.254

Morosini, P., Magliano, L., Brambilla, L., Ugolini, S., and Pioli, R. (2000). Development, reliability and acceptability of a new version of the DSM-IV Social and Occupational Functioning Assessment Scale (SOFAS) to assess routine social funtioning. *Acta Psychiatr. Scand.* 101, 323–329.

Morrison, A. P., French, P., Stewart, S. L., Birchwood, M., Fowler, D., Gumley, A. I., et al. (2012). Early detection and intervention evaluation for people at risk of psychosis: multisite randomised controlled trial. *BMJ* 344:e2233. doi: 10.1136/bmj.e2233

Murray, G., Corlett, P., Clark, L., Pessiglione, M., Blackwell, A., Honey, G., et al. (2008). Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Mol. Psychiatry* 13, 239–276. doi: 10.1038/sj.mp.4002157

Phillips, L. J., Nelson, B., Yuen, H. P., Francey, S. M., Simmons, M., Stanford, C., et al. (2009). Randomized controlled trial of interventions for young people at ultra-high risk of psychosis: study design and baseline characteristics. *Austr. N. Z. J. Psychiatry* 43, 818–829. doi: 10.1080/00048670903107625

Pinkham, A. E., Penn, D. L., Perkins, D. O., Graham, K. A., and Siegel, M. (2007). Emotion perception and social skill over the course of psychosis: a comparison of individuals "at-risk" for psychosis and individuals with early and chronic schizophrenia spectrum illness. *Cogn. Neuropsychiatry* 12, 198–212. doi: 10.1080/13546800600985557

Riccardi, I., Stratta, P., and Rossi, A. (2015). When economic theory meets the mind: neuroeconomics as a new approach to psychopathology. *J. Psychopathol.* 21, 141–144.

Ridderinkhof, K. R., Van Den Wildenberg, W. P., Segalowitz, S. J., and Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain Cogn.* 56, 129–140. doi: 10.1016/j.bandc.2004.09.016

Rietdijk, J., Klaassen, R., Ising, H., Dragt, S., Nieman, D., Van De Kamp, J., et al. (2012). Detection of people at risk of developing a first psychosis: comparison of two recruitment strategies. *Acta Psychiatr. Scand.* 126, 21–30. doi: 10.1111/j.1600-0447.2012.01839.x

Rilling, J. K., and Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annu. Rev. Psychol.* 62, 23–48. doi: 10.1146/annurev.psych.121208.131647

Ruhrmann, S., Schultze-Lutter, F., and Klosterkötter, J. (2010). Probably at-risk, but certainly ill—advocating the introduction of a psychosis spectrum disorder in DSM-V. *Schizophr. Res.* 120, 23–37. doi: 10.1016/j.schres.2010.03.015

Schilbach, L., Derntl, B., Aleman, A., Caspers, S., Clos, M., Diederen, K. M., et al. (2016). Differential patterns of dysconnectivity in mirror neuron and mentalizing networks in schizophrenia. *Schizophr. Bull.* 42, 1135–1148. doi: 10.1093/schbul/sbw015

Schmidt, A., Smieskova, R., Aston, J., Simon, A., Allen, P., Fusar-Poli, P., et al. (2013). Brain connectivity abnormalities predating the onset of psychosis: correlation with the effect of medication. *JAMA Psychiatry* 70, 903–912. doi: 10.1001/jamapsychiatry.2013.117

Seiferth, N. Y., Pauly, K., Habel, U., Kellermann, T., Shah, N. J., Ruhrmann, S., et al. (2008). Increased neural response related to neutral faces in individuals at risk for psychosis. *Neuroimage* 40, 289–297. doi: 10.1016/j.neuroimage.2007.11.020

Shim, G., Kang, D. H., Chung, Y. S., Yoo, S. Y., Shin, N. Y., and Kwon, J. S. (2008). Social functioning deficits in young people at risk for schizophrenia. *Austr. N. Z. J. Psychiatry* 42, 678–685. doi: 10.1080/00048670802203459

StataCorp (2013). *Stata Statistical Software*. College Station, TX: StataCorp LP.

Strauss, G. P., Waltz, J. A., and Gold, J. M. (2013). A review of reward processing and motivational impairment in schizophrenia. *Schizophr. Bull.* 40, S107–S116. doi: 10.1093/schbul/sbt197

Sugranyes, G., Kyriakopoulos, M., Corrigall, R., Taylor, E., and Frangou, S. (2011). Autism spectrum disorders and schizophrenia: meta-analysis of the neural correlates of social cognition. *PLoS One* 6:e25322. doi: 10.1371/journal.pone.0025322

Thompson, A. D., Papas, A., Bartholomeusz, C., Allott, K., Amminger, G. P., Nelson, B., et al. (2012). Social cognition in clinical "at risk" for psychosis and first episode psychosis populations. *Schizophr. Res.* 141, 204–209. doi: 10.1016/j.schres.2012.08.007

Valmaggia, L., Stahl, D., Yung, A., Nelson, B., Fusar-Poli, P., Mcgorry, P., et al. (2013). Negative psychotic symptoms and impaired role functioning predict transition outcomes in the at-risk mental state: a latent class cluster analysis study. *Psychol. Med.* 43, 2311–2325. doi: 10.1017/S0033291713000251

van der Gaag, M., Nieman, D. H., Rietdijk, J., Dragt, S., Ising, H. K., Klaassen, R. M., et al. (2012). Cognitive behavioral therapy for subjects at ultrahigh risk for developing psychosis: a randomized controlled clinical trial. *Schizophr. Bull.* 38, 1180–1188. doi: 10.1093/schbul/sbs105

Van Doesum, N. J., Tybur, J. M., and Van Lange, P. A. M. (2017). Class impressions: higher social class elicits lower prosociality. *J. Exp. Soc. Psychol.* 68, 11–20. doi: 10.1016/j.jesp.2016.06.001

Van Doesum, N. J., Van Lange, D. A. W., and Van Lange, P. A. M. (2013). Social mindfulness: skill and will to navigate the social world. *J. Pers. Soc. Psychol.* 105, 86–103. doi: 10.1037/a0032540

Van Doesum, N. J., Van Prooijen, J. W., Verburgh, L., and Van Lange, P. A. M. (2016). Social hostility in soccer and beyond. *PLoS One* 11:e0153577. doi: 10.1371/journal.pone.0153577

Van Lange, P. A., Ouwerkerk, J. W., and Tazelaar, M. J. (2002). How to overcome the detrimental effects of noise in social interaction: the benefits of generosity. *J. Pers. Soc. Psychol.* 82, 768–780. doi: 10.1037/0022-3514.82.5.768

Van Lange, P. A. M., and Van Doesum, N. J. (2015). Social mindfulness and social hostility. *Curr. Opin. Behav. Sci.* 3, 18–24. doi: 10.1016/j.cobeha.2014.12.009

van Os, J., Kenis, G., and Rutten, B. P. (2010). The environment and schizophrenia. *Nature* 468, 203–212. doi: 10.1038/nature09563

van Os, J., and Linscott, R. J. (2012). Introduction: the extended psychosis phenotype—relationship with schizophrenia and with ultrahigh risk status for psychosis. *Schizophr. Bull.* 38, 227–230. doi: 10.1093/schbul/sbr188

van Os, J., and Reininghaus, U. (2016). Psychosis as a transdiagnostic and extended phenotype in the general population. *World Psychiatry* 15, 118–124. doi: 10.1002/wps.20310

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

Velthorst, E., Fett, A. K. J., Reichenberg, A., Perlman, G., Van Os, J., Bromet, E. J., et al. (2016a). The 20-year longitudinal trajectories of social functioning in individuals with psychotic disorders. *Am. J. Psychiatry* 174, 1075–1085. doi: 10.1176/appi.ajp.2016.15111419

Velthorst, E., Nieman, D. H., Becker, H. E., Van De Fliert, R., Dingemans, P. M., Klaassen, R., et al. (2009). Baseline differences in clinical symptomatology between ultra high risk subjects with and without a transition to psychosis. *Schizophr. Res.* 109, 60–65. doi: 10.1016/j.schres.2009.02.002

Velthorst, E., Reichenberg, A., Kapara, O., Goldberg, S., Fromer, M., Fruchter, E., et al. (2016b). Developmental trajectories of impaired community functioning in schizophrenia. *JAMA Psychiatry* 73, 48–55. doi: 10.1001/jamapsychiatry.2015.2253

Voges, M., and Addington, J. (2005). The association between social anxiety and social functioning in first episode psychosis. *Schizophr. Res.* 76, 287–292. doi: 10.1016/j.schres.2005.01.001

Waltz, J. A., Schweitzer, J. B., Ross, T. J., Kurup, P. K., Salmeron, B. J., Rose, E. J., et al. (2010). Abnormal responses to monetary outcomes in cortex, but not in the basal ganglia, in schizophrenia. *Neuropsychopharmacology* 35, 2427–2439. doi: 10.1038/npp.2010.126

Wechsler, D. (1997). *WAIS-III Dutch Translation*. Lisse: Swets & Zeitlinger.

Wigman, J. T., Van Nierop, M., Vollebergh, W. A., Lieb, R., Beesdo-Baum, K., Wittchen, H. U., et al. (2012). Evidence that psychotic symptoms are prevalent in disorders of anxiety and depression, impacting on illness onset, risk, and severity—implications for diagnosis and ultra–high risk research. *Schizophr. Bull.* 38, 247–257. doi: 10.1093/schbul/sbr196

Wood, S. J., Yung, A. R., Mcgorry, P. D., and Pantelis, C. (2011). Neuroimaging and treatment evidence for clinical staging in psychotic disorders: from the at-risk mental state to chronic schizophrenia. *Biol. Psychiatry* 70, 619–625. doi: 10.1016/j.biopsych.2011.05.034

Woods, S. W., Addington, J., Cadenhead, K. S., Cannon, T. D., Cornblatt, B. A., Heinssen, R., et al. (2009). Validity of the prodromal risk syndrome for first psychosis: findings from the North American Prodrome Longitudinal Study. *Schizophr. Bull.* 35, 894–908. doi: 10.1093/schbul/sbp027

Woudstra, S., Van Tol, M.-J., Bochdanovits, Z., Van Der Wee, N. J., Zitman, F. G., Van Buchem, M. A., et al. (2013). Modulatory effects of the piccolo genotype on emotional memory in health and depression. *PLoS One* 8:e61494. doi: 10.1371/journal.pone.0061494

Yamagishi, T., Takagishi, H., Fermin Ade, S., Kanai, R., Li, Y., and Matsumoto, Y. (2016). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proc. Natl. Acad. Sci. U.S.A.* 113, 5582–5587. doi: 10.1073/pnas.1523940113

Yung, A. R., Phillips, L. J., Yuen, H. P., Francey, S. M., Mcfarlane, C. A., Hallgren, M., et al. (2003). Psychosis prediction: 12-month follow up of a high-risk ("prodromal") group. *Schizophr. Res.* 60, 21–32. doi: 10.1016/S0920-9964(02)00167-6

Yung, A. R., Yung, A. R., Pan Yuen, H., Mcgorry, P. D., Phillips, L. J., Kelly, D., et al. (2005). Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. *Austr. N. Z. J. Psychiatry* 39, 964–971. doi: 10.1080/j.1440-1614.2005.01714.x

Zhu, Y., Zhang, L., Fan, J., and Han, S. (2007). Neural basis of cultural influence on self-representation. *Neuroimage* 34, 1310–1316. doi: 10.1016/j.neuroimage.2006.08.047

Ziauddeen, H., and Murray, G. K. (2010). The relevance of reward pathways for schizophrenia. *Curr. Opin. Psychiatry* 23, 91–96. doi: 10.1097/YCO.0b013e328336661b

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

ORIGINAL RESEARCH
published: 26 June 2019
doi: 10.3389/fnins.2019.00660

Subjective Beliefs About Trust and Reciprocity Activate an Expected Reward Signal in the Ventral Striatum

Kim Fairley[1,2]*, *Jana Vyrastekova*[3], *Utz Weitzel*[3,4] *and Alan G. Sanfey*[2,5]

[1] Institute of Tax Law and Economics, Department of Economics, Leiden University, Leiden, Netherlands, [2] Donders Institute for Brain, Cognition and Behavior, Centre for Cognitive Neuroimaging, Radboud University, Nijmegen, Netherlands, [3] Institute for Management Research, Radboud University, Nijmegen, Netherlands, [4] Faculty of Law, Economics and Governance, Department of Economics, Utrecht University School of Economics, Utrecht, Netherlands, [5] Behavioral Science Institute, Radboud University, Nijmegen, Netherlands

There is overwhelming evidence that the evaluation of both reward decisions and their associated outcomes are closely linked with bilateral activation of the ventral striatum, with these insights stemming from tasks such as the monetary incentive delay task for lotteries and multiround Trust Games for social settings. The essential element in these tasks is an externally provided cue associated with specific gains/trustworthy partners and losses/non-trustworthy partners. However, in reality people typically use their own beliefs to guide their decision-making and assess the likelihood of positive or and negative outcomes. As when participants assess the relationship between cues and rewards, individuals should anticipate rewards in correspondence to their beliefs, i.e., the higher the belief of obtaining a reward in the future, the higher the anticipation of reward. In this study, we use decision-makers' own, naturally occurring, beliefs about both social and non-social contexts to examine the subsequent outcome of their choices. We hypothesize that mechanisms of belief-mediated reward processing are mediated by neural activation in the ventral striatum. An essential feature of our design is the elicitation of individuals' beliefs prior to the decision-making task itself. Furthermore, our incentivized, non-deceptive, decision-making task distinguishes between social – implemented by a Trust Game – and non-social sources, as well as risk and ambiguity as underlying types of uncertainty. Our main result shows that individual beliefs regarding reciprocity likelihoods in both the Trust Game and the lottery influence the amount invested. Subsequently, only the investment amount in the Trust Game parametrically modulates anticipatory reward and outcome evaluation in the ventral striatum. This study demonstrates a first approach at using participants' subjective sets of beliefs to examine reward processing. We discuss its potential promise, outline some limitations, and propose follow-up studies to extend the current approach.

Keywords: trust, beliefs, reward, risk, uncertainty, ventral striatum

OPEN ACCESS

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

Reviewed by:
Frauke Nees,
Central Institute for Mental Health,
Germany
Elsa Florence Fouragnan,
University of Plymouth,
United Kingdom

****Correspondence:***
Kim Fairley
k.fairley@law.leidenuniv.nl;
kim.fairley@gmail.com

Specialty section:
This article was submitted to
Decision Neuroscience,
a section of the journal
Frontiers in Neuroscience

Received: *30 January 2019*
Accepted: *07 June 2019*
Published: *26 June 2019*

Citation:
Fairley K, Vyrastekova J, Weitzel U and Sanfey AG (2019) Subjective Beliefs About Trust and Reciprocity Activate an Expected Reward Signal in the Ventral Striatum.
Front. Neurosci. 13:660.
doi: 10.3389/fnins.2019.00660

INTRODUCTION

Decision-making under conditions of uncertainty, that is, when we do not know the exact future outcome of our choices, is generally guided by beliefs we have about the world at large (Savage, 1954). These beliefs typically act as subjective probabilities, derived from a combination of specific prior knowledge, received information, and expertise in a particular domain

(Fox and Tversky, 1995; Fox and Weber, 2002). For example, imagine you want to invest part of your capital in a mutual fund. There are thousands of options, but after reading many sources of information you select a few mutual funds. You will only learn if your assessment of these selected mutual funds was satisfactory after evaluating their quarterly holdings. While this is a difficult task in itself, the choice process can be even more complicated when decisions are made in direct interaction with another person, for example when we are deciding to trust or reciprocate with other people. To extend our previous example, imagine, as an alternative to investing in mutual funds, you opt to invest part of your capital in some startup businesses. You carefully decide which proposals to select based on your expectations about which entrepreneurs will successfully execute their business plan. However, you will only learn if your expectations were correct when, after several months or years, you receive the financial statements of the companies you funded. In particular, you are interested in reviewing the projects you expected to do well and for which you anticipated a high return on investment.

Decision-makers only ever get a full picture of the relationship between their beliefs and decisions by examining the eventual outcomes of these choices, which then offer the opportunity to learn whether their initial expectations were met, or whether they had in fact incorrectly assessed them. Prior to learning the actual outcomes however, one can imagine that anticipated rewards might increase in correspondence to an individual's beliefs, i.e., the higher the beliefs of obtaining a reward in the future, the higher the anticipation of reward.

Though many studies have investigated the neural underpinnings of reward anticipation and outcome delivery with tasks such as the monetary incentive delay task (MID; Knutson et al., 2000) or the card guessing task (Delgado et al., 2000) in a lottery context, or a repeated Trust Game in a social context, to the best of our knowledge there has been no exploration to date of using decision-makers' own, subjective, beliefs about the evaluation and subsequent outcome of their choices. Therefore, in this study we investigate reward anticipation and outcome when the reward "cue" is a function of prior internal evaluations as opposed to the standard method of using an externally provided cue association.

When we refer to beliefs, specifically we mean participants' inherent priors, which are not manipulated in any way in this study, but which have formed based on previous personal, likely idiosyncratic, experiences. Our procedure is therefore different from studies which explicitly let participants form priors based on some experimental interaction, for instance a ball-tossing game with fictional players and specific behaviors (Fareri et al., 2012) or vivid descriptions of partners' life events as to establish moral characters (Delgado et al., 2005). Furthermore, our study does not investigate social learning *per se*, as here the decision-making and outcome phase are separated in time (Fareri et al., 2012, 2015). After the outcome phase, participants review their prior decisions, but, importantly, do so without the possibility to change these previous choices. This has the effect of minimizing learning processes that may take place during the experimental process, as this is not the key feature of our study.

In tasks like the aforementioned MID and card guessing task, the decision-maker must perform a certain action correctly – a rapid reaction time in the MID and a correct guess in the card task – in order to receive a monetary reward. The essential feature of these games is that, before the required action, players learn that certain visual cues are associated with specific gains or losses, indicating either how large the monetary reward is or how much they can avoid losing if they perform the required task successfully. In the social domain, similar cues can be provided to denote a good or bad social partner. For instance, in Fouragnan et al. (2013), participants were told that triangles indicated game partners with low scores in a social orientation task, whereas circles indicated high social orientation scores.

Here, we are interested in naturally occurring individual beliefs, not induced by establishing specific cue-outcome relations. We examined these beliefs in the context of a decision-making task which distinguishes between both sources and types of uncertainty. We define sources here as uncertainty measured in social and non-social settings, which we operationalize with a Trust Game and a lottery mechanism, respectively.

In the Trust Game a sender invests a certain amount with a receiver based on beliefs she has regarding the receiver's likely reciprocation, and therefore tries to reason about her partner's trustworthiness. In the lottery context, the investor will analyze how much to invest with a random mechanistic device and is likely to use introspection, based on (any) experience with outcomes decided via such mechanisms, for example roulette or a coin toss. By using participants' own belief sets it could be that participants rely more on these beliefs in a social context (Chang and Sanfey, 2011). That is, for example, correct beliefs regarding lottery outcomes are perceived as good luck, yet correct beliefs in a social situation are more likely perceived as a signal of personal success in properly assessing the social situation (Trautmann et al., 2008). Therefore, we are interested here in investigating whether social and non-social sources of uncertainty may influence belief-mediated anticipatory rewards in different ways.

In addition to exploring the relative *sources* of uncertainty, our study also distinguishes between *types* of uncertainty. By types of uncertainty, we refer to risk and ambiguity, which are events characterized by known objective probabilities and unknown probabilities, respectively (Wakker, 2010). A few studies have focused on the neural differences of anticipated rewards when cue-reward pairs are associated with either known probabilities (risk) or unknown probabilities (ambiguity). These studies show a distinct pattern of brain activation between anticipatory rewards under conditions of risk vs. ambiguity (Volz et al., 2003; Tobler et al., 2006), and are in line with primate studies which show that dopaminergic modulation of rewards varies across probability distributions (Fiorillo et al., 2003). By employing two types of uncertainty in this study, we can investigate both anticipated rewards that are a function of participants' subjective beliefs (ambiguity), as well as objective probabilities we provide (risk).

In humans, the neural mechanisms of both the evaluation of reward decisions and their associated outcomes are mostly observed by bilateral activation of the ventral striatum (Bartra et al., 2013). This activity has been observed in a wide variety of outcome modalities. For example, activation in the ventral striatum, whose axons receive dopaminergic input from the ventral tegmental area (VTA) in the midbrain (Schultz, 1998), has been observed for monetary rewards (Knutson et al., 2001; Knutson and Greer, 2008), food (O'Doherty et al., 2002; Hare et al., 2008, 2009), social cooperation (Rilling et al., 2004; Davey et al., 2009; Jones et al., 2011; Korn et al., 2012; Lin et al., 2012; Powers et al., 2013) and even the punishment of others (Singer et al., 2006). Relatedly, in multiround trust games, Bellucci et al. (2017) found in a meta-analysis that the decision to trust also activated the ventral striatum, which they inferred to be likely associated with reward prediction error signals. However, during the feedback stage of this task, the dorsal striatum was active, which according to the authors was likely related to reinforcement learning processes.

In a similar manner to how anticipatory reward mechanisms operate when a previously learned cue is presented, we expect that people anticipate rewards when awaiting outcomes of decisions that were mediated by specific internal beliefs. When the investor in our earlier example anticipates a higher return from certain business projects, we would predict that these expectations would lead to increased reward anticipation prior to learning how these particular projects fared. Mechanistically, we hypothesize that this process is mediated by activation in the ventral striatum when participants are anticipating the potential outcome of their rewards. Though anticipating rewards in both social and non-social settings are thought to be processed in the striatum (Lin et al., 2012), our earlier hypothesis that participants might rely more on their beliefs in a social context could imply that we find stronger activation in the ventral striatum comparing between the Trust Game and a matched lottery task. With regard to types of uncertainty, there is evidence that predicting outcomes under various levels of uncertainty as compared to certainty activates the ventral striatum bilaterally (Volz et al., 2003). Therefore, with regard to ambiguity, we hypothesize greater ventral striatal activation for the anticipation of ambiguous as compared to risky outcomes. Lastly, during outcome delivery, we expect to observe activation in the ventral striatum as a function of the magnitude of the reward, that is, as a function of participant's own earlier investment choices.

To examine this question experimentally, namely the neural mechanisms of belief-mediated anticipatory rewards and reward outcomes, an essential feature of our design is the careful elicitation of individual beliefs prior to decision-making. If we observe that participants' decisions are indeed guided by their beliefs, we can then investigate the associated neural response as participants await and receive the respective outcomes. Importantly, this also optimally requires a clear and non-deceptive incentive scheme, as dopaminergic modulation is primarily observed when rewards are actually valuable in an uncertain environment (Schultz, 2010).

Taken together, this study aims to test how internally constructed beliefs, as opposed to objective cue-outcome associations, impact the neural mechanisms of reward anticipation and the subsequent delivery of rewards. Based on substantial pre-existing evidence that both reward anticipation and reward receipt are coded in the ventral striatum (Bartra et al., 2013), we hypothesize that both belief-mediated anticipatory rewards as well as reward receipt itself will activate the ventral striatum. We explore this question using a novel incentivized decision-making task that distinguishes between both types and sources of uncertainty.

MATERIALS AND METHODS

Participants

A total of 26 participants (mean age = 22, 50% female) were recruited for this study via the online recruitment system SONA of the Donders Institute for Brain, Cognition and Behavior. Students with a psychology or economics background were excluded due to concerns about, respectively, suspicions regarding the veracity of the actual social interaction and a prior detailed understanding of game theoretic behavior.

Three of the 26 participants were excluded from our sample prior to analysis. One participant said that he did not believe the real human interaction and the incentive scheme after the experiment. Data for two participants were lost due to technical issues; the head coil was not applied correctly and the MRI data was not transferred appropriately. Furthermore, three participants were removed after analyzing all behavioral data as responses were very erratic, differed more than two standard deviations from mean responses and revealed clear misunderstandings (e.g., betting on scenarios with 0% chance to win) in one case and no variation in investment levels in the other two cases. Therefore, unless explicitly noted, analyses reported here are based on 20 participants (mean age = 22, 11 females and 9 males). Finally, this study was approved by the local ethical committee.

Design and Procedures

The full experiment consisted of two parts, a decision phase and an outcome phase, separated by a short break. The focus of this manuscript is on the outcome phase. As the outcomes stem from the decision-making phase, we explain the setup below to be able to explain how outcomes were presented to participants.

On each trial, participants received an endowment of 10 tokens (which were later exchanged for cash). Participants could decide to invest any number of these tokens in either a human partner (social source) or a lottery (non-social source), depending on the experimental condition, with the investment amount then tripled by the experimenter. Additionally, there were two different types of uncertainty regarding the likelihood of their investment being repaid, that of risk and of ambiguity. This resulted in a total of four experimental conditions, explained in detail below.

In the *social* condition, we employed a standard Trust Game (Berg et al., 1995). The fMRI participant, termed the sender, had their (tripled) investment transferred to another player, known as the receiver. This receiver could then decide to either keep all this investment, or return half of it to the sender. If half was sent back, the sender was obviously better off than if they had transferred nothing, but at the time of decision faced uncertainty as to whether the receiver would reciprocate his or her trust. In the social context, participants placed an investment under two types of uncertainty: they explicitly knew the probability of being paired with a reciprocating receiver, known as the risky trust game (RTG), or they did not receive any probabilistic information regarding reciprocity, known as the ambiguous trust game (ATG).

Receivers' choices were collected during a behavioral session prior to the fMRI experiment. Receivers made a binary choice to either return or keep the investment should a positive investment be received from the sender. Receivers could not condition their choice on the different investment amounts the sender could potentially invest with the receiver. Thereby our fMRI participants, in their role as sender, only acted upon beliefs regarding receivers' trustworthy behavior, and their decisions were not confounded by other potential motives, for example signaling trust to receivers (McCabe et al., 2003) or eliciting positive reciprocity (Houser et al., 2010).

In the *non-social* condition, participants' outcomes were resolved via a typical Ellsberg lottery design (Ellsberg, 1961). They bet on the color of a marble drawn from an urn, with this marble either a "winning" or "losing" color. Again, the fMRI participant decided on a transfer, receiving back either half of the tripled investment (if a winning colored marble was drawn), or alternatively losing their entire investment (if a losing colored marble was drawn). In this condition participants also faced two types of uncertainty. In the risky lottery (RLOT), participants knew the probabilities of drawing a marble with a winning color, whereas in the ambiguous lottery (ALOT) this probability was unknown.

We created risky and ambiguous trials in both social and non-social contexts by introducing a group principle to the general feature of the games discussed above. In the Trust Game, we grouped nine decisions made by nine different receivers. One receiver was randomly drawn from the pool of nine and matched to the MRI participant's investment choice. In the lottery, there were nine marbles in the urn. One randomly drawn marble from this urn determined if the participant received half of his tripled investment.

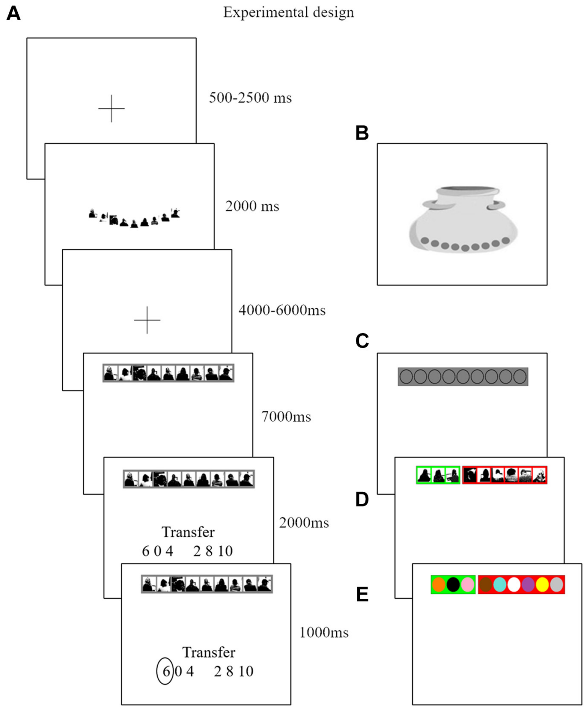


FIGURE 1 | Each trial consists of six screens. Panel **(A)** is an example of a trial from the ATG. The second screen indicates the source of uncertainty. Nine silhouettes are displayed when participants are in a social context. Nine marbles are displayed when participants face a lottery context Panel **(B)**. The fourth screen is the decision screen. They are instructed to decide how much to transfer here. As the six possible transfer options appear in a random order on the next screen, they are unable to prepare for a specific button press. On the last screen we confirm their choice. In the ATG Panel **(A)** nine silhouettes on a gray background indicate that no information is given about the distribution of receivers that decided to send back half or keep the investment. To illustrate the tailor-made structure of our design, we assume a participant who believes three out of nine receivers will reciprocate. In the ALOT Panel **(C)** the participant receives instruction that three out of nine colors that can be used in any combination in this lottery are winning colors. In this way we align underlying subjective probabilities between the ATG and ALOT. In the risky trials we align individual's beliefs to objective probabilities. A participant who believes three out of nine receivers will reciprocate, will most often face a RTG, which is composed of three receivers (green background) that decided to send back half of any received investment versus six receivers (red background) that decided to keep their investment Panel **(D)**. Finally, in the RLOT the urn is composed of all nine colors out of which three are winning colors (green background) and six are losing colors (red background) Panel **(E)**.

In the social context participants have underlying prior beliefs about the reciprocal behavior of receivers in general, and receive the following information as part of the instructions. We provided basic information regarding the pool of trustees, e.g., age, gender, study, hobbies – which were answered by the trustees after they had placed their reciprocating decision. Any difference fMRI participants, in their role as trustor, reveal about trustees' reciprocating behavior is based on the same information all of them received and is therefore likely the result of differences in reciprocating behavior in general. Therefore it is important that we control for these beliefs in order to rule out inconsistencies in these underlying likelihoods and objective probabilities across our four experimental settings. For instance, imagine a sender who thinks that five out of nine receivers are likely to reciprocate. If this participant is confronted with a RTG where six out of nine receivers decided to transfer back half of the investment, we cannot assess whether differences in investment behaviors between both scenarios are caused by the type of uncertainty, or by a mismatch between subjective probability of 5/9 in the ATG and the objective probability of 6/9 in the RTG. Therefore, we elicited individual beliefs in the ATG before participants made decisions in our experimental setting. With an incentive-compatible belief elicitation technique (quadratic scoring rule, e.g., see Schlag et al., 2015), we asked how many receivers out of the pool of nine they thought would reciprocate their investment. This belief is then used to present participants with *belief-corresponding* scenarios in the experimental settings. Essentially, individual beliefs entailed a

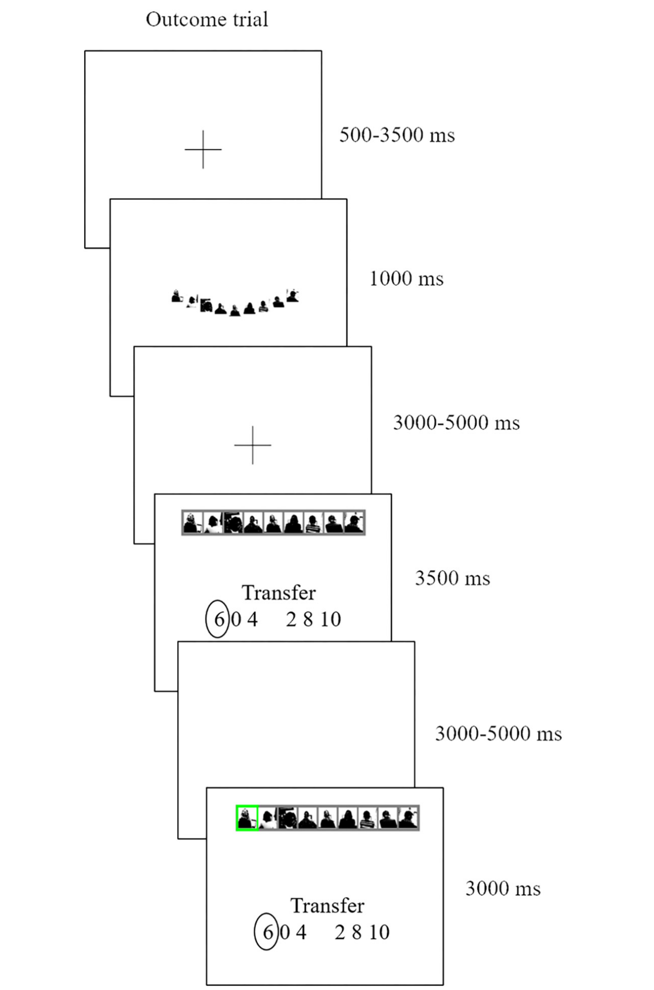


FIGURE 2 | An outcome of an ATG trial. We took pictures of receivers while they were seated behind a laptop. The pictures only show receivers' silhouettes in black and white and no facial features are shown.

tailor-made trial structure for each participant (see **Figure 1** for an overview and example of the experimental setup). By implementing this feature, we made sure that beliefs are aligned in our four settings. This enabled us to investigate expected reward signals by examining the effect of both source and type of uncertainty, taking into account participants' naturally occurring beliefs.

To reiterate, we focus here solely on the outcome phase, that is, the revealing of decision (either trust or lottery) outcomes after all decisions have been made. Participants passively reviewed their previously made choices and then saw the respective outcome (see **Figure 2** for a trial). During this outcome phase our primary focus is on the 3500 ms time period when participants are reminded of their earlier investment choice, and then await the outcome. We term this moment the *anticipation screen.* They then see the actual outcome of that trial, when a randomly selected receiver (social condition) or marble (non-social condition) is selected (final screen **Figure 2**, henceforth referred to as the *outcome screen)*. A selected receiver or marble highlighted in green indicates a winning trial, and when colored red indicates a losing trial.

Receivers' decisions were collected during behavioral sessions, which took place at the Nijmegen School of Management decision laboratory. The fMRI experiment took place at the Centre for Cognitive Neuroimaging at the Donders Institute for Brain, Cognition and Behavior. The fMRI task was presented using Matlab Psychtoolbox (Kleiner et al., 2007). Participants read instructions and performed a belief elicitation task as part of the instructions (75 mins in total) before they were placed in the MRI scanner for approximately 60 min. The fMRI experiment consisted of the decision-making phase and outcome phase. After the decision-making phase, they saw a total of 88 outcome trials in the scanner, equally divided between trust and lottery outcomes (during the decision-making phase, participants also made choices when the chance of reciprocation was 0%, respectively, 100% chance. We excluded these decisions during the outcome phase as there is no uncertainty and thus no influence of individuals' beliefs regarding their outcome). There were 15 outcome trials for each experimental condition and in addition there were filler trials for other probabilities in the RTG and RLOT that did not match participants' beliefs. This provided greater variety in decision contexts, and also made it more difficult for participants to assess the individually tailor-made structure.

The outcomes were presented in 18 blocks, with each block consisting of five trials of either trust or lottery outcomes (four outcome trials for block 17 and 18). Within each block, both risky and ambiguous trials were presented in a random order. To enhance attention to the outcome phase, we introduced payment screens. After every two blocks, two outcomes were randomly selected, one from the lottery and one from the trust condition, which counted toward participants' earnings. Each token was converted to 10 eurocents.

After the experiment subjects were paid out in cash dependent on their choices and randomly selected outcomes, and the accuracy of their stated beliefs. Notably, no deception was used in this experiment. Please see the appendix for the instructions and a detailed explanation of the payment scheme.

Image Acquisition and Preprocessing

Functional neuroimaging data was collected on a 3-Tesla Siemens MRI system (Skyra) at the Donders Centre for Cognitive Neuroimaging in Nijmegen, Netherlands. Images were acquired using a 32-channel head coil, with a standard multi-echo imaging pulse T2*-weighted sequence (field of view = 224 mm, matrix = 64 × 64, repetition time (TR) = 2390 ms; echo times (TE) = 9.4, 20.6, 32.0, 43.0, and 54.0 ms, flip angle = 90°, slice gap = 0.5 mm). Using a multi-echo sequence provides a better signal-to-noise ratio for brain areas susceptible to dropout, while allowing for scanning of the whole brain (Poser et al., 2006). One whole-brain volume consisted of thirty-one ascending slices (slice thickness = 3.0 mm, voxel size = 3.5 mm × 3.5 mm × 3.0 mm).

For each participant we acquired a high-resolution anatomical T1-weighted image (MPRAGE; 192 slices; TR = 2300 ms, voxel size = 1 mm × 1 mm × 1 mm). Participants' heads were loosely taped to the coil within the scanner in order to limit movement during image acquisition.

fMRI data analysis was performed using SPM12 (Statistical Parametric Mapping; Friston et al., 2007). Prior to preprocessing we combined and realigned the five read-outs acquired via the multi-echo sequence by using standard procedures described by Poser et al. (2006). The first five volumes, acquired prior to task initiation, were used to estimate the weighted echo time per voxel for optimal echo combination including allowing T1 equilibration effects. These five volumes were then discarded from the analysis (Poser et al., 2006). After echos were combined, preprocessing consisted of slice-timing to the middle slice, co-registration of the functional images to the anatomical images, segmentation of the functional and anatomical image, and normalization to the Montreal Neurological Institute (MNI) template using the segmentation parameters. Functional images were then smoothed with a Gaussian kernel of 8 mm full-width at half maximum (FWHM).

Data Analyses

Behavioral Parameters

In this study we were interested in the question of whether decision-makers' beliefs about the outcomes of their choices would act as a cue for reward anticipation, and whether this might differ across conditions. In the RTG and RLOT participants do not face uncertainty as they receive objective probabilities (in line with the beliefs we elicit in the ATG), which naturally act as cue for reward anticipation.

In the ambiguous social context (the ATG), we elicited beliefs regarding the reciprocity of receivers before participants made investment decisions during the experiment. During the decision-making phase of this experiment, we observe participants' actual investment choices and assume they stem from their individual subjective beliefs. It is therefore crucial that we establish a relationship between participants' *a priori* beliefs and the investment choices they make in the ATG and the ALOT. Therefore, we will first examine whether indeed participants base their investment choices in the ATG and ALOT on their subjective expectations, and subsequently test whether participants' investment levels different across our experimental conditions. These analyses consist of a linear mixed effects model (estimated with the toolboxes lme4 and lmerTest in R). The results section details the variables, random intercepts, and slopes included in this model.

Neuroimaging Analyses

To study the neural mechanisms of reward anticipation and outcome delivery, the primary explanatory variables (EV) of our general linear model (GLM) examined the BOLD response during trials in which participants reviewed their previously made choices and awaited their outcome (fourth screen in **Figure 2**). Four EV's indicated the onset of the anticipation screens, modeled for a duration of 3500 ms, when participants reviewed decisions from the RTG (belief-corresponding risky trials), ATG, RLOT (belief- corresponding risky trials), and ALOT. To examine whether participant's investment behavior served as a cue that would trigger expected rewards, we added this variable as parametric modulator to these four EV's.

Other EV's in this model included the other review decisions from the RTG and RLOT filler trials (not corresponding to participants' beliefs), the trust or lottery cue (second screen in **Figure 2**), trials in which participants had not made a choice within the required 2 s (modeled at the onset of the anticipation screen for the full duration of the remainder of the trial), one outcome screen that coded a "win" (investment gets transferred back), one outcome screen that coded a 'loss' (participant loses investment), and finally one EV that modeled the nine payment information screens. The remaining events are the fixation and blank screen, which are therefore considered the implicit baseline.

When we were specifically interested in analyzing the BOLD responses of the actual outcomes, separated as wins and losses, we added the investment choices as parametric modulators to the outcome period, and entered these as the first variables to our model, otherwise similar as the model discussed above, in order to allow for sufficient explanatory variance regarding these parametric modulators.

All regressors were modeled with a canonical hemodynamic response function. To account for motion, we included the six head movement parameters together with their squared value and the temporal derivatives as nuisance regressors. A standard high-pass filter (cut-off 128 s) and auroregressive AR (1) model were used during the GLM analysis to account for possible slow-frequency drifts and temporal autocorrelation, respectively.

Our primary contrasts of interest are the anticipated rewards, as a function of the earlier chosen investment levels, re-evaluated during anticipation compared to implicit baseline, the specific neural mechanisms of anticipating outcomes as a function of source (social: anticipation ATG and RTG as compared to non-social: anticipation ALOT and RLOT), and comparing types of uncertainty (risk: RTG and RLOT vs. ambiguity: ATG and ALOT). Furthermore, we examine the amount won (lost) during the outcome phase, indicated by the investment level being reciprocated (held), compared to implicit baseline.

For the specified contrasts outlined above, one-sample *t*-tests were performed as second-level models to analyze group effects. Participants' beliefs were added as a covariate at the group level. Statistical maps with an initial threshold of uncorrected $p < 0.001$ were established and were subsequently corrected for multiple comparisons using a Family Wise Error corrected cluster threshold of $p < 0.05$. As our hypotheses are centered on the role of the striatum during belief-mediated anticipation and outcome, we apply a small volume correction based on an *a priori* region defined by meta-analysis (Bartra et al., 2013), using specific coordinates for left striatum [−12, 12, −6] and right striatum [12, 10, −6], each with a radius of 10 mm.

Finally, the raw data and code used here will be made available by the authors to any qualified researcher.

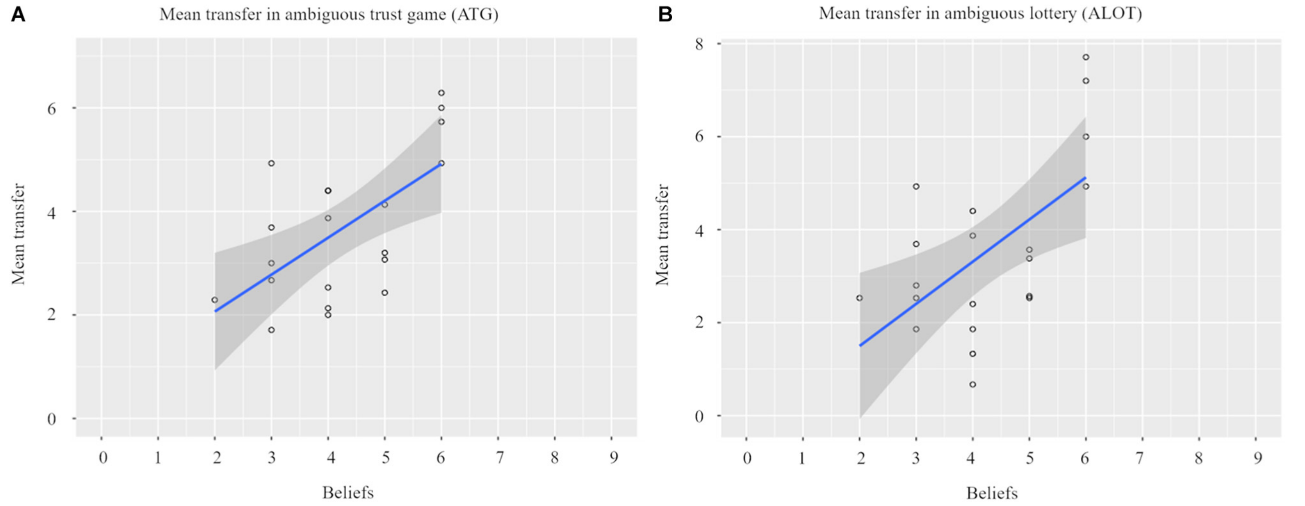


FIGURE 3 | Elicited beliefs regarding receivers' reciprocity influenced chosen transfer in the ambiguous trust game Panel **(A)**. Based on individual beliefs, participants received a matching amount of winning colors in the ambiguous lottery. Participants used this information, given during instructions prior to the experiment in the MRI scanner, as transfer positively increased as a function of the amount of winning colors Panel **(B)**.

RESULTS

Beliefs and Decision-Making Under Ambiguity

Individual beliefs regarding the likelihood that receivers will reciprocate varied substantially. Some participants indicated quite low belief in receiver reciprocity, expecting only two or three of nine receivers to reciprocate their investment. On the other hand, some participants believed that six of nine receivers would return their investment.

Figure 3A illustrates that individual beliefs, elicited prior to the investment choice, positively correlated with the amount they subsequently invested in the ATG (*Pearson's* $r = 0.620$, $p = 0.004$). That is, the larger the number of reciprocators that our participants thought would be present in a group of nine receivers, the more tokens they were willing to invest.

We also found a positive relationship between the amount of winning colors and participants' investment choices in the ALOT (*Pearson's* $r = 0.587$, $p = 0.006$, see **Figure 3B**). Thus, as expected, in both social and non-social contexts, the higher the subjective probability of receiving half of the tripled investment back, the more tokens participants were prepared to invest.

Although these results may appear intuitive, they are important for the neuroimaging analyses. When we add participant's investment choices to our fMRI models we can reliably state that these investments are guided by their individual beliefs. Any difference we find across conditions is therefore unlikely to be the result of a mismatch between subjective probabilities (based on participants' beliefs from the ATG), the underlying likelihood in the ALOT, or objective probabilities in the risk treatments.

Participants' beliefs also interacted with our experimental conditions resulting in interesting investment patterns in the Trust Game and lotteries. In a companion paper we focus exclusively on the decision-making phase and present its neuroimaging analyses – here we only look at the outcome phase in relation to beliefs – but for clarity we provide a short behavioral overview of investment behavior here. The mean transfer in the experiment, across conditions and subjects, was 3.83 tokens. In **Figure 4**, participants' transfers are shown across conditions. In general, participants invested more in the risky conditions than in the ambiguous conditions, illustrating ambiguity aversion. This general pattern, however, was strongly influenced by individual beliefs, namely that the higher were beliefs regarding reciprocity

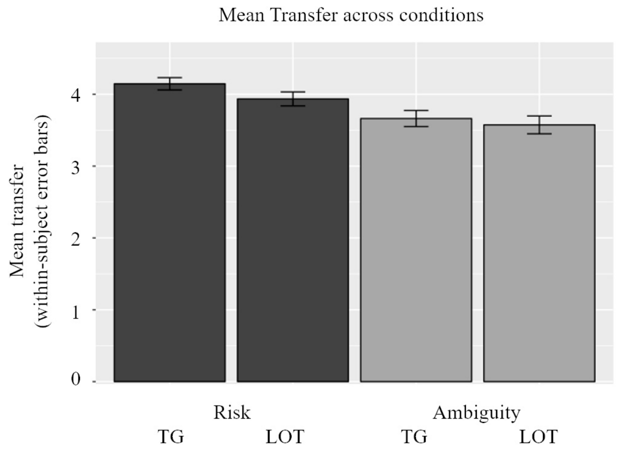


FIGURE 4 | Overview of transfer choices across conditions. Participants invest less in the ambiguous conditions [both in Trust Game (TG) and lottery (LOT)] than the risky conditions. This effect is highly influenced by beliefs. The general pattern of ambiguity aversion only holds for participants with low beliefs in trustees (and number of winning colors). There was no effect of sources of uncertainty: participants do not alter their transfer between the TG and the LOT.

in the ATG (and number of winning colors in the ALOT), the more ambiguity averse behavior was displayed. This result is similar to findings from experimental economics, which show variability in ambiguity aversion along the probability distribution (Trautmann and van de Kuilen, 2014). These results are confirmed by a linear mixed-effects model which consisted of participants' transfers as the dependent variable, and type (risk vs. ambiguity) and source (Trust Game vs. lottery) of uncertainty as independent factors, along with gender, participants' beliefs, trial number, and an interaction of beliefs and both experimental factors. A random intercept and two random slopes accounted for clustering at the participant level and repeated trials within experimental conditions. Confirming the bivariate correlation between beliefs and investment choice, the mixed effects model underlined the significance of participants' beliefs ($\beta = 0.891$, $p = 0.002$ via Satterthwaite's method) and their interaction with the type of uncertainty ($\beta = 0.538$, $p = 0.025$ via Satterthwaite's method). Although the variable trial number was also negatively significant ($p = 0.027$) – indicating that as participants progress through the experiment they transfer less – its economic significance was rather small ($\beta = -0.006$).

Imaging Data

We first focused on the observed neural activity during the anticipation phase. To check whether participants were actually observing their previously made choices, we first examined the BOLD signal during all anticipation screens (fourth screen in **Figure 2**). Various brain regions were active – parahippocampal cortex (peak activation: −8, −52, 0, 1282 voxels, $p < 0.001$), dorsolateral prefrontal cortex (peak activation: −32, 10, 56, 63 voxels, $p = 0.007$), and control network (peak activation: −50, −52, 38, 64 voxels, $p = 0.007$) – which, based on the NeuroSynth database (Yarkoni et al., 2011), are very likely to be involved in attentional and memory processes.

More importantly, we then added participants' investment level as parametric modulator, allowing us to ask whether trials on which the most tokens were invested showed a greater expected reward signal while participants reviewed their chosen investment prior to seeing the outcome. When we focused on investment choices during the anticipation phase across all experimental conditions (ALOT, RLOT, ATG, RTG), no subthreshold clusters were found. We then looked at the social and non-social anticipatory outcomes separately. Analyses here demonstrated that the more that was invested in the Trust Game, the more activation was observed bilaterally in the ventral striatum (peak activations: −4, 7, −7 and 6, 4, −7, 11 voxels, $p = 0.025$ after small volume correction, see **Figure 5**), whereas no regions surpassed this threshold during the lottery outcomes. A direct comparison of investment levels in the Trust Game versus the lottery also revealed an area in the ventral striatum bilaterally, as part of an area which extended into the orbitofrontal cortex (peak activations: −8, 21, −4 and −15, 35, −7, 18 voxels, $p = 0.020$ after small volume correction).

Next, we explored the different types of uncertainty anticipation, namely comparing risky versus ambiguous trials, but found no significant neural effects for this contrast. Even when we restricted the analysis to a functional ROI based on the contrast which described investment levels between Trust Game and lottery, we did not observe activation in this area.

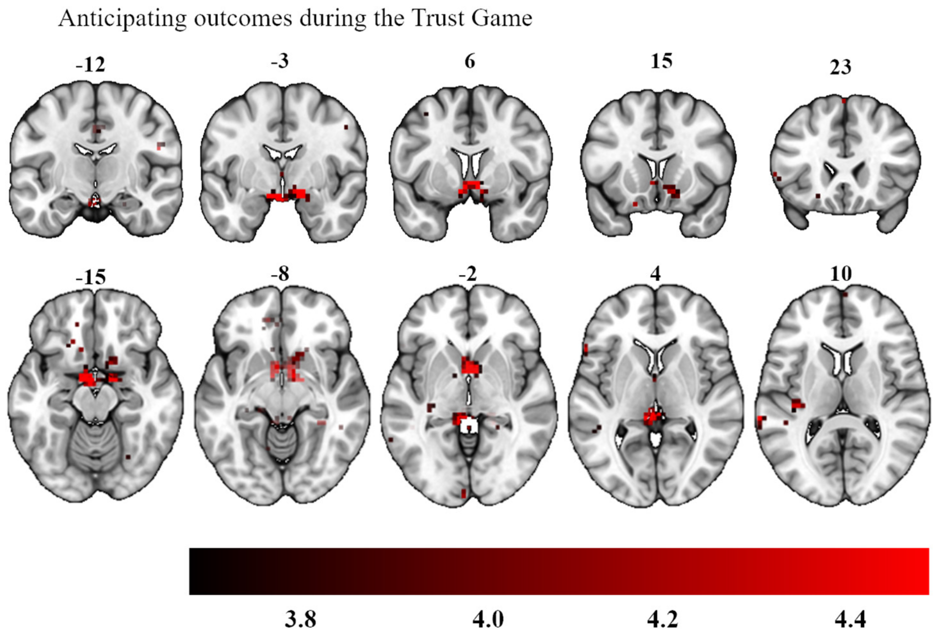


FIGURE 5 | Bilateral activation in the ventral striatum as a function of participants' investment choices during the anticipation of their social outcomes. The color map defines the strength of the contrast's *T*-scores (Statistical maps with an initial threshold of uncorrected $p < 0.001$ were established and were subsequently corrected for multiple comparisons using a Family Wise Error corrected cluster threshold of $p < 0.05$).

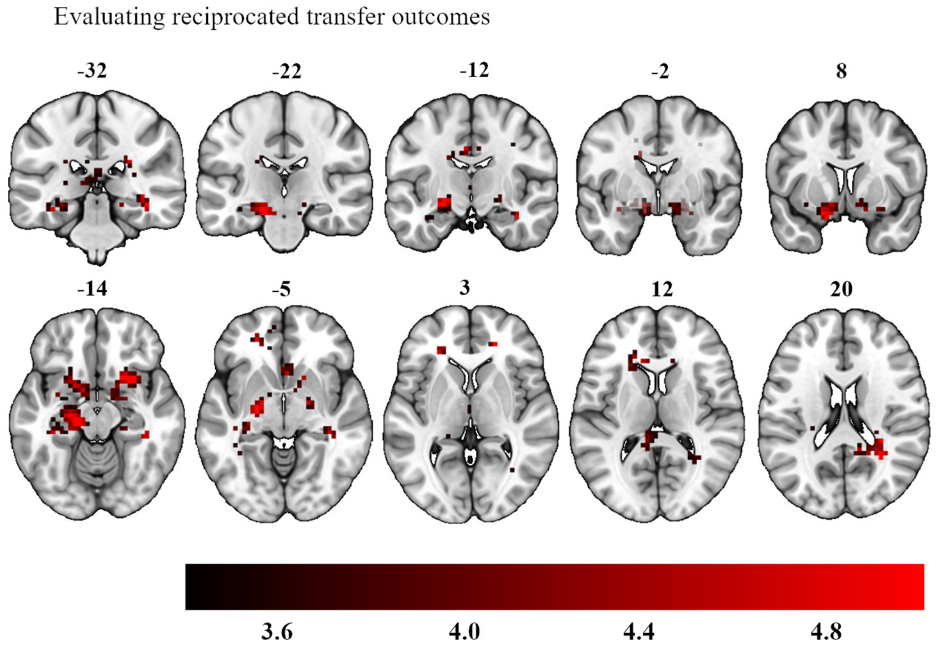


FIGURE 6 | Bilateral activation in the amygdala, left ventral striatum and right dorsal striatum as a function of participants' investment choices during the delivery of their outcomes. The color map defines the strength of the contrast's *T*-scores (Statistical maps with an initial threshold of uncorrected $p < 0.001$ were established and were subsequently corrected for multiple comparisons using a Family Wise Error corrected cluster threshold of $p < 0.05$).

We also examined the question of neural differences when outcomes were finally resolved. We investigated the investment amount as a parametric modulator when experiencing a win during the outcome phase, collapsed across experimental conditions (last screen in **Figure 2**). This contrast yielded strong bilateral activation in an area encompassing the amygdala bilaterally, left ventral striatum and right dorsal striatum (left hemisphere peak activations: −22, −14, −10 and −18, 7, −18, 193 voxels, $p < 0.001$ whole brain analysis; right hemisphere peak activations: 20, −7, −7 and 20, 18, −10, 84 voxels, $p = 0.003$ whole brain analysis, see **Figure 6**). The investment amount as parametric modulator for a loss did not show any significant activation patterns.

We further investigated whether individual differences in attitudes toward social and ambiguity preferences might explain variance in neural data. Individuals' social preferences were defined as a normalized score between −1 to 1 where a score above (below) 0 indicated a person who invested more (less) with a person in the TG than the lottery. Individuals' ambiguity preferences were also defined as a normalized score between −1 to 1 where a score above (below) 0 indicated a person who was ambiguity averse (seeking). When we added social preferences as a covariate to the contrast which investigated neural differences in investment levels in the Trust Game versus the lottery, we observed the right motor and somatosensory cortex activation ($p = 0.015$ whole brain). Individuals' ambiguity preferences as covariates for the contrast investment levels in the ambiguous versus the risky settings did not yield any significant neural findings.

DISCUSSION

Reward is an important and well-studied topic in the field of Neuroscience (Bartra et al., 2013). Initiated by innovative primate studies, a growing literature has emerged examining the putative dopaminergic modulation of reward (Schultz et al., 1997; Schultz, 1998). Our study sought to address scenarios when anticipated rewards stem from individuals' own beliefs and subsequent decision-making, instead of relying on cue-outcome associations that are typically evident in tasks such as the MID and multiround Trust Games. In this experiment, we examined the strength of individual beliefs, their relationship with subsequent decisions, and their associated neural mechanisms when anticipating their outcomes. These questions were explored in a real-life decision-making context, in which outcomes were clearly (and non-deceptively) resolved. We asked whether these belief-mediated anticipated rewards were neurally processed in the manner of an expected reward signal, similar to how rewards are evoked through abstract cue-outcome associations.

Our decision-making task distinguished between social (Trust Game context) and non-social (lottery context) sources of uncertainty, as well as risk and ambiguity as types of uncertainty. Choices made by participants in both the Trust Game and the lottery tasks indicated clearly that underlying beliefs did in fact guide participants' decision-making. Participants invested more when they expected a greater number of their potential game partners to reciprocate their investment in the ATG. Similarly, participants in the ALOT invested most when they knew a greater number of colors out of the nine possible colors would lead to a return on their investment. Subsequently, individuals' investment behavior is also influenced by their beliefs: the higher beliefs regarding reciprocity were in the ATG (and number of winning colors in the ALOT), the more ambiguity averse behavior was displayed.

Our neuroimaging analyses then focused on whether these belief-related expectation signals were evident in brain regions related to standard cue-based reward anticipation. We found confirmatory evidence of this in the Trust Game. The greater the expectation of receiving a back-transfer in the Trust Game, the greater the investment amount that was made, and in turn the greater the activation in bilateral ventral striatum prior to the outcome being presented, as compared to anticipation in the lottery context. Anticipating the outcome of whether your investment is reciprocated by another person versus a lottery is likely more salient as it depends on subjective assessments of trusting and engaging with another people and its outcome results from their intentional behavior, which aspects are of course absent when interacting with a mechanistic device. Also, one consequence of our experimental approach is that participants in the ALOT did not actively have to form a prior belief. A feature of dopaminergic modulation of reward is that the more uncertain a reward is, the more information the consequent outcome will allow for updating of priors (Schultz, 2010). Although a different ambiguous urn was constructed on every trial in the ALOT, participants knew how many colors were winning colors. This feature might have reduced the uncertainty in the ALOT as compared to the ATG.

Our novel result illustrates that one's own investment choice, modulated by one's expectations regarding receivers' reciprocating behavior, can serve as an anticipatory cue. Here though, the cue was neither externally created by character vignettes (Delgado et al., 2005) nor learned in a Pavlovian manner by pairing shapes to more or less trustworthy persons in a social context (Fouragnan et al., 2013), but was rather internally generated via participants' own beliefs about the world. This finding illustrates that eliciting participants' beliefs can be just as powerful in evoking anticipated reward signals as specifically pairing abstract cues with explicit (social) gains and losses.

We also showed that when participants were informed about a positive outcome – that their trust decision was reciprocated by a receiver in the Trust Game or that their marble was drawn in the lottery – the degree of their chosen investment level modulated the reward signal in the left ventral striatum and right dorsal striatum. These effects also highlight the potential of using participants' own beliefs in a real-life decision-making task when examining reward and subjective value. Some other effects are also worth exploring further.

A well-established finding is that losses are coded in the ventral striatum (Bartra et al., 2013). However, experiencing a loss in this study, that is, when the amount invested was not returned, did not activate similar brain regions as compared to when a trial was "won." Notably though, participants in this task did not actually lose money, but rather they lost the opportunity of winning more money by receiving a part of the tripled investment. When they lost, they still retained the non-invested number of tokens, thus perhaps minimizing the effect of the virtual loss. Moreover, it is found that positive effects are more likely to be coded in the striatum than negative effects (Bartra et al., 2013). These factors might explain this null finding with regard to experiencing losses.

Secondly, we also did not find neural differences in the anticipation of outcomes between ambiguous and risky contexts. Our experimental design differs from earlier explorations showing that various levels of uncertainty modulate expected reward in the ventral striatum (Fiorillo et al., 2003; Volz et al., 2003; Tobler et al., 2006). Namely, following standard practices in Economics (Wakker, 2010), here we clearly distinguish between risk and ambiguity, instead of varying uncertainty along a continuous distribution. Although decision-making under risk and ambiguity appear to be processed independently (Hsu et al., 2005; Huettel et al., 2006), anticipating their respective outcomes does not appear to differentially modulate neural processes. It might be that passively observing prior decisions does not sufficiently highlight the distinction between the types of uncertainty. Whereas revealing outcomes of social vs. non-social contexts emphasizes the role of the receiver and his intentions as compared to a non-intentional random mechanistic device, separating outcomes by types of uncertainty is likely not as compelling.

In a broader context, this is also a limitation of our experimental setup. We purposely separated the decision-making phase from the outcome phase, as we did not want participants to learn from the outcomes of their choices which could lead to potential belief adaptation across the experiment. While this means that our design can rule out learning effects, and that we can reliably use the beliefs elicited prior to decision-making, a downside of this procedure is that the re-evaluation of the choices that participants undertake is quite passive. Although we endeavored to enhance attention by including payment screens, we would ideally engage participants more intensively. Additionally it is worth noting that these results are based on a rather small sample size, and as such deserve follow-up exploration.

One interesting potential follow-up could be to design a dynamic experiment in which participants would be able to change future decision-making as a function of beliefs, which would presumably be updated as participants learned about the outcomes of prior choices, and beliefs could thus be elicited at various moments throughout the fMRI experiment. This would promote active engagement of both decision-making and outcome attention as well as the interaction between both phases as a function of belief updating, which moves experimental approaches closer to how trust and reciprocity are experienced in everyday life. This method could bridge two important directions in the field of Decision Neuroscience: namely, explorations of reward processing, which to date have rather neglected the role of participants' inherent beliefs, and the analyses of beliefs, which have focused on how beliefs emerge and are shaped (Vilares and Kording, 2011) but have examined less the interaction with expected value processing. Our study offers a first attempt as to how participants' own belief sets are employed in the reward processing in the context of trust and risky choice.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the local ethical committee with written informed consent from all subjects. The protocol was approved by the local ethical committee (CMO Arnhem-Nijmegen).

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study, and revised, read, and approved the submitted version of the manuscript. KF carried out the statistical analysis and wrote the first draft of the manuscript.

FUNDING

This research was partially supported by an European Research Council Grant to AS (ERC StG 313454).

ACKNOWLEDGMENTS

The material collected for this study was part of KF's dissertation (Fairley, 2016). Her thesis is the only medium it has appeared in, publishing this material is in line with the university policy and the thesis can be accessed online.

REFERENCES

Bartra, O., McGuire, J. T., and Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* 76, 412–427. doi: 10.1016/j.neuroimage.2013.02.063

Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., and Krueger, F. (2017). Neural signatures of trust in reciprocity: a coordinate-based meta-analysis. *Hum. Brain Mapp.* 38, 1233–1248. doi: 10.1002/hbm.23451

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

Chang, L. J., and Sanfey, A. G. (2011). Great expectations: neural computations underlying the use of social norms in decision-making. *Soc. Cogn. Affect. Neurosci.* 8, 277–284. doi: 10.1093/scan/nsr094

Davey, C. G., Allen, N. B., Harrison, B. J., Dwyer, D. B., and Yücel, M. (2009). Being liked activates primary reward and midline self-related brain regions. *Hum. Brain Mapp.* 31, 660–668. doi: 10.1002/hbm.20895

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

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., and Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* 84, 3072–3077. doi: 10.1152/jn.2000.84.6.3072

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Q. J. Econ.* 75, 643–669.

Fairley, K. (2016). *Behavioral and Neuroscientific Essays on Decision-Making Under Uncertainty*. Ph.D. thesis, Radboud University, Nijmegen.

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

Fiorillo, C. D., Tober, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898–1902. doi: 10.1126/science.1077349

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

Fox, C., and Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *Q. J. Econ.* 110, 585–603. doi: 10.2307/2946693

Fox, C. R., and Weber M. (2002). Ambiguity aversion, comparative ignorance, and decision context. *Organ. Behav. Hum. Process.* 88, 476–498. doi: 10.1006/obhd.2001.2990

Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E., and Penny, W. D. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Cambridge: Academic Press.

Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324, 646–648. doi: 10.1126/science.1168450

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J. Neurosci.* 28, 5623–5630. doi: 10.1523/JNEUROSCI.1309-08.2008

Houser D., Schunk D., and Winter J. (2010). Distinguishing trust from risk: an anatomy of the investment game. *J. Econ. Behav. Organ.* 74, 72–81. doi: 10.1016/j.jebo.2010.01.002

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–1683. doi: 10.1126/science.1115327

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., and Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron* 49, 765–775. doi: 10.1016/j.neuron.2006.01.024

Jones, R. M., Somerville, L. H., Li, J., Ruberri, E. J., Libby, V., Glover, G., et al. (2011)Behavioral and neural properties of social reinforcement learning. *J. Neurosci.* 31, 13039–13045. doi: 10.1523/JNEUROSCI.2972-11.2011

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. *Perception* 36, 1–16.

Knutson, B., Adams, C. M., Fong, G. W., and Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* 21, RC159–RC163.

Knutson, B., and Greer, S. M. (2008). Anticipatory affect: neural correlates and consequences for choice. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3771–3786. doi: 10.1098/rstb.2008.0155

Knutson, B., Westdorp, A., Kaiser, E., and Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* 12, 20–27. doi: 10.1006/nimg.2000.0593

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., and Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *J. Neurosci.* 32, 16832–16844. doi: 10.1523/JNEUROSCI.3016-12.2012

Lin, A., Adolphs, R., and Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Soc. Cogn. Affect. Neurosci.* 7, 274–281. doi: 10.1093/scan/nsr006

McCabe, K. A., Rigdon, M. L., and Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *J. Econ. Behav. Organ.* 52, 267–275. doi: 10.1016/s0167-2681(03)00003-9

O'Doherty, J. P., Deichmann, R., Critchley, H. D., and Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron* 33, 815–826. doi: 10.1016/s0896-6273(02)00603-7

Poser, B. A., Versluis, M. J., Hoogduin, J. M., and Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: parallel-acquired inhomogeneity-desensitizied fMRI. *Magn. Reson. Med.* 55, 1227–1235. doi: 10.1002/mrm.20900

Powers, K. E., Somerville, L. H., Kelley, W. M., and Heatherton, T. F. (2013). Rejection sensitivity polarizes striatal-medial prefrontal activity when anticipating social feedback. *J. Cogn. Neurosci.* 25, 1887–1895. doi: 10.1162/jocn_a_00446

Rilling, J. K., Sanfey, A. G., Aronson J. A., Nystrom L. E., and Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* 15, 2539–2543.

Savage, L. (1954). *The Foundations of Statistics*. New York, NY: John Wiley.

Schlag, K. H., Tremewan, J., and Van der Weele, J. J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Exp. Econ.* 18, 457–490. doi: 10.1007/s10683-014-9416-x

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27. doi: 10.1152/jn.1998.80.1.1

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behav. Brain Funct.* 6, 24–33. doi: 10.1186/1744-9081-6-24

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

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439, 466–469. doi: 10.1038/nature04271

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2006). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems, *J. Neurophysiol.* 97, 1621–1632. doi: 10.1152/jn.00745.2006

Trautmann, S. T., and van de Kuilen, G. (2014). "Ambiguity attitudes," in *Blackwell Handbook of Judgment and Decision Making*, eds G. Keren and G. Wu (New York, NY: John Wiley & Sons, Ltd).

Trautmann, S. T., Vieider, F. M., and Wakker, P. P. (2008). Causes of ambiguity aversion: known versus unknown preferences. *J. Risk Uncertain.* 36, 225–243. doi: 10.1007/s11166-008-9038-9

Vilares, I., and Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior and the brain. *Ann. N. Y. Acad. Sci.* 1224, 22–39. doi: 10.1111/j.1749-6632.2011.05965.x

Volz, K. G., Schubotz, R. I., and von Cramon, D. Y. (2003). Predicting events of varying probability: uncertainty investigated by fMRI. *Neuroimage* 19, 271–280. doi: 10.1016/s1053-8119(03)00122-8

Wakker, P. P. (2010). *Prospect Theory for Risk and Ambiguity*. Cambridge: Cambridge University Press.

Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., and Wager, T., (2011). NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670 doi: 10.1038/nmeth.1635

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

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 09 July 2019
doi: 10.3389/fnhum.2019.00230

Association of Polymorphism of Arginine-Vasopressin Receptor 1A (*AVPR1a*) Gene With Trust and Reciprocity

Kuniyuki Nishina[1], Haruto Takagishi[2], Hidehiko Takahashi[3], Masamichi Sakagami[2] and Miho Inoue-Murayama[4]*

[1]Graduate School of Brain Sciences, Tamagawa University, Tokyo, Japan, [2]Brain Science Institute, Tamagawa University, Tokyo, Japan, [3]Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan, [4]Wildlife Research Center, Kyoto University, Kyoto, Japan

Oxytocin (OXT) is known to play an important role in trust, whereas the involvement of other peptide hormones has not been evaluated. In this study, we focused on microsatellite polymorphisms in the intron of the arginine-vasopressin receptor 1a (*AVPR1a*) gene and examined whether the association between the repeat lengths in the intron of *AVPR1a* is associated with trust and reciprocity in humans. Four-hundred and thirty-three participants played the trust game, answered the attitudinal trust question, and their buccal cells were collected. Results showed that men with a short form of *AVPR1a* tend to send more money to the opponent, even if there is a possibility of being betrayed by the opponent. Additionally, people with a short form of *AVPR1a* tended to return money to the opponent who trusts them. However, attitudinal trust was not associated with *AVPR1a*. These results indicate that arginine-vasopressin receptor 1a plays an important role in trust and reciprocal behaviors.

Keywords: trust game, trust, reciprocity, economic game, AVPR1A gene, gene

OPEN ACCESS

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

Reviewed by:
*Raffaele Palmirotta,
University of Bari Aldo Moro, Italy
Masahiro Matsunaga,
Aichi Medical University, Japan*

****Correspondence:***
*Haruto Takagishi
haruharry@gmail.com*

***Received:** 27 February 2019*
***Accepted:** 21 June 2019*
***Published:** 09 July 2019*

Citation:
Nishina K, Takagishi H, Takahashi H, Sakagami M and Inoue-Murayama M (2019) Association of Polymorphism of Arginine-Vasopressin Receptor 1A (AVPR1a) Gene With Trust and Reciprocity. Front. Hum. Neurosci. 13:230. doi: 10.3389/fnhum.2019.00230

INTRODUCTION

Trust is an indicator of social capital reflecting the efficiency of society, and numerous studies in the field of social science have examined human trust (Putnam et al., 1994; Fukuyama, 1995; Yamagishi, 2011). In recent years, attention has focused on the biological foundation of trust, which revealed that the peptide hormone oxytocin (OXT) synthesized in the hypothalamus regulates trust (Kosfeld et al., 2005). OXT functions in various parts of the body such as the uterus and mammary glands after transport through blood vessels from the posterior pituitary gland. OXT is axon-projected in various regions of the central nervous system, such as the striatum, amygdala, hippocampus, and others (Meyer-Lindenberg et al., 2011). Previous studies showed that OXT attenuates the stress response and enhances the reward system, as well as regulates social cognition and behavior (Domes et al., 2007; Feldman, 2017). In trust, OXT attenuates anxiety related to social risk and promotes trust by suppressing the activity of the amygdala, which is the center of emotional processing (Baumgartner et al., 2008). Additionally, because twin studies have shown that trust is inherited (Cesarini et al., 2008; Reimann et al., 2017), genetic approaches have been used to identify candidate genes of trust. Some studies showed that a polymorphism in the oxytocin receptor gene (*OXTR* rs53576) in human chromosome 3p.25.3 is associated with trust behavior and trust attitude (Krueger et al., 2012; Nishina et al., 2015), and that the amygdala volume mediates the association between *OXTR* rs53576 and trust attitude (Nishina et al., 2018). OXT is known to play an important role in trust, whereas the involvement of other peptide hormones has not been evaluated.

Arginine-vasopressin (AVP) is a peptide hormone synthesized in the hypothalamus and exerts its effects in the central nervous system (Meyer-Lindenberg et al., 2011). The AVP receptors, V1a and V1b, are distributed in the prefrontal cortex, hippocampus, amygdala, and various other regions of the brain and regulate anxiety and pair bonding behavior (Young and Wang, 2004). According to nonhuman primate studies, V1a knockout mice show low levels of anxiety (Bielsky et al., 2004) and that V1 receptor antagonist reduces anxiety-related behavior in rats (Liebsch et al., 1996). In human studies, administration of AVP increased the stress response (Shalev et al., 2011) and enhanced activation of the amygdala response to negative emotional stimuli (Brunnlieb et al., 2013). Because previous studies of OXT revealed that OXT inhibits social stress (Heinrichs et al., 2003) and attenuates activation of the amygdala response to a fearful face (Kirsch et al., 2005), OXT and AVP have opposite effects on the brain. If so, AVP inhibits trust in contrast to OXT.

The arginine-vasopressin receptor 1a (*AVPR1a*) gene is on human chromosome 12q14.2 and has two exons (Thibonnier et al., 1996). *AVPR1A* has three microsatellite polymorphisms in the promoter region, repeating two bases of $(GT)_{25}$, a complex repeat of $(CT)_4$-TT-$(CT)_8$-$(GT)_{24}$ [RS3], and a repetition of the four-nucleotide sequence GATA [RS1]. Previous studies showed that the repeat length in RS3 is associated with autism (Yirmiya et al., 2006), pair bonding behavior (Walum et al., 2008), maternal behavior (Avinun et al., 2012), and altruistic behavior in the economic game (Knafo et al., 2008; Avinun et al., 2011; Wang et al., 2016). Additionally, Meyer-Lindenberg et al. (2009) found that the repeat length in RS1 and RS3 is associated with activation of the amygdala response to emotional facial expression. Nonhuman primate studies showed that the repeat length in RS3 is associated with personality (Hopkins et al., 2012; Staes et al., 2016) and social cognition (Hopkins et al., 2014; Mahovetz et al., 2016). These results indicate that *AVPR1a* plays a role in social cognition and social behavior not only in humans but also in various other species.

Although many studies have examined microsatellite polymorphisms (RS1 and RS3) in the promoter region of *AVPR1a*, a recent study detected an association between the repeat length in the intron and personality in common marmoset. Inoue-Murayama et al. (2018) examined the association between microsatellite polymorphisms in the intron of *AVPR1a* and personality scores rated by humans for common marmoset and revealed that the short form of *AVPR1a* is associated with a high level of sociability. These results indicate that not only the repeat length in the promoter region but also that in the intron is related to sociality in nonhuman primates. However, it remains unclear whether microsatellite polymorphisms in the intron of *AVPR1a* are associated with sociality in humans. Considering the common role of *AVPR1a* in sociality with other animals, it is important to understand the evolution of human sociality.

In this study, we focused on microsatellite polymorphisms in the intron of *AVPR1a* and examined whether the association between the repeat lengths in the intron of *AVPR1a* is associated with trust and reciprocity in humans. To clarify the biological basis of trust, it is necessary to evaluate whether OXT and AVP are related to trust behavior.

MATERIALS AND METHODS

Participants

Six-hundred non-student residents living in Tokyo suburbs were selected from a list of 1,670 applicants who responded to a brochure distributed to approximately 180,000 households. These 600 individuals consisted of 75 men and 75 women in each 10-year age group from 20 to 59 years in the first wave (May 17, 2012). The study was conducted in ten waves for 7 years (from 2012 to 2018) and the participants repeatedly participated in the experiment. Findings concerning some of the data collected during the ten phases have been previously reported (Yamagishi et al., 2014, 2015, 2016a,b, 2017a,b; Nishina et al., 2015, 2018; Matsumoto et al., 2016). An overview of the whole research project is provided in **Figure 1**.

Trust Game

The trust game was conducted in the fifth wave (December 16, 2013 to February 23, 2014). The procedures of the trust game are similar to those reported previously (Nishina et al., 2015). Participants played the trust game in a situation where anonymity was fully guaranteed. The trust game was played between pairs of participants randomly matched from among the 6–12 participants who attended the same experimental session. One member of the pair played the role of truster and the other the role of trustee. The truster was provided with JPY 1,000 by the experimenter and decided how much of these funds to transfer to the trustee in increments of JPY 100. The transferred money was then tripled and provided to the trustee. The trustee then decided how much of the tripled money to transfer back to the truster. The endowment money of JPY 1,000 was provided only to the truster, and not to the trustee. All participants were told that they would play the game twice, each time with a different partner and that their role would change. All participants played the truster role in the first game and trustee role in the second game. Trustees' responses in the second game were measured using the strategy method. We averaged the amount of money the trustee had returned to the opponent when receiving more than 60% of the endowment and defined the behavior as reciprocity.

Attitudinal Trust

The question of attitudinal trust was used in the first wave and in the seventh wave. Participants answered the

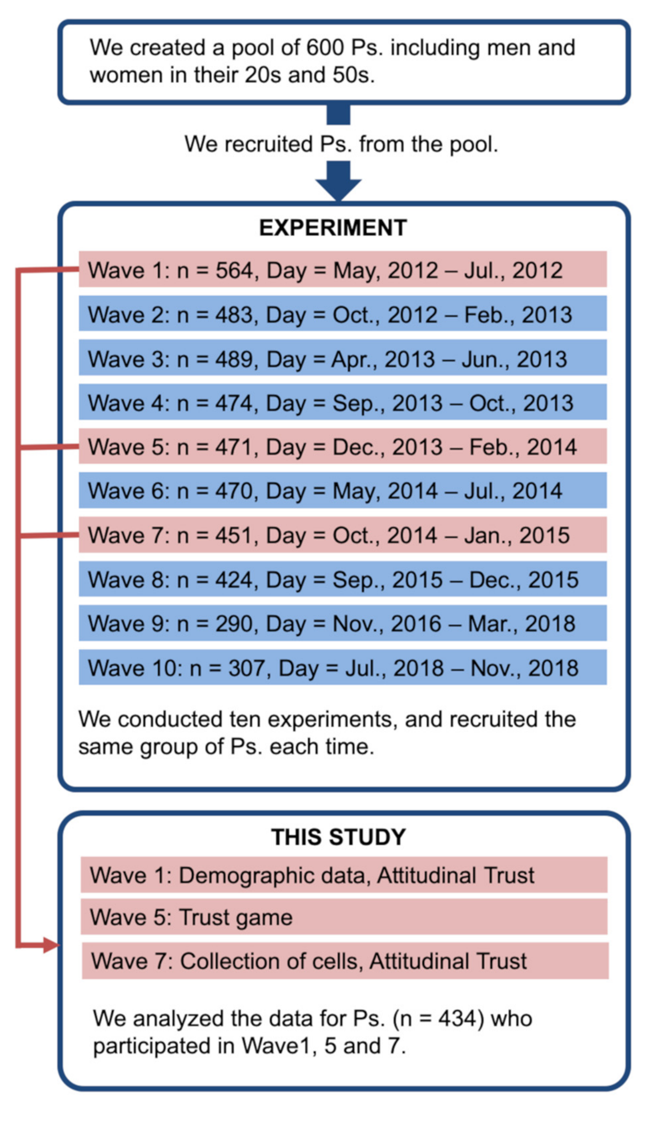

We created a pool of 600 Ps. including men and women in their 20s and 50s.
We recruited Ps. from the pool.
EXPERIMENT
Wave 1: n = 564, Day = May, 2012 – Jul., 2012
Wave 2: n = 483, Day = Oct., 2012 – Feb., 2013
Wave 3: n = 489, Day = Apr., 2013 – Jun., 2013
Wave 4: n = 474, Day = Sep., 2013 – Oct., 2013
Wave 5: n = 471, Day = Dec., 2013 – Feb., 2014
Wave 6: n = 470, Day = May, 2014 – Jul., 2014
Wave 7: n = 451, Day = Oct., 2014 – Jan., 2015
Wave 8: n = 424, Day = Sep., 2015 – Dec., 2015
Wave 9: n = 290, Day = Nov., 2016 – Mar., 2018
Wave 10: n = 307, Day = Jul., 2018 – Nov., 2018
We conducted ten experiments, and recruited the same group of Ps. each time.
THIS STUDY
Wave 1: Demographic data, Attitudinal Trust
Wave 5: Trust game
Wave 7: Collection of cells, Attitudinal Trust
We analyzed the data for Ps. (n = 434) who participated in Wave1, 5 and 7.

FIGURE 1 | Overview of the whole research project. Ps, participants.

following question, "Do you think most people would try to take advantage of you if they got a chance or would they not?" The form of answer was a binary; 0 indicated low trust and 1 indicated high trust. This question was used in two large scale social surveys: the General Social Survey and World Value Survey. We averaged the two scores of this question. The score was associated with the polymorphism of the oxytocin receptor gene (*OXTR* rs53576) reported in our previous studies (Nishina et al., 2015; Nishina et al., 2018).

Genotyping

Participants' buccal cells were collected in the seventh wave (October 25th, 2014 to January 25, 2015) and preserved in 90% ethanol until DNA extraction. DNA was extracted using the DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. DNA was amplified by PCR. To amplify the microsatellite polymorphism in the intron [$(GT)_{14}(GA)_{13}(A)_6$] (Accession No. DQ177277; **Figure 2**), we used primers 5′-ATGTGGTCTGTCTGGGATGC-3′ (forward) and 5′-GGGTGCGACTGTAGTACACA-3′ (reverse; Inoue-Murayama et al., 2018). PCR amplification conditions were as follows: 94°C for 1 min, and then 94°C for 30 s, 60°C for 30 s, 74°C for 1 min) × 35 cycles, and final extension at 74°C for 10 min. The PCR products were analyzed with an ABI 3130*xl* DNA Sequencer and GeneMapper Software (Applied Biosystems, Foster City, CA, USA).

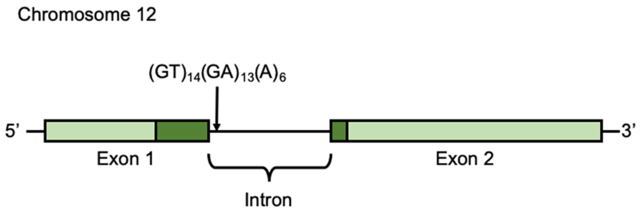


FIGURE 2 | Arginine-vasopressin receptor 1a (*AVPR1a*) gene *AVPR1a* is located in chromosome 12q14.2 in humans and it has two exons. Light green shows non-coding region and Dark green shows coding region.

Analysis

A total of 470 participants (male = 228, female = 242) played the trust game and we genotyped 449 participants (male = 221, female = 228). We analyzed 434 participants (male = 213, female = 221) for whom both behavioral and genetic data were available. Since this study was a part of the large-scale research project, we could not design a sample size suitable for this study. Instead, we report the power of analysis used in this study. The power was calculated by G*Power 3.1 software (Faul et al., 2009).

RESULTS

Genotype Distribution

The distribution of the number of alleles is shown in **Table 1**. We defined an allele of repeat length 217 and greater than 217 as "long" (L) and that less than 217 as "short" (S). The genotype distribution of the 434 participants was 17.1% SS (N = 74), 50.5% SL (N = 219), and 32.5% LL (N = 141). This distribution did not significantly differ from Hardy-Weinberg equilibrium ($\chi^2_{(1)}$ = 0.497, p = 0.481). The demographic data for 434 participants are shown in **Supplementary Tables S1–S5**. We did not find significant differences in the proportion of sex ($\chi^2_{(2)}$ = 2.07, p = 0.356), generation ($\chi^2_{(6)}$ = 5.04,

TABLE 1 | Frequency for *AVPR1a* genotype.

Allele	*n*	% Carriers
211	5	0.6
213	211	24.3
215	151	17.4
217	490	56.5
219	10	1.2
221	1	0.1

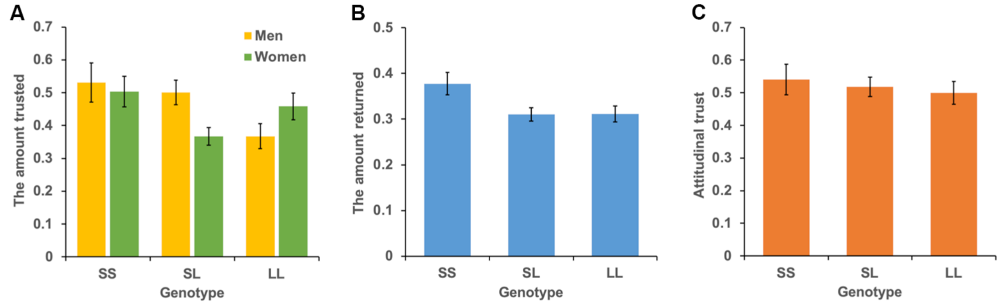


FIGURE 3 | Mean levels of behavioral trust, reciprocity, and attitudinal trust for each genotype. The vertical bar represents the amount sent to the second player **(A)**, the amount returned to the first player **(B)**, and the level of attitudinal trust **(C)**. Error bars show standard error.

$p = 0.538$), education level ($\chi^2_{(2)} = 4.33$, $p = 0.115$), annual income ($\chi^2_{(12)} = 7.98$, $p = 0.787$), and subjective social class ($\chi^2_{(8)} = 8.95$, $p = 0.346$), genotype.

Trust

The mean levels of trust for the three genotypes are shown in **Figure 3A**. We conducted a multiple regression analysis of trust behavior. Age, sex (men = 1), dummy variable of SL genotype (= 1), and dummy variable of SS genotype (= 1) were used as independent variables. By setting the LL genotype as the baseline, this model can be used to examine the differences in trust behavior between LL vs. SL and LL vs. SS. The dependent variable is the ratio of money sent by the first player to the second player. The results showed that the SS genotype ($b = 0.004$, $SE = 0.002$, $p = 0.015$, $\beta = 0.116$) and age positively affected trust behavior ($b = 0.101$, $SE = 0.048$, $p = 0.034$, $\beta = 0.114$; model 1 in **Table 2**). However, the SL genotype ($b = 0.019$, $SE = 0.036$, $p = 0.601$, $\beta = 0.028$) and sex ($b = 0.041$, $SE = 0.032$, $p = 0.203$, $\beta = 0.061$) did not significantly affect trust behavior. Additionally, we examined the interaction effect of genotype and sex in model 2. The interaction effect of SL genotype and sex ($b = 0.218$, $SE = 0.071$, $p = 0.002$, $\beta = 0.275$) and the effect of age ($b = 0.004$, $SE = 0.002$, $p = 0.020$, $\beta = 0.111$) were significant. The effect of sex ($b = -0.09$, $SE = 0.055$, $p = 0.101$, $\beta = -0.136$), the SL genotype ($b = -0.09$, $SE = 0.050$, $p = 0.074$, $\beta = -0.134$) and SS genotype ($b = 0.034$, $SE = 0.069$, $p = 0.616$, $\beta = 0.039$), the interaction effect of SS genotype and sex ($b = 0.128$, $SE = 0.094$, $p = 0.177$, $\beta = 0.109$) were not significant. Since the interaction effect of the SL genotype and sex was significant, we analyze the effect of genotype for each sex. In men, the SL genotype ($b = 0.125$, $SE = 0.054$, $p = 0.022$, $\beta = 0.173$), the SS genotype ($b = 0.161$, $SE = 0.070$, $p = 0.022$, $\beta = 0.173$), and age ($b = 0.006$, $SE = 0.002$, $p = 0.014$, $\beta = 0.166$) positively affected trust behavior. In women, age ($b = 0.001$, $SE = 0.001$, $p = 0.440$, $\beta = 0.052$), the SL genotype ($b = -0.090$, $SE = 0.046$, $p = 0.051$, $\beta = -0.148$), and the SS genotype ($b = 0.40$, $SE = 0.063$, $p = 0.523$, $\beta = 0.048$) did not have an effect. The power of analysis calculated by the ΔR_2 of tested predictors was 0.48 in model 1 and 0.87 in model 2.

Reciprocity

The mean levels of reciprocity for the three genotypes are shown in **Figure 3B**. We conducted the same analyses used for trust behavior analysis. The dependent variable was the ratio of money returned by the second player to the first player. The results showed that the SS genotype ($b = 0.058$, $SE = 0.030$, $p = 0.048$, $\beta = 0.104$) and age ($b = 0.005$, $SE = 0.0009$, $p < 0.001$, $\beta = 0.241$) positively affected reciprocity (model 1 in **Table 3**). The SL genotype ($b = -0.005$, $SE = 0.022$, $p = 0.838$, $\beta = -0.011$) and sex ($b = -0.018$, $SE = 0.020$, $p = 0.369$, $\beta = -0.042$) did not significantly affect reciprocity. Additionally, we found no significant interaction effect of genotype and sex on reciprocity (model 2 in **Table 3**). The power of analysis calculated by the ΔR_2 of tested predictors was 0.52 in model 1 and 0.50 in model 2.

Attitudinal Trust

The mean levels of attitudinal trust for the three genotypes are shown in **Figure 3C**. We conducted the same analytic model. The dependent variable was the level of attitudinal trust. We did not find an effect of SS genotype ($b = 0.027$, $SE = 0.059$, $p = 0.647$, $\beta = 0.024$) and SL genotype ($b = 0.016$, $SE = 0.045$,

TABLE 2 | Results of multiple regression analysis of trust.

	Model 1				Model 2			
Variables	***b***	***SE***	***p***	***β***	***b***	***SE***	***p***	***β***
Intercept	0.239	0.069	0.001	0.000	0.315	0.073	<0.0001	0.000
Age	0.004	0.002	0.015	0.116	0.004	0.002	0.020	0.111
Sex	0.041	0.032	0.203	0.061	−0.091	0.055	0.101	−0.136
SL	0.019	0.036	0.601	0.028	−0.090	0.050	0.074	−0.134
SS	0.101	0.048	0.034	0.114	0.034	0.069	0.616	0.039
SL × Sex		-			0.218	0.071	0.002	0.275
SS × Sex		-			0.128	0.094	0.177	0.109

TABLE 3 | Results of multiple regression analysis of reciprocity.

	Model 1				Model 2			
Variables	***b***	***SE***	***p***	***β***	***b***	***SE***	***p***	***β***
Intercept	0.123	0.043	0.004	0.000	0.129	0.046	0.005	0.000
Age	0.005	0.001	<0.0001	0.241	0.005	0.001	<0.0001	0.237
Sex	−0.018	0.020	0.369	−0.042	−0.023	0.035	0.514	−0.053
SL	−0.005	0.022	0.838	−0.011	−0.016	0.031	0.621	−0.037
SS	0.058	0.030	0.048	0.104	0.080	0.043	0.064	0.141
SL × Sex			-		0.023	0.044	0.599	0.046
SS × Sex			-		−0.040	0.059	0.497	−0.054

TABLE 4 | Results of multiple regression analysis of attitudinal trust.

	Model 1				Model 2			
Variables	***b***	***SE***	***p***	***β***	***b***	***SE***	***p***	***β***
Intercept	0.143	0.086	0.098	0.000	0.176	0.092	0.057	0.000
Age	0.009	0.002	<0.0001	0.213	0.009	0.002	<0.0001	0.213
Sex	0.019	0.040	0.626	0.023	−0.044	0.070	0.533	−0.052
SL	0.016	0.045	0.725	0.019	−0.028	0.063	0.658	−0.033
SS	0.027	0.059	0.647	0.024	−0.033	0.086	0.704	−0.029
SL × Sex			-		0.086	0.090	0.336	0.086
SS × Sex			-		0.114	0.119	0.338	0.078

$p = 0.725$, $\beta = 0.019$) on attitudinal trust (model 1 in **Table 4**). Additionally, we did not find an interaction effect of SS genotype and sex ($b = 0.114$, $SE = 0.119$, $p = 0.338$, $\beta = 0.078$) and SL genotype and sex ($b = 0.086$, $SE = 0.090$, $p = 0.336$, $\beta = 0.086$; model 2 in **Table 4**).

DISCUSSION

Trust behavior is associated with microsatellite polymorphisms in the intron of *AVPR1a*. Men with a short form of *AVPR1a* tend to send more money to the opponent, even if there is a possibility of being betrayed by the opponent. In contrast, men with a long form of *AVPR1a* tend to keep their money. This is the first study to reveal an association between the microsatellite polymorphism in the intron of *AVPR1a* and trust behavior in humans. Our results indicate that the microsatellite polymorphism in the intron of *AVPR1a* reflects the function of arginine-vasopressin receptor 1a as well as *AVPR1a* RS1 and RS3. Previous studies showed that AVP neurons in the hypothalamus are axon-projected to the amygdala, which is the center of anxiety and fear processing (Huber et al., 2005). Additionally, AVP promotes anxiety and fearful response to emotional stimuli (Shalev et al., 2011; Brunnlieb et al., 2013). Such anxiety regulating the action of AVP can explain the results for trust behavior observed in this study. People with a long form of *AVPR1a* experience a strong effect from AVP and may be fearful of being betrayed by others. Inoue-Murayama et al. (2018) examined the association between the repeat length of the intron of *AVPR1a* and personality in the common marmoset and found that individuals with a long form of *AVPR1a* had high levels of neuroticism. These findings support the hypothesis that anxiety plays a role in trust in those with a long form of *AVPR1a*. In our previous study (Nishina et al., 2015), we found an association between the polymorphism of the oxytocin receptor gene and behavioral trust and attitudinal trust in men. A sex difference of the association of gene polymorphism and trust was also observed in the current study. These results indicate that OXT, as well as, AVP play important roles in trust in men.

Reciprocity was also associated with the repeat length of *AVPR1a*. People with a short form of *AVPR1a* tended to return money to the opponent who trusts them. The association between *AVPR1a* and reciprocity cannot be explained by the hypothesis that AVP enhances anxiety related to exploitation by others, as there is no risk of being betrayed by others in this case. AVP may not regulate the anxiety related to exploitation by others, but rather anxiety regarding the loss of money. Thus, as people with a long form of *AVPR1a* show high levels of anxiety related to the loss of money, they keep the money in either role, distrust the first player, and do not reciprocate in the second player. Huber et al. (2005) found that OXT neurons and AVP neurons differ in the location of the projection to the amygdala. This suggests that OXT and AVP regulate different types of anxiety. To evaluate this possibility, further studies are needed to examine whether *AVPR1a* is related to behavior in a trust game with a computer partner reflecting non-social risk avoidance.

AVP is related to not only anxiety but also reward processing (Meyer-Lindenberg et al., 2011). Vasopressin neurons from the hypothalamus are projected to the ventral pallidum, which forms the dopamine pathway. Avinun et al. (2011) found that individuals with the 334 allele of RS3 in *AVPR1a* showed low levels of generosity and that the 334 allele carriers maximized their self-interests through the reward system enhancement effect of AVP. However, other studies showed that AVP does not affect the motivation of maximizing self-interest, but motivation of social reward such as mutual cooperation (Rilling et al., 2012, 2014). Whether AVP affects social rewards or non-social rewards require further examination.

There were two differences between the results observed in this study and those observed in the *OXTR* study. First, we did not find an association of the polymorphism of *AVPR1a* and attitudinal trust. An important difference between behavioral trust and attitudinal trust is whether there is financial damage if the trust is betrayed. One possibility is that OXT affects attitudes like general trust by acting continuously, while AVP influences actual decision making by acting acutely depending on the situation. Another possibility is that vasopressin plays an important role in money-related decision making. Further studies are needed to examine the differences in the effects of oxytocin and vasopressin on trust. Second, while *OXTR* was related to trust but not to reciprocity (Nishina et al., 2015), *AVPR1a* was related to both trust and reciprocity. This result shows that OXT acts on trust-specific factors and that AVP acts on factors related to overall pro-social behavior. As described above, different types of anxiety may be regulated by OXT and AVP.

We found similar results as a previous study that examined the association between microsatellite polymorphisms in the intron of *AVPR1a* and sociality (Inoue-Murayama et al., 2018). In common marmoset, a short form of *AVPR1a* was related to a high level of sociality. In humans, a short form of *AVPR1a* was also related to a high level of trust and reciprocity. The common features between types of *AVPR1a* and sociality shows that the role of AVP system in sociality is an evolutionarily old issue.

We did not examine which brain function and structure mediates the association between *AVPR1a* and trust behavior. As many imaging genetic approaches are available for evaluating humans (Saito et al., 2014; Wang et al., 2016; Nishina et al., 2018), further studies are needed to determine the neural mechanism relationship between *AVPR1a* and trust behavior.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors to any qualified researcher.

ETHICS STATEMENT

All experimental protocols were approved by the Ethics Committee of Tamagawa University, where the study was conducted, and ethics committee of Kyoto University Graduate School and Faculty of Medicine, where genotyping analysis was conducted. Each participant signed an informed consent form before the experiment.

AUTHOR CONTRIBUTIONS

KN, HaT, HiT, MS, and MI-M designed research. KN and HaT performed research. KN, HaT, and MI-M analyzed data and wrote the article.

FUNDING

This work was supported by JSPS KAKENHI Grant Numbers JP15H05730, JP16H06324, JP25118004, JP25290082, JP19H04915, Cooperative Research Program of Wildlife Research Center, Kyoto University and AMED Grant Number JP18dm0307006.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2019.00230/full#supplementary-material

REFERENCES

Avinun, R., Ebstein, R. P., and Knafo, A. (2012). Human maternal behaviour is associated with arginine vasopressin receptor 1A gene. *Biol. Lett.* 8, 894–896. doi: 10.1098/rsbl.2012.0492

Avinun, R., Israel, S., Shalev, I., Gritsenko, I., Bornstein, G., Ebstein, R. P., et al. (2011). AVPR1A variant associated with preschoolers' lower altruistic behavior. *PLoS One* 6:e25274. doi: 10.1371/journal.pone.0025274

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., and Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58, 639–650. doi: 10.1016/j.neuron.2008.04.009

Bielsky, I. F., Hu, S. B., Szegda, K. L., Westphal, H., and Young, L. J. (2004). Profound impairment in social recognition and reduction in anxiety-like behavior in vasopressin V1a receptor knockout mice. *Neuropsychopharmacology* 29, 483–493. doi: 10.1038/sj.npp.1300360

Brunnlieb, C., Münte, T. F., Tempelmann, C., and Heldmann, M. (2013). Vasopressin modulates neural responses related to emotional stimuli in the right amygdala. *Brain Res.* 1499, 29–42. doi: 10.1016/j.brainres.2013.01.009

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., and Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proc. Natl. Acad. Sci. U S A* 105, 3721–3726. doi: 10.1073/pnas.0710069105

Domes, G., Heinrichs, M., Michel, A., Berger, C., and Herpertz, S. C. (2007). Oxytocin improves "mind-reading" in humans. *Biol. Psychiatry* 61, 731–733. doi: 10.1016/j.biopsych.2006.07.015

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Feldman, R. (2017). The neurobiology of human attachments. *Trends Cogn. Sci.* 21, 80–99. doi: 10.1016/j.tics.2016.11.007

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. New York, NY: Free Press Paperbacks.

Heinrichs, M., Baumgartner, T., Kirschbaum, C., and Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biol. Psychiatry* 54, 1389–1398. doi: 10.1016/s0006-3223(03)00465-7

Hopkins, W. D., Donaldson, Z. R., and Young, L. J. (2012). A polymorphic indel containing the RS3 microsatellite in the 5′ flanking region of the vasopressin V1a receptor gene is associated with chimpanzee (Pan troglodytes) personality. *Genes Brain Behav.* 11, 552–558. doi: 10.1111/j.1601-183x.2012.00799.x

Hopkins, W. D., Keebaugh, A. C., Reamer, L. A., Schaeffer, J., Schapiro, S. J., and Young, L. J. (2014). Genetic influences on receptive joint attention in chimpanzees (Pan troglodytes). *Sci. Rep.* 4:3774. doi: 10.1038/srep03774

Huber, D., Veinante, P., and Stoop, R. M. (2005). Vasopressin and oxytocin excite distinct neuronal populations in the central amygdala. *Science* 308, 245–248. doi: 10.1126/science.1105636

Inoue-Murayama, M., Yokoyama, C., Yamanashi, Y., and Weiss, A. (2018). Common marmoset (Callithrix jacchus) personality subjective well-being hair cortisol level and AVPR1a OPRM1 and DAT genotypes. *Sci. Rep.* 8:10255. doi: 10.1038/s41598-018-28112-7

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., et al. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *J. Neurosci.* 25, 11489–11493. doi: 10.1523/JNEUROSCI.3984-05.2005

Knafo, A., Israel, S., Darvasi, A., Bachner-Melman, R., Uzefovsky, F., Cohen, L., et al. (2008). Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes Brain Behav.* 7, 266–275. doi: 10.1111/j.1601-183x.2007.00341.x

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

Krueger, F., Parasuraman, R., Iyengar, V., Thornburg, M., Weel, J., Lin, M., et al. (2012). Oxytocin receptor genetic variation promotes human trust behavior. *Front. Hum. Neurosci.* 6:4. doi: 10.3389/fnhum.2012.00004

Liebsch, G., Wotjak, C. T., Landgraf, R., and Engelmann, M. (1996). Septal vasopressin modulates anxiety-related behaviour in rats. *Neurosci. Lett.* 217, 101–104. doi: 10.1016/s0304-3940(96)13069-x

Mahovetz, L. M., Young, L. J., and Hopkins, W. D. (2016). The influence of AVPR1A genotype on individual differences in behaviors during a mirror self-recognition task in chimpanzees (Pan troglodytes). *Genes Brain Behav.* 15, 445–452. doi: 10.1111/gbb.12291

Matsumoto, Y., Yamagishi, T., Li, Y., and Kiyonari, T. (2016). Prosocial behavior increases with age across five economic games. *PLoS One* 11:e0158671. doi: 10.1371/journal.pone.0158671

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

Meyer-Lindenberg, A., Kolachana, B., Gold, B., Olsh, A., Nicodemus, K. K., Mattay, V., et al. (2009). Genetic variants in AVPR1A linked to autism predict amygdala activation and personality traits in healthy humans. *Mol. Psychiatry* 14, 968–975. doi: 10.1038/mp.2008.54

Nishina, K., Takagishi, H., Fermin, A. S. R., Inoue-Murayama, M., Takahashi, H., Sakagami, M., et al. (2018). Association of the oxytocin receptor gene with attitudinal trust: role of amygdala volume. *Soc. Cogn. Affect. Neurosci.* 13, 1091–1097. doi: 10.1093/scan/nsy075

Nishina, K., Takagishi, H., Inoue-Murayama, M., Takahashi, H., and Yamagishi, T. (2015). Polymorphism of the oxytocin receptor gene modulates behavioral and attitudinal trust among men but not women. *PLoS One* 10:e0137089. doi: 10.1371/journal.pone.0137089

Putnam, R. D., Leonardi, R., and Nanetti, R. Y. (1994). *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton, NJ: Princeton University Press.

Reimann, M., Schilke, O., and Cook, K. S. (2017). Trust is heritable whereas distrust is not. *Proc. Natl. Acad. Sci. U S A* 114, 7007–7012. doi: 10.1073/pnas.1617132114

Rilling, J. K., DeMarco, A. C., Hackett, P. D., Chen, X., Gautam, P., Stair, S., et al. (2014). Sex differences in the neural and behavioral response to intranasal oxytocin and vasopressin during human social interaction. *Psychoneuroendocrinology* 39, 237–248. doi: 10.1016/j.psyneuen.2013.09.022

Rilling, J. K., DeMarco, A. C., Hackett, P. D., Thompson, R., Ditzen, B., Patel, R., et al. (2012). Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. *Psychoneuroendocrinology* 37, 447–461. doi: 10.1016/j.psyneuen.2011.07.013

Saito, Y., Suga, M., Tochigi, M., Abe, O., Yahata, N., Kawakubo, Y., et al. (2014). Neural correlate of autistic-like traits and a common allele in the oxytocin receptor gene. *Soc. Cogn. Affect. Neurosci.* 9, 1443–1450. doi: 10.1093/scan/nst136

Shalev, I., Israel, S., Uzefovsky, F., Gritsenko, I., Kaitz, M., and Ebstein, R. P. (2011). Vasopressin needs an audience: neuropeptide elicited stress responses are contingent upon perceived social evaluative threats. *Horm. Behav.* 60, 121–127. doi: 10.1016/j.yhbeh.2011.04.005

Staes, N., Weiss, A., Helsen, P., Korody, M., Eens, M., and Stevens, J. M. (2016). Bonobo personality traits are heritable and associated with vasopressin receptor gene 1a variation. *Sci. Rep.* 6:38193. doi: 10.1038/srep38193

Thibonnier, M., Graves, M. K., Wagner, M. S., Auzan, C., Clauser, E., and Willard, H. F. (1996). Structure sequence expression and chromosomal localization of the human V1a vasopressin receptor gene. *Genomics* 31, 327–334. doi: 10.1006/geno.1996.0055

Walum, H., Westberg, L., Henningsson, S., Neiderhiser, J. M., Reiss, D., Igl, W., et al. (2008). Genetic variation in the vasopressin receptor 1a gene (AVPR1A) associates with pair-bonding behavior in humans. *Proc. Natl. Acad. Sci. U S A* 105, 14153–14156. doi: 10.1073/pnas.0803081105

Wang, J., Qin, W., Liu, F., Liu, B., Zhou, Y., Jiang, T., et al. (2016). Sex-specific mediation effect of the right fusiform face area volume on the association between variants in repeat length of AVPR1A RS3 and altruistic behavior in healthy adults. *Hum. Brain Mapp.* 37, 2700–2709. doi: 10.1002/hbm.23203

Yamagishi, T. (2011). *Trust: The Evolutionary Game of Mind and Society.* New York, NY: Springer.

Yamagishi, T., Akutsu, S., Cho, K., Inoue, Y., Li, Y., and Matsumoto, Y. (2015). Two-component model of general trust: predicting behavioral trust from attitudinal trust. *Soc. Cogn.* 33, 436–458. doi: 10.1521/soco.2015.33.5.436

Yamagishi, T., Li, Y., Fermin, A. S., Kanai, R., Takagishi, H., Matsumoto, Y., et al. (2017a). Behavioural differences and neural substrates of altruistic and spiteful punishment. *Sci. Rep.* 7:14654. doi: 10.1038/s41598-017-15188-w

Yamagishi, T., Matsumoto, Y., Kiyonari, T., Takagishi, H., Li, Y., Kanai, R., et al. (2017b). Response time in economic games reflects different types of decision conflict for prosocial and proself individuals. *Proc. Natl. Acad. Sci. U S A* 114, 6394–6399. doi: 10.1073/pnas.1608877114

Yamagishi, T., Li, Y., Matsumoto, Y., and Kiyonari, T. (2016a). Moral bargain hunters purchase moral righteousness when it is cheap: within-individual effect of stake size in economic games. *Sci. Rep.* 6:27824. doi: 10.1038/srep27824

Yamagishi, T., Takagishi, H., Fermin, A. S. R., Kanai, R., Li, Y., and Matsumoto, Y. (2016b). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proc. Natl. Acad. Sci. U S A* 113, 5582–5587. doi: 10.1073/pnas.1523940113

Yamagishi, T., Li, Y., Takagishi, H., Matsumoto, Y., and Kiyonari, T. (2014). In search of homo economicus. *Psychol. Sci.* 25, 1699–1711. doi: 10.1177/0956797614538065

Yirmiya, N., Rosenberg, C., Levi, S., Salomon, S., Shulman, C., Nemanov, L., et al. (2006). Association between the arginine vasopressin 1a receptor (AVPR1a) gene and autism in a family-based study: mediation by socialization skills. *Mol. Psychiatry* 11, 488–494. doi: 10.1038/sj.mp.4001812

Young, L. J., and Wang, Z. (2004). The neurobiology of pair bonding. *Nat. Neurosci.* 7, 1048–1054. doi: 10.1038/nn1327

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

ORIGINAL RESEARCH
published: 17 July 2019
doi: 10.3389/fnins.2019.00741

Left Amygdala and Putamen Activation Modulate Emotion Driven Decisions in the Iterated Prisoner's Dilemma Game

Iveta Eimontaite[1], Igor Schindler[1], Matteo De Marco[2], Davide Duzzi[3], Annalena Venneri[2] and Vinod Goel[4,5]**

[1] Department of Psychology, University of Hull, Hull, United Kingdom, [2] Department of Neuroscience, The University of Sheffield, Sheffield, United Kingdom, [3] IRCCS San Camillo Hospital Foundation, Venice, Italy, [4] Department of Psychology, York University, Toronto, ON, Canada, [5] Capital Normal University, Beijing, China

Although economic decision-making is commonly characterized as a purely rational phenomenon, it is clear that real-world decision-making is influenced by emotions. Yet, relatively little is known about the neural correlates of this process. To explore this issue, 20 participants underwent fMRI scanning while engaged in the Prisoner's Dilemma game under partner-directed sympathy, anger and neutral emotion conditions. Participants were most and least likely to cooperate after sympathy and anger induction, respectively, with the neutral condition eliciting intermediate cooperation rates. Moreover, the sympathy condition elicited quicker responses for cooperation than defection choices, whereas this pattern was reversed in the anger and neutral conditions. Left amygdala activation showed a positive correlation with cooperation rates and self-reports of partner directed sympathy in the sympathy condition. In the anger condition, left putamen activation was positively correlated with cooperation rates and negatively correlated with self-reports of partner directed anger strength. These findings indicate that while the left amygdala activation may be indicative of emotion enhancement and increase of cooperative behavior, the left putamen may help to suppress an emotion to overcome anger and engage in cooperation.

Keywords: prisoner's dilemma, sympathy, anger, amygdala, putamen, cooperation, decision-making

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

Reviewed by:
Jan B. Engelmann,
University of Amsterdam, Netherlands
Thomas Zoëga Ramsøy,
Neurons Inc., Denmark

****Correspondence:***
Iveta Eimontaite
iveta.eimontaite@cranfield.ac.uk;
iveta.eimontaite@gmail.com
Vinod Goel
vgoel@yorku.ca

Specialty section:
This article was submitted to
Decision Neuroscience,
a section of the journal
Frontiers in Neuroscience

Received: *12 February 2019*
Accepted: *03 July 2019*
Published: *17 July 2019*

Citation:
Eimontaite I, Schindler I, De Marco M, Duzzi D, Venneri A and Goel V (2019) Left Amygdala and Putamen Activation Modulate Emotion Driven Decisions in the Iterated Prisoner's Dilemma Game. Front. Neurosci. 13:741. doi: 10.3389/fnins.2019.00741

INTRODUCTION

Human choice often involves tension between cooperation and non-cooperation. Actions to combat climate change provide a relevant real-world example. As a society we and our children would be much better off if we all cooperated in reducing carbon emissions (and jointly bear the costs). However, as an individual, if I bear the cost and reduce my carbon footprint (cooperate), and my neighbor continues to pollute (defect), he will reap a greater benefit than myself. If, however, I choose to continue polluting (defect), and my neighbor bears the cost of reducing carbon emissions (cooperate), I will reap the greatest benefit. If we both choose to continue polluting (defecting), we will both suffer equally. These types of choices are often formulated and studied in the laboratory as variations of the Prisoner's Dilemma game (Oskamp and Perlman, 1965).

For much of the 20th century, the dominant view of humans, embodied in the "homo economicus" model was as a utility maximiser as a consumer, and a profit maximiser as a producer.

On this model, decision-makers will exhibit perfect self-interested rationality and select the choice most advantageous for them. What makes such game theoretic tasks interesting is that the advantageous choice is dependent upon predicting the choice made by your opponent. Economic theory argues that defecting or not cooperating with your partner in the iterated Prisoner's Dilemma game can be consistent with utility maximizing behavior (Neyman, 1985). Data on such tasks show that participants will typically cooperate 40% of the time, while defecting approximately 60% of the time (Jones et al., 1968; Bó and Fréchette, 2011).

The "homo economicus" model is slowly changing as we begin to accept and accommodate the reality that various factors including emotions (Goel and Dolan, 2003; Goel and Vartanian, 2011; Halperin et al., 2013; Smith et al., 2014, 2015; Goel et al., 2017; Levine et al., 2017; Eimontaite et al., 2018), reward processing (Sanfey et al., 2003), Theory of Mind (Camerer, 2003) and individual differences in cognitive inhibition (De Neys et al., 2011), social orientation (Emonds et al., 2014), and trust (Chaudhuri et al., 2002; Lambert et al., 2017) modulate our decision-making. In fact, trust and cooperation are hard to separate and quite often these terms are used interchangeably while investigating social interaction games (Yamagishi et al., 2005). However, the attempts to separate cooperation and trust show that cooperation leads to trust (Chaudhuri et al., 2002; Yamagishi et al., 2005). Our focus here is on the effect of emotions on rational choice in the Prisoner's Dilemma game and development of cooperation as predecessor of trust in social interactions.

Common sense tells us that emotions should drive decisions by modulating subjective experiences (Scherer, 1982, 2005). Behavioral data, unsurprisingly, indicate that sympathy can encourage higher cooperation levels, even if it is costly/detrimental to the decision-maker (Bloom, 2017). Anger can trigger higher defection rates, again, even at a cost to the decision-maker (Bosman and Van Winden, 2002; Ben-Shakhar et al., 2004; Duersch and Servátka, 2007). In a study by Kopelman et al. (2006), participants in the role of sellers made higher demands while interacting with buyers displaying negative emotions by asking higher prices, and provided shorter warranty periods, etc. Buyers were less likely to sign a deal in the negative emotion condition compared to positive and neutral emotion conditions (Kopelman et al., 2006). The same pattern of behavior is observed in the iterated Prisoner's Dilemma with anger and sympathy emotions felt toward the other: sympathy toward the opponent increases cooperation, while anger toward the opponent increases defection compared to the neutral condition (Eimontaite et al., 2013). On the other hand, it is not only the valence of the emotion which needs to be considered, but also the motivation which is triggered by induced emotion. Engelmann and Hare (2018) review studies where withdrawal-related emotional states, such as sadness, fear and empathy, lead toward risk averse choices, while approach-related emotions, such as anger, lead to more risky decisions. These results, as Engelmann and Hare (2018) note, are also reflected in the neuroimaging study findings: choices under safety show activation in the ventromedial prefrontal cortex and ventral striatum, but not the insula. Yet, insula activation is evident under conditions of sadness and the perception of fairness.

Although emotions are important in decision-making, they are not the only factors determining the choice one will make. The perception of the possible rewards/gains or losses also affect decision-making processes (McCabe et al., 2001; Sanfey et al., 2003; King-Casas, 2005). Reward processing in the brain is marked by striatum activation (including putamen and caudate) and in economic games has shown an increase in activation associated with winnings (Elliott et al., 2003; Haruno, 2005; Hsu et al., 2008), and the decrease in activation with losses (Verney et al., 2003; Bjork, 2004).

Strategic thinking is also an important factor in decision-making and seems to be represented by medial prefrontal cortex activation (Blair et al., 1999; Frith, 2001; McCabe et al., 2001; Decety et al., 2004). High co-operators in the Trust Game showed stronger medial prefrontal cortex activation whilst interacting with human opponents as opposed to interacting with a computer. Yet for high defectors, activation of this region did not depend on the type of the opponent – human or computer (McCabe et al., 2001). Furthermore, deciding to trust individuals from the same racial group or not involved the striatum and amygdala (Stanley et al., 2012). In particular, striatum activation was recorded during representation of race-based reputations that shape trust decisions, while the amygdala was involved in processing emotionally relevant social group information. The amygdala is also critical for forming trust: patients with lesions to the amygdala tended to increase trust in response to betrayals in the Trust Game, while neurologically normal adults and patient controls show a decrease in trust after betrayals (Koscik and Tranel, 2011).

Tasks like the Prisoner's Dilemma can be presented either as single shot trials or multiple trials involving extended social interaction. The latter, iterated version of the task, presents the outcome of the interaction after each trial. This introduces complexity in terms of social context and reputation building, requiring additional strategizing (Camerer, 2003; Cuesta et al., 2015; Levine et al., 2017; Li et al., 2017). Reputation building involves monitoring the choices of your opponent in the context of your choices. If you cooperate on a particular trial, but your opponent chooses to defect, this will affect your decision on subsequent trials. But it will also have an emotional impact in terms of making you angry, upset, disappointed, or feeling cheated. In such a case, it is not clear how one would separate the effects of reputation building from emotions. Separating the influences of the emotional state of the decision-maker from strategic thinking in decision-making processes would allow further understanding of how various social influences shape decision-making. Separating the influences of emotional states of the decision-maker from strategic thinking in decision-making processes would allow further understanding of how various social influences shape decision-making. Some research has used iterated single shot games with unknown opponents to avoid reputation building effects (Ramsøy et al., 2015; Macoveanu et al., 2016), and this paradigm allows to investigate decision-making without prior emotion induction. However, adding emotion induction unrelated to the interaction in the game would be complicated in the context of iterated single shot games and difficult for the participant to keep track of.

The goal of the present study was to identify brain regions associated with decision-making in the Prisoner's Dilemma under the influence of three partner-directed emotion conditions: sympathy, anger, and neutral. Several previous neuroimaging studies have explored the effect of emotions on decision-making in the Prisoner's Dilemma game. In a study by Singer et al. (2004) participants and their opponents (who were not real) had to interact on a Prisoner's Dilemma type game. They had two conditions – make decisions by themselves (intentional) or follow predetermined decision by a computer. After the interaction, participants were asked to evaluate the other players and the results revealed sympathetic responses with cooperative opponents, and anger toward defecting opponents when these decisions were intentional (participants decided by themselves and were not determined by computer). In a follow-up study, after interaction in the Prisoner's Dilemma game, participants had to observe pain induction to cooperative opponents and this increased their anterior insula and anterior cingulate cortex activation (Singer et al., 2006). However, during the same pain induction to the unfair opponent, male participants showed increased activity in nucleus accumbens.

Rilling et al. (2008) looked at the interaction between reciprocated and unreciprocated cooperation in the iterated Prisoner's Dilemma game. In particular, opponent's defection after participants cooperation showed greater activation in bilateral anterior insula, left hippocampus and left amygdala, while bilateral ventral striatum showed deactivation. Furthermore, unreciprocated cooperation after previous cooperation compared to defection showed increased activity in anterior insula and left hippocampus. These results indicate that these areas are responsive to unreciprocated cooperation and anger emotion as reported by participants in a post-experiment questionnaire. Although these studies provide some insight into how emotions affect decision-making in socio-economic games, the emotion is triggered by the game play and it is hard to disengage whether emotions were driving the decision or they were incidental to the outcomes.

Our study differs from previous efforts in three respects. First, we avoided the potential confound of reputation building by keeping participants blind to the outcome of each trial in addition to avoiding one shot games. Second, participants' knowledge of the other player was built by emotion induction prior to the interaction in the Prisoner's Dilemma game. That is, emotions were triggered by an event that was incidental to the decision situation, but the emotion was decision-relevant as it was triggered by and directed toward their opponent in the Prisoner's Dilemma game. Finally, we compared the effect of two distinct emotions, sympathy and anger, in a within-subject design, allowing us to investigate cooperation and defection choices while controlling for individual differences. We predicted participants would show more cooperation in the sympathy condition compared to the neutral condition, and more defection in the anger compared to the neutral condition (Eimontaite et al., 2013). At the neural level we were interested in the interaction between emotion and choice and expected activation in the brain areas previously identified to be involved in emotional stimuli processing, processing of trustworthiness and decision-making. In particular, we predicted amygdala activation during processes where participants would embrace emotions and emotionally relevant information about individuals (Koscik and Tranel, 2011; Stanley et al., 2012), and striatum, and in particular putamen, activation for overcoming emotion effects (Padmala and Pessoa, 2010; Cutler and Campbell-Meiklejohn, 2019). That is, increased cooperation in sympathy condition would result in activation in the amygdala, while decreased defection in anger condition would show putamen involvement.

MATERIALS AND METHODS

Participants

Twenty-two Italian health care professionals employed at the IRCCS San Camillo Hospital Foundation (Venice, Italy) voluntarily took part in the study. Two participants were removed due to awareness of the deception and extensive head movement in the scanner, leaving 20 participants (6 males, 14 females) in the final analysis. Mean age of the participants was 29 years (*SD* = 5.68), and mean education was 16.4 years (*SD* = 3.54). Participants had normal or corrected to normal vision, 18 were right-handed. The study was approved by the University of Hull (United Kingdom) and the IRCCS San Camillo (Italy) ethics committees. Each participant provided written informed consent.

Task

The Prisoner's Dilemma game simulated a hypothetical situation whereby you and a partner are bankers suspected in corporate malfeasance. The police interrogate both of you separately, and offer each of you the following deal: If you provide the missing facts to the police (i.e., defect on your partner), and your partner stays silent, you will get a reward of €50,000 and your partner will pay a fine of €50,000. If both of you confess (defect on each other) and fill in the facts for the police, you will both pay a fine of €10,000. If you choose to stay silent (cooperate with your partner), and your partner fills in the facts for the police (defects), he/she will receive a reward of €50,000 and you will pay a fine of €50,000. If both of you choose to stay silent (cooperate), the police will not have enough evidence to convict either of you and will be forced to pay you €30,000 each for wrongful arrest. The payoff matrix is presented in **Figure 1**. The payoffs are a function of, not only the participant's selection, but also the selection of their partner. The task requires participants to make decisions that will maximize their hypothetical gains and minimize hypothetical losses. Each participant plays the game with three different partners, under three different emotion conditions.

Emotion Inducement

Steps were also taken to make participants feel anger or sympathy toward two of the three partners and remain neutral toward the third partner prior to the commencement of the game via an essay writing and evaluation task. Participants were asked to write a short essay describing something important to them. The

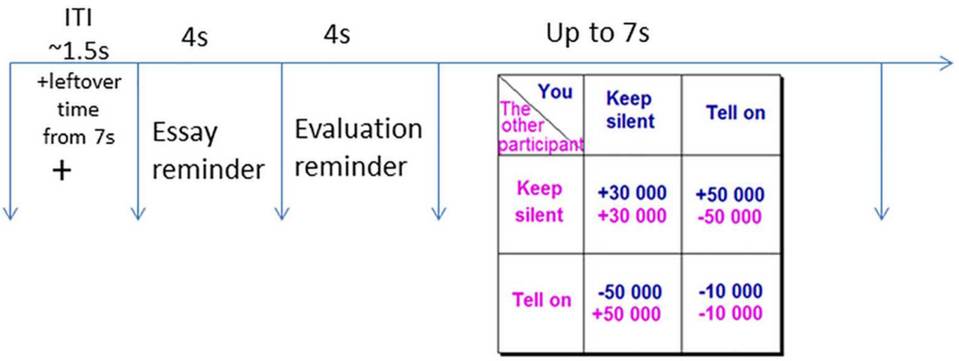


FIGURE 1 | A single trial of the Prisoner's Dilemma would start with a fixation cross (1.5 s), followed by a reminder of the other "participant's" essay and evaluation (each for 4 s). In the end the participant would make a choice (up to 7 s) on the Prisoner's Dilemma outcome table by pressing either the Index finger (cooperation) or the Middle finger (defection).

experimenter would take the essay out of the room, explaining that it would be given to their "partner" for comments/evaluation and that they would be required to evaluate the partner's essays. Approximately 5 min later the experimenter would return with one of the "partner's" essays for evaluation. After the evaluation was completed, the experimenter would take the evaluation and leave the room to retrieve the participant's essay evaluated by their "partner" (Eimontaite et al., 2013).

In actuality, the participant was being deceived. There were no other participants. The experimenter would return with the participant's essay, purportedly evaluated by their "partner." These evaluations consisted of the ratings of the essays on six 9-point bipolar scales (unintelligent–intelligent; thought provoking–boring; friendly–unfriendly; illogical–logical; respectable–unrespectable; irrational–rational), along with a space for free comments. In the sympathy condition the emotion was induced with an essay written by a young person coping with cancer [modified from Harmon-Jones et al. (2003)]. In this condition the evaluation of the participant's essay was rated neutrally (between 4 and 7 on the evaluation scales) and a hand-written positive comment "I can understand why a person would think like this" was left underneath the evaluation. In the anger condition, emotion was mainly triggered by the negative evaluation consisting of ratings that were weighted toward negative words (e.g., illogical or unacceptable). An insulting comment was also hand-written underneath the evaluation ("This is the stupidest thing I have ever read"). The essay in this condition was neutral in content, but a poorly written (grammatical mistakes, badly structured arguments). Finally, neutral emotion induction consisted of a neutral content essay, written in an unemotional and grammatically correct way, followed by a neutral (evaluations between 4 and 6) evaluation of the participant's own essay with no hand-written comments.

The procedure was repeated three times (once per each emotion condition) and photographs of both the opponent's essays and the evaluations of the participant's own essay were taken to strengthen the deception (all photographs were prefabricated before the experiment). These photographs of essay and evaluation were later presented before each Prisoner's Dilemma trial so that participants would know with whom they were interacting. Following emotion induction, participants previewed the uploaded photographs of their "partner's" essays and the evaluations they received on their own essays, to familiarize them with the digital versions.

The essays and evaluations were hand-written on different color paper (light blue, light purple, and light green) so that participants would learn to associate a color with a particular "partner." Colors associated with particular conditions hence the conditions were counterbalanced across participants.

The Task Presentation

An iterated 108-trial version (36 per emotion condition) of the game was used in the experiment. Each individual trial of the game would start with a fixation cross remaining on screen for 1.5 s on average followed by the scanned essay and the evaluation from the emotion induction 4 s each, both color-coded to provide content cues; serving as a rumination helping to prolong the emotion duration (Sbarra and Emery, 2005; Verduyn et al., 2009). Finally, the payoff matrix was presented for 7 s during which participants had to choose between cooperation and defection. If the participant made their decision in less than 7 s, the remaining time was added to the inter trial interval (ITI).

The order of the emotion conditions was pseudo-randomized, allowing a maximum of three consequent trials of the same emotion condition. Six different payoff matrices were presented and the amount possible to gain and lose in each had the same proportions (3: 5: −5: −1; i.e., participant cooperates/other cooperates: +€30,000/+€30,000; participant cooperates/other defects: +€50,000/−€50,000; participant defects/other cooperates: −€50,000/+€50,000; participant defects/other defects: −€10,000/−€10,000; **Figure 1**). Three pre-determined outcomes of the interaction ("You get €315,500 out of overall €730,000 possible earnings," "You get €396,000 out of overall €849,000 possible earnings," or "You get €745,000 out of overall €900,000 possible earnings") were counterbalanced between three runs. These outcomes were presented only after 36 trials to avoid a reputation effect, and were independent of the participants' responses (**Figure 1**). Participants were not provided information about the opponent earnings. The dependent measure was the mean number of defection and cooperation per emotion condition.

In addition to the iterated Prisoner's Dilemma game participants completed a self-report emotion questionnaire to evaluate the success of emotion induction [adapted from Harmon-Jones et al. (2003), and Harmon-Jones and Sigelman (2001)]. Words being semantically related to sympathy, compassion, and sadness were pooled into a sympathy word group (Cronbach's Alpha 0.914, n = 17). Similarly, words indicative of anger and fear emotions were combined to an anger emotion word list (Cronbach's Alpha 0.875, n = 17), and the neutral emotion word list contained adjectives associated with positive affect (Cronbach's Alpha 0.834, n = 18). Further, a mixed ANOVA confirmed that each emotion was successfully induced as planned: in the sympathy condition, the sympathy word group was rated highest as well as the anger word group in the anger emotion condition (**Supplementary Materials**).

Procedure

Before signing informed consent and agreeing to take part in the study, participants were informed that the purpose of the study is to investigate various reasoning processes. They were told that they will need to interact with other individuals in this study on some of the tasks, however, other tasks will be completed just on their own. After this, participants took part in the essay writing/emotion induction task. After emotion induction, participants were taken to the fMRI room, where they were reminded of the rules of the Prisoner's Dilemma before playing it. The experiment consisted of three runs with 36 trials per run (12 trials of sympathy, 12 of anger, and 12 of neutral emotion condition). Each run lasted for 11.5 min. Participants did not receive the reimbursement depending on their performance in the Prisoner's Dilemma game. After the scanning procedure, participants filled in the Self-Report Emotion Questionnaire. Finally, questions establishing the participant's belief in the deception were asked and the full debrief was given providing the true aims of the experiment.

fMRI Acquisition

Scanning was performed at the IRCCS San Camillo using a 1.5T Phillips Achieva MRI scanner operated with a Sense eight channel head coil. The experiment was divided into three functional runs, with time to rest between runs. Functional scans were acquired by using manufacturers standard single shot EPI sequence [TR = 2060 ms, echo time (TE) = 45 ms, flip angle = 90°, 25 slices, slice thickness = 5 mm, no gap, matrix size 80 × 80, voxel size 2.88 × 2.88 × 5 mm, FOV = 230 × 230 mm]. At the start of the scanning each participants' fieldmap was acquired (T1 weighted fast field echo sequence, TE long = 7.6 ms, TE short 4.9 ms, slice thickness 5 mm, matrix size 72 × 60, no gap, voxel size 0.8 × 0.8 × 5 mm). Fieldmaps were used to correct EPI images for static geometric distortions caused by susceptibility-induced field inhomogeneities and head movement (Andersson et al., 2001; Hutton et al., 2002). To aid intersubject registration, at the end of each scanning session, a 3D T1-weighted structural scan was acquired for each participant (Fast field gradient echo sequence, TR = 7.4 ms, TE = 3.4 ms, 280 slices, slice thickness = 0.6 mm, matrix 240 × 240, voxel size 1.04 × 1.06).

fMRI Analysis

Image pre-processing and data analysis were carried out using Statistical Parametric Mapping software in Matlab 2016a (SPM12; Wellcome Centre for Human Neuroimaging at UCL). The first 6 dummy volumes of each run were discarded to allow for T1 equilibration, and then the EPI images were corrected for geometric distortions caused by susceptibility-induced field inhomogeneities. Field maps were first brain extracted using FSL BET (Smith, 2002) and then processed for each participant using the FieldMap toolbox in SPM (Hutton et al., 2004). The EPI images were then realigned and unwarped (Andersson et al., 2001). Each participant's structural image was coregistered to the mean of the motion-corrected functional images using a 12-parameter affine transformation, and segmented according to the default procedure in SPM12 (Ashburner and Friston, 2005). The spatial normalization parameters resulting from the previous step were applied to the functional images to allow for intersubject analysis. Finally, these images were smoothed using a 6 mm full width at half maximum Gaussian kernel.

For each participant, an event-related general linear model (GLM) was designed. The GLM consisted of regressors of interest: the onsets of the Prisoner's Dilemma payoff matrix separately for cooperation and defection in each emotion condition (sympathy, anger and neutral; at the time when the payoff matrix appeared on the screen until participants made their choice and pressed the button, on average lasting 1.96 s, SD = 1.22). Motion parameters defined by the realignment procedure were entered as regressors of no interest, separately for each run.

Time derivatives were used and runs where either of the emotion conditions did not have a single defection or cooperation were removed (eight runs overall). Statistical parametric maps were generated from contrasts of interest: [sympathy (defection vs. cooperation) vs. neutral (defection vs. cooperation)], and [anger (defection vs. cooperation) vs. neutral (defection vs. cooperation)].

A random-effects group-level analysis using one-sample t-tests on the contrast images obtained from each contrast of interest for each participant was used with peak uncorrected $p \leq 0.005$ and extent threshold of $k = 20$ (multiple testing was accounted for on cluster level based corrected $pFWE$ of 0.05). This threshold was suggested to be comparable to FWE corrected thresholds according to Lieberman and Cunningham (2009), and Lieberman et al. (2009), however, further discussion by Eklund et al. (2016) shows that clusterwise inferences increase false positive error.

RESULTS

Behavioral Impact of Emotion on Social Decision-Making

To investigate the effect of sympathy, anger and neutral emotion on defection and cooperation rates, a repeated measures *ANOVA*,

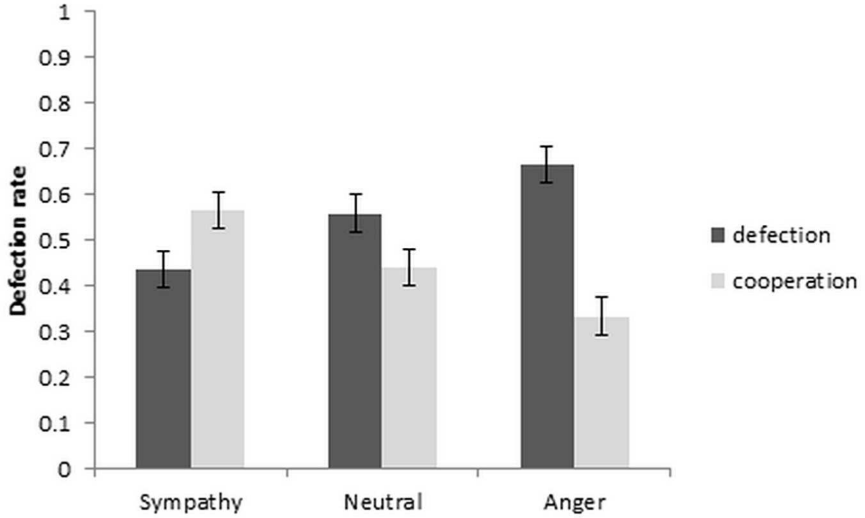


FIGURE 2 | Defection and cooperation rate as a function of emotion condition (±1SEM).

with independent variable of emotion condition and dependent variable of defection rate was used. The results are depicted in **Figure 2**. A significant repeated measures *ANOVA* [$F(2, 36) = 6.97$, $p = 0.003$, $\eta_p{}^2 = 0.279$] with *post hoc* comparisons between the emotion conditions showed that the cooperation rate increased significantly from neutral to sympathy [$t(19) = 2.79$, $p = 0.012$, $d_z = 0.624$], and decreased from neutral to anger at a trend level [$t(19) = -2.07$, $p = 0.052$, $d_z = 0.463$]. Cooperation also increased from anger to sympathy conditions [$t(19) = 4.13$, $p = 0.001$, $d_z = 0.923$]. Within-subject contrast showed a significant linear trend [$F(1, 19) = 7.02$, $p = 0.016$, $\eta_p{}^2 = 0.270$].

Further analysis of the reaction times with a repeated measures *ANOVA* showed the main effect of emotion as well as the emotion by choice interaction to be significant [$F(2,38) = 6.23$, $p = 0.005$, $\eta_p{}^2 = 0.247$ and $F(2,38) = 4.55$, $p = 0.017$, $\eta_p{}^2 = 0.193$, respectively]. The paired t-tests between the defection and cooperation choice in each emotion condition revealed significantly quicker RT's in the sympathy condition for cooperation than defection, and also significantly quicker RT's in defection than cooperation in the neutral condition [$t(19) = -2.15$, $p = 0.045$, $d_z = 0.481$ and $t(19) = -3.20$, $p = 0.005$, $d_z = 0.716$, respectively]. Although the reaction time in the anger condition increased from defection to cooperation choice, the increase was not significant [$t(19) = -1.13$, $p = 0.274$, $d_z = 0.253$; **Table 1**].

TABLE 1 | Mean response time (seconds) in cooperation and defection choices (SD) and mean defection rates (SD) as a function of the emotion condition.

	Emotion condition		
	Sympathy	**Neutral**	**Anger**
Defection response time	1.90 (0.17)	1.71 (0.15)	1.57 (0.13)
Cooperation response time	1.60 (0.13)	1.93 (0.18)	1.71 (0.16)
Defection rate	0.44 (0.20)	0.56 (0.20)	0.67 (0.19)

Imaging Results

The behavioral results indicated that anger directed at the other player increases defection, while sympathy directed at the other player increases cooperation. To isolate the neural basis of increased defection responses in the anger condition we undertook Emotion by Choice interaction analysis, comparing the BOLD signal change in the various emotion conditions as a function of defection and cooperation. We present the following three interaction contrasts (and their reverse) below: (1) [anger (defection – cooperation) – neutral (defection – cooperation)]; (2) [sympathy (defection – cooperation) – neutral (defection – cooperation)]; (3) [sympathy (defection – cooperation) – anger (defection – cooperation)]. The neural activations associated with the decision making, independent of emotions, are included in the **Supplementary Materials**.

Activation Associated With Defection in Anger Condition

We used the contrast [anger (defection-cooperation) – neutral (defection-cooperation)] to compare the differential effects of Defection and Cooperation in Anger and Neutral conditions. It showed activation in the bilateral putamen, and the right posterior cingulate BA 23 ($P_{\text{FWE}} < 0.05$, **Table 2** and **Figures 3A,C,D**).

We reasoned that if this activation is a reflection of the participants' choice to defect because of anger directed at their partner, then there should be a significant correlation between percent signal change and cooperation in the anger condition, but not in the sympathy or neutral conditions. Furthermore, the subjective rating from the self-report emotion questionnaire for anger words should correlate with the percent signal change. In fact, Pearson's correlation coefficient showed a positive correlation between interaction contrast percentage signal change in the left putamen and the cooperation rate in the anger condition cooperation trials ($r = 0.45$, $p = 0.045$, respectively;

TABLE 2 | Regions of increased activation in the contrasts comparing the sympathy, anger and neutral emotion conditions between each other.

Brain region	Brodmann area	Hemisphare	# of voxels	peak T	MNI coordinates		
					x (mm)	y (mm)	z (mm)
Activation associated with defection in anger condition: a(d-c)-n(d-c)							
Sub-lobar							
Lentiform nucleus, Putamen*		L	56	3.45	−18	5	−8
Lentiform Nucleus, Putamen*		R	41	3.36	24	11	−5
Limbic Lobe							
Posterior Cingulate**	BA 23	R	58	4.3	3	−37	22
Activation associated with defection in Sympathy condition: s(d-c)-n(d-c)							
Limbic Lobe							
Uncus, Superior Temporal Pole*	BA 28	R	321	4.84	27	5	−23
Uncus, Amygdala**		L	108	4.66	−21	−1	−23
Cingulate Gyrus*	BA 23	R	691	7.33	3	−28	34
Activation associated with defection in anger vs. sympathy condition: a(d-c)-s(d-c)							
Sub-lobar							
Lenntiform Nucleus, Putamen**		L	27	2.94	−18	5	−8
Activation associated with defection in sympathy vs. anger condition: s(d-c)-a(d-c)							
Frontal Lobe							
Medial Frontal Gyrus**	BA 10	L	15	−4.05	−9	56	−8

**Cluster – level $P_{FWE} \leq 0.05$. **Cluster – level $P_{uncorrected} \leq 0.05$.*

Figure 3B). Furthermore, anger emotion strength as measured with the self-report emotion questionnaire negatively correlated with percent signal change in the left putamen (r = −0.50, p = 0.047). The correlation between anger emotion word ratings and the behavioral cooperation was negative but not significant (r = −0.30, p = 0.207).

Finally, there was no significant correlation between defection/cooperation and interaction contrast percent signal change in the left putamen in the neutral condition (r = −0.08, p = 0.751). Correlations between cooperation/defection in the anger condition and percent signal change in the left cingulate gyrus (BA 23), as well as between the cooperation/defection in the neutral condition and the percent signal change in the neutral condition in the left putamen, and the left posterior cingulate gyrus (BA 23) were not significant ($r < 0.25$, $p > 0.288$).

The reverse contrast [neutral (defection-cooperation) – anger (defection-cooperation)] did not show any significant activations.

Sympathy and Neutral Interaction With Defection and Cooperation Choice

The sympathy condition results in increased levels of cooperation. To isolate the neural basis of increased cooperation (decreased defection) in the sympathy condition we utilized the following contrast: [sympathy (defection – cooperation) – neutral (defection - cooperation)]. The contrast revealed activation in the right superior temporal pole (BA 28) (cluster level $P_{\text{FWE}} < 0.05$), and activation in the left amygdala (cluster level $P_{\text{uncorrected}} < 0.05$; **Table 2** and **Figures 4A,C,D**).

Again correlation analyses were performed to test for a relationship between the cooperation in the sympathy and neutral emotion conditions in the activated areas. The left amygdala interaction contrast percent signal change positively correlated with cooperation in the sympathy condition (r = 0.57, p = 0.009; **Figure 4B**). In addition, the increase in the interaction contrast percent signal change in the left amygdala was positively correlated with self-report scores for sympathy words (r = 0.46, p = 0.043). In contrast, the positive correlation between self-report words and cooperation was not significant (r = 0.28, p = 0.234). Correlations between percent signal change in the left amygdala and the corresponding decisions were not significant ($r \leq 0.37$, $p \geq 0.110$) in the neutral emotion condition. The correlation in the right superior temporal pole and the left putamen with the defection in the sympathy and neutral conditions were not significant ($r < 0.33$, $p > 0.15$).

The reverse contrasts [neutral (defection – cooperation) – sympathy (defection – cooperation)] showed no significant activation.

Sympathy and Anger Interaction With Defection and Cooperation Choice

Finally, we examined the response by emotion (sympathy and the anger) interaction, [anger (defection – cooperation) – sympathy (defection – cooperation)], revealing activation in the left putamen, cluster level- $P_{\text{uncorrected}} \leq 0.05$. The reversed contrast [sympathy (defection-cooperation) – anger (defection-cooperation)] showed activation in the left medial frontal gyrus (BA 10) (Cluster level- $P_{\text{uncorrected}} \leq 0.05$).

The correlation between percent signal change in the left putamen with cooperation in the anger condition was a trend (r = 0.44, p = 0.052), while in the defection trials and in the sympathy condition cooperation and defection trials correlation was not significant ($r \leq -0.36$, $p \geq 0.12$). The left middle frontal gyrus (BA 10) activation in the anger and the sympathy emotion conditions did not correlate with the defection rate neither in defection nor in cooperation trials ($r \leq -0.38$, $p \geq 0.10$).

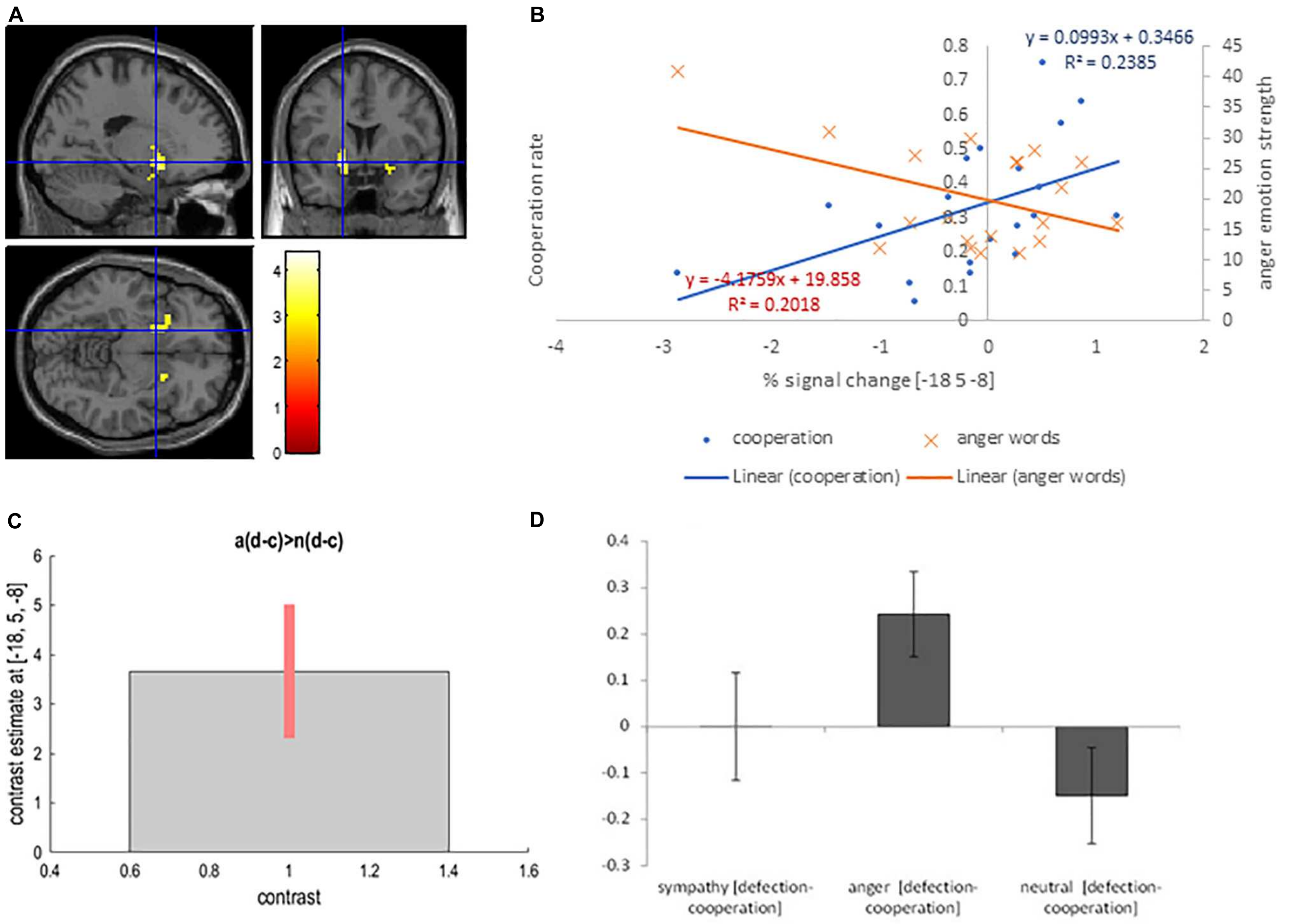


FIGURE 3 | (A) Bilateral putamen activation overlaid on the MNI single subject template based on the interaction contrast [anger (defection – cooperation) – neutral (defection-cooperation)]; **(B)** Correlation between the interaction contrast percent signal change in the left putamen and the cooperation rate in the anger condition (blue) and correlation between the percent signal change in the left putamen and the strength of anger word rating in the anger condition (orange); **(C)** Contrast estimate and 90% Confidence Interval for interaction contrast; **(D)** Difference in percent signal change between defection and cooperation trials in sympathy, anger and neutral emotion condition for interaction contrast.

DISCUSSION

The current study investigated the functional neuroanatomy of cooperation and defection responses in the Prisoner's Dilemma game under conditions of partner directed sympathy, anger, or neutral emotions. The outcome of the game was presented after each run (36 trials with three opponents) and did not provide information about what the opponents received. The reputation building was induced prior to the Prisoner's Dilemma game via the emotion induction task. The behavioral results confirmed the effectiveness of the manipulation. As expected, participants' cooperation rates increased significantly from the neutral to the sympathy condition and decreased from the neutral to the anger condition (trend level). Consistent with this, the sympathy condition elicited quicker responses for cooperation than defection choices, whereas this pattern was reversed in the anger and neutral conditions. Imaging results showed (relative) greater activation in the left putamen, in the anger condition, and in left amygdala, in the sympathy condition, compared to the neutral condition, in response to cooperation choices.

Left putamen percent signal change positively correlated with cooperation rate. Furthermore, self-reported anger emotion strength was negatively correlated with percent signal change in this area. These results suggest that relative increase in left putamen activation corresponds to more cooperative behavior, and given the negative correlation with self-report anger words strength, the putamen activation may be important for overcoming the desire to retaliate.

Previous studies have documented the role of striatum, and in particular, left putamen, in emotion regulation. In one study, participants were shown emotionally neutral faces and asked to engage either in positive emotion reappraisal (think positively about the face) or negative emotion reappraisal (think negatively) (Richey et al., 2015). Left putamen activation was observed during positive reappraisal trials. Furthermore, not only

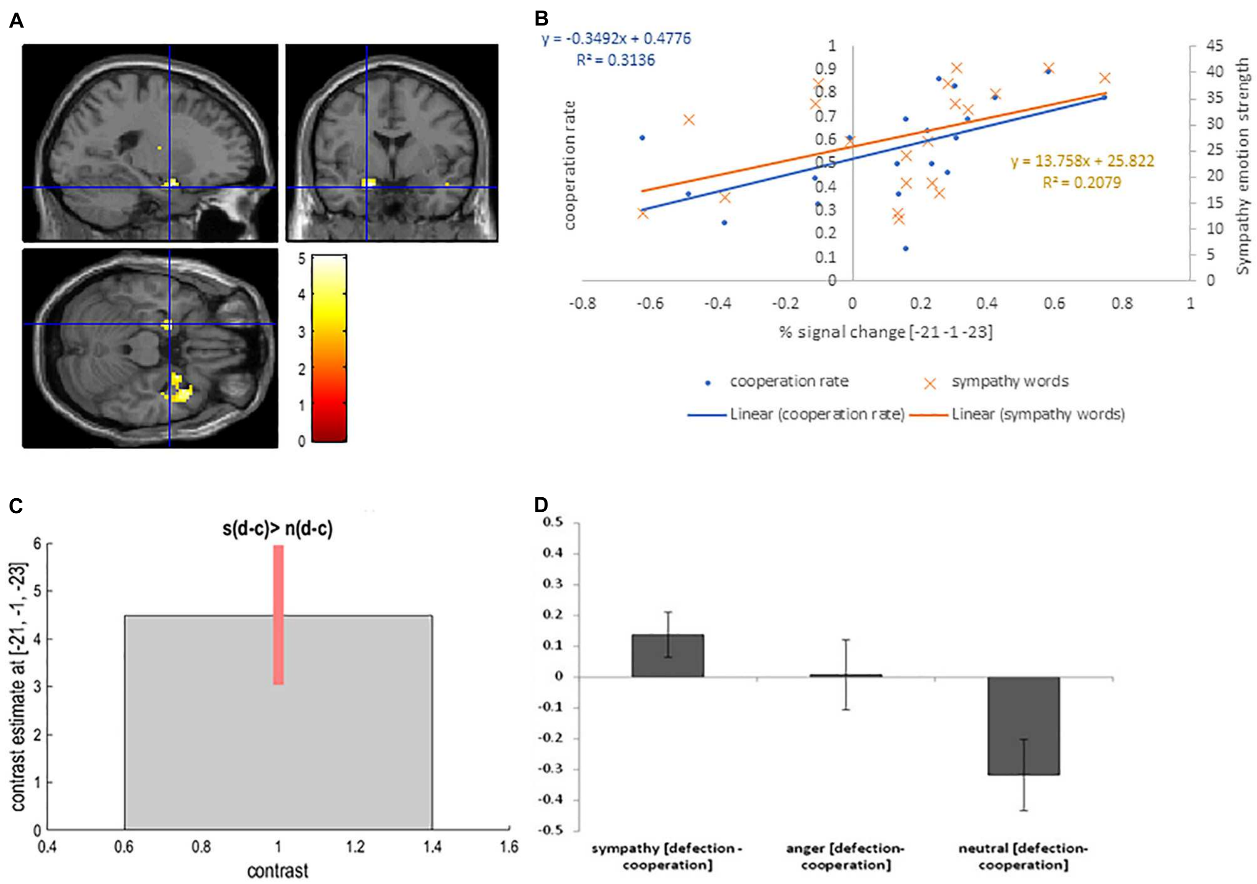


FIGURE 4 | (A) Left amygdala activation overlaid on the MNI single subject template based on the interaction contrast [sympathy (defection – cooperation) – neutral (defection-cooperation)]; **(B)** Correlation between the interaction contrast percent signal change in the left amygdala and the cooperation rate in the sympathy condition (blue), and correlation between the percent signal change in the left putamen and the strength of sympathy word rating in the sympathy condition (orange); **(C)** Contrast estimate and 90% Confidence Interval for interaction contrast; **(D)** Difference in percent signal change between defection and cooperation trials in sympathy, anger and neutral emotion condition for interaction contrast.

reappraisal, but also emotion suppression elicits activation in the left (and also right) putamen. Vanderhasselt et al. (2013) asked participants to view negative and high arousing images and either suppress the emotion or engage in negative emotion reappraisal. Negative emotion suppression, but not reappraisal, showed increased bilateral putamen activation. This is consistent with our suggestion that the putamen activation in the anger condition may be linked to overcoming the emotion and cooperating despite the anger directed at the partner.

The contrast sympathy (defection – cooperation) – neutral (defection – cooperation) showed activation in the left amygdala. This activation was positively correlated with both cooperation rates and self-reported sympathy emotion ratings: higher amygdala activation related to higher cooperation rates and higher scores on self-report sympathy emotion strength. These findings suggest that relative activation of the amygdala in the sympathy condition corresponds to increased cooperating responses and the use of more sympathy words to describe the participant. The finding is consistent with past studies. In the Singer et al. (2004) study, a cooperative opponent in the Prisoner's Dilemma triggered sympathetic responses from the participants (as revealed by the post-trial questionnaire). Furthermore, intentional decision to cooperate by the opponent in this study was associated with increased amygdala activation; the left amygdala was activated when participants were presented with a photo of an intentional cooperator (person who decided to cooperate themselves instead of being assigned this decision by a computer). In a study investigating incidental fear during the Trust Game with social (with a human opponent) and non-social (decisions generated by computer) trials, results showed that a strong unexpected electrical shock (Threat of Shock, ToS) can reduce trust transfer rates in both social and non-social conditions (Engelmann et al., 2019). Furthermore, in the absence of ToS, a significant connectivity was observed between the temporo parietal junction (TPJ) and amygdala during social trust trials, but this connectivity was disrupted by the introduction of ToS. The authors suggest that the TPJ-amygdala connectivity present when there are no aversive emotional stimuli reflects information pre-processing occurring both cognitively (i.e., mentalizing, TPJ) and emotionally (trustworthiness of the opponent assessment, amygdala). However, once threatening stimuli are introduced, this connectivity is broken: the amygdala shows suppression and therefore breaks its communication with the TPJ, reducing one's ability to mentalize. This suggests that increased amygdala activity indicates not only of emotional stimuli preprocessing, but also shows mentalizing processes. In another study, participants making altruistic decisions (cooperation) as opposed to selfish decisions (defection) also showed amygdala activation in the Prisoner's Dilemma game (Cutler and Campbell-Meiklejohn, 2019). Increased amygdala activation might have been the result of participants not expecting their cooperation to be reciprocated: opponent's unreciprocated cooperation toward participants' resulted in increased activation of participants left amygdala (Rilling et al., 2008). Another explanation might come from the research exploring hippocampus and amygdala connectivity in episodic emotional memories (Phelps, 2004). Participants under the condition of receiving instructed anticipated emotional stimuli (indication of possible electric shock) showed increased left amygdala activation. This suggests that episodic memories can influence an individual's emotional reactions in part by modulating amygdala activation. In the current study, it is possible that participants had episodic memories about the sympathy-triggering stimuli and experienced sadness toward the other. Therefore, they were possibly anticipating to feel guilty if they would choose defection and this resulted in the choice of cooperation and showed an increased left amygdala activation during these choices.

Additionally, the amygdala is part of the human reinforcement expectancies system which is involved in learning the signs of distress of others and in this way guiding individuals from antisocial behavior (Ray et al., 2005; Blair, 2007) and helping to solve moral dilemmas (Greene et al., 2004). As anticipatory emotions can guide individuals from antisocial behavior (Rick and Loewenstein, 2008), expectation of the guilt arising from their decision results in higher cooperation rates, which is in line with the withdrawal emotion function (Engelmann and Hare, 2018). Incidental sadness, which is related to sympathy emotion in our study, does not show the same reward processing as in the neutral emotion condition (Harlé et al., 2012). Furthermore, at the neural level, researchers found that the left ventral striatum showed stronger activation in the neutral condition (indicating reward processing) but in the sad condition, this pattern was not observed. Behaviorally, sad participants had a stronger preference toward fair offers during social interactions. Therefore, it is possible to assume that in our sympathy condition, participants were proposing fairer decisions in the Prisoner's Dilemma game.

One unexpected finding was that the decrease in cooperation rates from neutral to anger was only a trend (Eimontaite et al., 2013). One possible explanation for the lack of significance may be that participants were medical personnel. Compassion and empathy are desirable skills in nurses and health care workers as they need to interpret and understand the feelings of their patients as well as demonstrate compassion for their condition (Morse, 1991) in addition to being able to restrain negative actions, remain calm and in control of their behavior in a stressful situation (Zhang et al., 2001). Due to these professional characteristics, the participants might have shown a strong response toward the partner in the sympathy condition, and may have been able to control their negative emotions in the anger condition. However, this requires further exploration.

Furthermore, not having actual reimbursement for the Prisoner's Dilemma game might make the participant feel like the interactions are without real consequences and could be criticized on the grounds of not including choices with real outcomes and consequences. This could potentially have led to a stronger cooperation response in the sympathy condition compared to defection in anger. The results should be explored further within an environment where participants are reimbursed based on their decisions. Furthermore, the current study used an extent $k = 20$ and peak level $p = 0.005$ (uncorrected) combination threshold. This threshold was comparable to the Family Wise Error-corrected thresholds of Lieberman and Cunningham (2009), and Lieberman et al. (2009). Further discussion on the subject by Eklund et al. (2016) considered clusterwise thresholds to be invalid, and the results of the current study can be considered, therefore, exploratory. Finally, the small sample size is a limitation of the current study in terms of correlation results. Yarkoni (2009) argues that small sample size correlations results in power issues. A solution for this issue is a recommended increase in sample size ($N > 50$) (Yarkoni, 2009).

The current study investigated the effect of emotions on the iterated Prisoner's Dilemma game, with the outcome of multiple interactions with three opponents revealed only after 36 trials, thus avoiding reputation building through game, and reputation building was only induced through emotion manipulation before participants performed on the Prisoner's Dilemma game. The results show that the effects of partner directed sympathy and anger emotions on decision-making are represented by modulation of activation in the left putamen and the left amygdala. In particular, increased (relative) activation in left putamen is associated with increased cooperation decisions, even in the face of partner directed anger. Left amygdala activation increased (relatively) in response to increased number of cooperation responses in the partner sympathy directed condition. In addition, reaction times increased for decisions where participants went against their emotional impulse, providing further support, showing the conflict between emotional and rational. These results are important as they provide further evidence for the role of the left putamen and left amygdala in social exchange decision-making under the influence of partner directed emotion, yet without reputation building effects.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding authors.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the University of Hull (United Kingdom) and the IRCCS San Camillo (Italy) with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the departmental ethics committee, University of Hull (United Kingdom) and the IRCCS San Camillo (Italy).

AUTHOR CONTRIBUTIONS

IE, MDM, and DD performed the measurements. VG, IS, and AV were involved in planning and supervised the work. IE processed the experimental data, performed the analysis, drafted the manuscript, and designed the figures. VG aided in interpreting the results and worked on the manuscript. All authors discussed the results and commented on the manuscript.

FUNDING

This work was supported, in part, by a Wellcome Trust grant (#089233) to VG.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.00741/full#supplementary-material

REFERENCES

Andersson, J. L. R., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage* 13, 903–919. doi: 10.1006/nimg.2001.0746

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., and Van Winden, F. (2004). *Reciprocity and Emotions: Arousal, Self-Reports, and Expectations.* CESifo Working Papers, 1298, Amsterdam: University of Amsterdam.

Bjork, J. M. (2004). Incentive-elicited brain activation in adolescents: similarities and differences from young adults. *J. Neurosci.* 24, 1793–1802. doi: 10.1523/JNEUROSCI.4862-03.2004

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends Cogn. Sci.* 11, 387–392. doi: 10.1016/j.tics.2007.07.003

Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I., and Dolan, R. J. (1999). Dissociable neural responses to facial expressions of sadness and anger. *Brain* 122, 883–893. doi: 10.1093/brain/122.5.883

Bloom, P. (2017). Empathy and its discontents. *Trends Cogn. Sci.* 21, 24–31. doi: 10.1016/j.tics.2016.11.004

Bó, P. D., and Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: experimental evidence. *Am. Econ. Rev.* 101, 411–429. doi: 10.1257/aer.101.1.411

Bosman, R., and Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *Econ. J.* 112, 147–169. doi: 10.1111/1468-0297.0j677

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends Cogn. Sci.* 7, 225–231. doi: 10.1016/S1364-6613(03)00094-9

Chaudhuri, A., Sopher, B., and Strand, P. (2002). Cooperation in social dilemmas, trust and reciprocity. *J. Econ. Psychol.* 23, 231–249. doi: 10.1016/S0167-4870(02)00065-X

Cuesta, J. A., Gracia-Lázaro, C., Ferrer, A., Moreno, Y., and Sánchez, A. (2015). Reputation drives cooperative behaviour and network formation in human groups. *Sci. Rep.* 5:7843. doi: 10.1038/srep07843

Cutler, J., and Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *NeuroImage* 184, 227–241. doi: 10.1016/j.neuroimage.2018.09.009

De Neys, W., Novitskiy, N., Geeraerts, L., Ramautar, J., and Wagemans, J. (2011). Cognitive Control and Individual Differences in Economic Ultimatum Decision-Making. *PLoS One* 6:e27107. doi: 10.1371/journal.pone.0027107

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., and Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *NeuroImage* 23, 744–751. doi: 10.1016/j.neuroimage.2004.05.025

Duersch, P., and Servátka, M. (2007). *Risky Punishment and Reward in the Prisoner's Dilemma.* Heidelberg: University of Heidelberg.

Eimontaite, I., Goel, V., Raymont, V., Krueger, F., Schindler, I., and Grafman, J. (2018). Differential roles of polar orbital prefrontal cortex and parietal lobes in logical reasoning with neutral and negative emotional content. *Neuropsychologia* 119, 320–329. doi: 10.1016/j.neuropsychologia.2018.05.014

Eimontaite, I., Nicolle, A., Schindler, I., and Goel, V. (2013). The effect of partner-directed emotion in social exchange decision-making. *Front. Psychol.* 4:469

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113

Elliott, R., Newman, J. L., Longe, O. A., and Deakin, J. W. (2003). Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. *J. Neurosci.* 23, 303–307. doi: 10.1523/jneurosci.23-01-00303.2003

Emonds, G., Declerck, C. H., Boone, C., Seurinck, R., and Achten, R. (2014). Establishing cooperation in a mixed-motive social dilemma, an fMRI study investigating the role of social value orientation and dispositional trust. *Soc. Neurosci.* 9, 10–22. doi: 10.1080/17470919.2013.858080

Engelmann, J B., Meyer, F., Ruff, C. C., and Fehr, E. (2019). The neural circuitry of affect-induced distortions of trust. *Sci. Adv.* 5:eaau3413. doi: 10.1126/sciadv.aau3413

Engelmann, J. B., and Hare, T. A. (2018). "Emotions can bias decision-making processes by promoting specific behavioral tendencies". in *The Nature of Emotion: Fundamental Questions*. eds. A S. Fox, R C. Lapate, A J. Shackman, and R J. Davidson (New York, NY: Oxford University Press).

Frith, U. (2001). Mind blindness and the brain in autism. *Neuron* 32, 969–979. doi: 10.1016/S0896-6273(01)00552-9

Goel, V., and Dolan, R. J. (2003). Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning. *Neuroimage* 20, 2314–2321. doi: 10.1016/j.neuroimage.2003.07.027

Goel, V., and Vartanian, O. (2011). Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cogn. Emot.* 25, 121–131. doi: 10.1080/02699931003593942

Goel, V., Lam, E., Smith, K. W., Goel, A., Raymont, V., Krueger, F., et al. (2017). Lesions to polar/orbital prefrontal cortex selectively impair reasoning about emotional material. *Neuropsychologia* 99, 236–245. doi: 10.1016/j.neuropsychologia.2017.03.006

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 389–400. doi: 10.1016/j.neuron.2004.09.027

Halperin, E., Porat, R., Tamir, M., and Gross, J. J. (2013). Can emotion regulation change political attitudes in intractable conflicts? from the laboratory to the field. *Psychol. Sci.* 24, 106–111. doi: 10.1177/0956797612452572

Harlé, K. M., Chang, L. J., van 't Wout, M., and Sanfey, A. G. (2012). The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. *Neuroimage* 61, 32–40. doi: 10.1016/j.neuroimage.2012.02.027

Harmon-Jones, E., and Sigelman, J. (2001). State anger and prefrontal brain activity: evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression. *J. Pers. Soc. Psychol.* 80, 797–803. doi: 10.1037/0022-3514.80.5.797

Harmon-Jones, E., Peterson, H., and Vaughn, K. (2003). The dissonance-inducing effects of an inconsistency between experienced empathy and knowledge of past failures to help: support for the action-based model of dissonance. *Basic Appl. Soc. Psychol.* 25, 69–78. doi: 10.1207/S15324834BASP2501_5

Haruno, M. (2005). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *J. Neurophysiol.* 95, 948–959. doi: 10.1152/jn.00382.2005

Hsu, M., Anen, C., and Quartz, S. R. (2008). The Right and the good: distributive justice and neural encoding of equity and efficiency. *Science* 320, 1092–1095. doi: 10.1126/science.1153651

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., and Turner, R. (2002). Image distortion correction in fMRI: a quantitative evaluation. *NeuroImage* 16, 217–240. doi: 10.1006/nimg.2001.1054

Hutton, C., Deichmann, R., Turner, R., & Andersson, J. L. R. (2004). "Combined correction for geometric distortion and its interaction with head motion in fMRI". in *Proceedings of the. ISMRM 12*, Kyoto.

Jones, B., Steele, M., and Gahagan, J. (1968). Matrix values and cooperative behavior in the prisoner's dilemma game. *J. Pers. Soc. Psychol.* 8, 148–153. doi: 10.1037/h0025299

King-Casas, B. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83. doi: 10.1126/science.1108062

Kopelman, S., Rosette, A. S., and Thompson, L. (2006). The three faces of eve: strategic displays of positive, negative, and neutral emotions in negotiations. *Organ. Behav. Hum. Decis. Process.* 99, 81–101. doi: 10.1016/j.obhdp.2005.08.003

Koscik, T. R., and Tranel, D. (2011). The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia* 49, 602–611. doi: 10.1016/j.neuropsychologia.2010.09.023

Lambert, B., Declerck, C. H., Emonds, G., and Boone, C. (2017). Trust as commodity: social value orientation affects the neural substrates of learning to cooperate. *Soc. Cogn. Affect. Neurosci.* 12, 609–617. doi: 10.1093/scan/nsw170

Levine, E. E., Barasch, A., Rand, D. G., Berman, J. Z., and Small, D. A. (2017). Signaling emotion and reason in cooperation. *J. Exp. Psychol. Gen.* 147, 702–719. doi: 10.2139/ssrn.2922765

Li, J., Zhang, C., Sun, Q., Chen, Z., and Zhang, J. (2017). Changing the intensity of interaction based on individual behavior in the iterated prisoner's dilemma game. *IEEE Trans. Evol. Comput.* 21, 506–517. doi: 10.1109/TEVC.2016.2628385

Lieberman, M D., Berkman, E. T., and Wager, T. D. (2009). Correlations in social neuroscience aren't voodoo: commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* 4, 299–307. doi: 10.1111/j.1745-6924.2009.01128.x

Lieberman, M. D., and Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Soc. Cogn. Affect. Neurosci.* 4, 423–428. doi: 10.1093/scan/nsp052

Macoveanu, J., Ramsoy, T. Z., Skov, M., Siebner, H. R., and Fosgaard, T. R. (2016). The neural bases of framing effects in social dilemmas. *J. Neurosci. Psychol. Econ.* 9, 14–28. doi: 10.1037/npe0000050

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11832–11835. doi: 10.1073/pnas.211415698

Morse, J. M. (1991). Negotiating commitment and involvement in the nurse-patient relationship. *J. Adv. Nurs.* 16, 455–468. doi: 10.1111/j.1365-2648.1991.tb03436.x

Neyman, A. (1985). Bounded complexity justifies cooperation in prisoners' dilemma. *Econ. Lett.* 18, 227–229. doi: 10.1016/0165-1765(85)90026-6

Oskamp, S., and Perlman, D. (1965). Factors affecting cooperation in a prisoner's dilemma game. *J. Conflict Resolut.* 9, 359–374. doi: 10.1177/002200276500900308

Padmala, S., and Pessoa, L. (2010). Interactions between cognition and motivation during response inhibition. *Neuropsychologia* 48, 558–565. doi: 10.1016/j.neuropsychologia.2009.10.017

Phelps, E. (2004). Human emotion and memory: interactions of the amygdala and hippocampal complex. *Curr. Opin. Neurobiol.* 14, 198–202. doi: 10.1016/j.conb.2004.03.015

Ramsøy, T. Z., Skov, M., Macoveanu, J., Siebner, H. R., and Fosgaard, T. R. (2015). Empathy as a neuropsychological heuristic in social decision-making. *Soc. Neurosci.* 10, 179–191. doi: 10.1080/17470919.2014.965341

Ray, R. D., Ochsner, K. N., Cooper, J. C., Robertson, E. R., Gabrieli, J. D., and Gross, J. J. (2005). Individual differences in trait rumination and the neural systems supporting cognitive reappraisal. *Cogn. Affect. Behav. Neurosci.* 5, 156–168. doi: 10.3758/cabn.5.2.156

Richey, J. A., Damiano, C. R., Sabatino, A., Rittenberg, A., Petty, C., Bizzell, J., et al. (2015). Neural Mechanisms of emotion regulation in autism spectrum disorder. *J. Autism Dev. Disord.* 45, 3409–3423. doi: 10.1007/s10803-015-2359-z

Rick, S., and Loewenstein, G. (2008). "The role of emotion in economic behavior", in *Handbook of Emotions*. eds. M. Lewis, J. M. Haviland-Jones, and L. F. Barrett (New York, NY: The Guilford Press).

Rilling, J. K., Goldsmith, D. R., Glenn, A. L., Jairam, M. R., Elfenbein, H. A., Dagenais, J. E., et al. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia* 46, 1256–1266. doi: 10.1016/j.neuropsychologia.2007.11.033

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The Neural basis of economic decision-making in the ultimatum game. *Science* 300, 1755–1758. doi: 10.1126/science.1082976

Sbarra, D. A., and Emery, R. E. (2005). The emotional sequelae of nonmarital relationship dissolution: analysis of change and intraindividual variability over time. *Pers. Relatsh.* 12, 213–232. doi: 10.1111/j.1350-4126.2005.00112.x

Scherer, K. R. (1982). Emotion as a process: function, origin and regulation. *Soc. Sci. Inf.* 21, 555–570. doi: 10.1177/053901882021004004

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Soc. Sci. Inf.* 44, 695–729. doi: 10.1177/0539018405058216

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., and Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron* 41, 653–662. doi: 10.1016/s0896-6273(04)00014-5

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439, 466–469. doi: 10.1038/nature04271

Smith, K. W., Balkwill, L. -L., Vartanian, O., and Goel, V. (2015). Syllogisms delivered in an angry voice lead to improved performance and engagement of a different neural system compared to neutral voice. *Front. Hum. Neurosci.* 9:273. doi: 10.3389/fnhum.2015.00273

Smith, K. W., Vartanian, O., and Goel, V. (2014). Dissociable neural systems underwrite logical reasoning in the context of induced emotions with positive and negative valence. *Front. Hum. Neurosci.* 8:736

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Stanley, D. A., Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., Phelps, E. A. (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

Vanderhasselt, M. -A., Kühn, S., and De Raedt, R. (2013). 'Put on your poker face': neural systems supporting the anticipation for expressive suppression and cognitive reappraisal. *Soc. Cogn. Affect. Neurosci.* 8, 903–910. doi: 10.1093/scan/nss090

Verduyn, P., Delvaux, E., Van Coillie, H., Tuerlinckx, F., and Van Mechelen, I. (2009). Predicting the duration of emotional experience: two experience sampling studies. *Emotion* 9, 83–91. doi: 10.1037/a0014610

Verney, S. P., Brown, G. G., Frank, L., and Paulus, M. P. (2003). Error-rate-related caudate and parietal cortex activation during decision making. *Neuroreport* 14, 923–928. doi: 10.1097/01.wnr.0000072842.93264.b6

Yamagishi, T., Kanazawa, S., Mashima, R., and Terai, S. (2005). Separating trust from cooperation in a dynamic relationship: prisoner's dilemma with variable dependence. *Ration. Soc.* 17, 275–308. doi: 10.1177/1043463105055463

Yarkoni, T. (2009). Big correlations in little studies: inflated fmri correlations reflect low statistical power-commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* 4, 294–298. doi: 10.1111/j.1745-6924.2009.01127.x

Zhang, Z., Luk, W., Arthur, D., and Wong, T. (2001). Nursing competencies: personal characteristics contributing to effective nursing performance. *J. Adv. Nurs.* 33, 467–474. doi: 10.1046/j.1365-2648.2001.01688.x

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

frontiers
in Human Neuroscience

ORIGINAL RESEARCH
published: 02 August 2019
doi: 10.3389/fnhum.2019.00257

Girls-Boys: An Investigation of Gender Differences in the Behavioral and Neural Mechanisms of Trust and Reciprocity in Adolescence

Imke L. J. Lemmers-Jansen[1], Anne-Kathrin J. Fett[1,2,3], Sukhi S. Shergill[3], Marlieke T. R. van Kesteren[4] and Lydia Krabbendam[1,3]*

[1] Department of Clinical, Neuro and Developmental Psychology, Faculty of Behavioral and Movement Sciences, Institute for Brain and Behavior, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [2] Department of Psychology, City, University of London, London, United Kingdom, [3] Department of Psychosis Studies, King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [4] Department of Education Sciences, Faculty of Behavioral and Movement Sciences, Institute for Brain and Behavior, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

OPEN ACCESS

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

Reviewed by:
Haruto Takagishi,
Tamagawa University, Japan
Shixuan Fu,
University of International Business and Economics, China

****Correspondence:***
Imke L. J. Lemmers-Jansen
imke.jansen@vu.nl

Received: *31 January 2019*
Accepted: *09 July 2019*
Published: *02 August 2019*

Citation:
Lemmers-Jansen ILJ, Fett A-KJ, Shergill SS, van Kesteren MTR and Krabbendam L (2019) Girls-Boys: An Investigation of Gender Differences in the Behavioral and Neural Mechanisms of Trust and Reciprocity in Adolescence.
Front. Hum. Neurosci. 13:257.
doi: 10.3389/fnhum.2019.00257

Background: Trust and reciprocity toward others have often been found to increase from childhood to adulthood. Gender differences in these social behaviors have been reported in adults. While adolescence is a key-period of change in social behavior, gender differences in trust and reciprocity during this developmental stage have rarely been investigated.

Methods: Here we investigate age-related gender differences in trust and reciprocity (n = 100, 51 female) and associated neural mechanisms (n = 44, 20 female) in adolescents between 13 and 19 years of age. Participants played two multi-round trust games with a pre-programmed cooperative and an unfair partner. Forty-four of 100 participants completed the trust game while undergoing functional brain imaging.

Results: Participants' investments were greater toward a cooperative than unfair game partner ($p < 0.01$), showing sensitivity to the degree of trustworthiness. There were no gender or age or related differences in baseline trust. In repeated cooperative interactions no gender differences were found, but younger adolescents showed slightly steeper increase of investments than older adolescents. In unfair interactions, younger males reacted with stronger decrease of investments than older males. Region of interest analysis of brain areas associated with in mentalizing, reward learning, conflict processing, and cognitive control revealed gender-by-age interactions on trusting behavior in the temporo-parietal junction (TPJ) and the caudate, showing stronger influence of age in males than in females during cooperation, and the reverse in unfair interactions. Additionally, main effects of gender were found in the TPJ, with higher activation in males, and in the caudate, with females showing greater activation.

Conclusion: In first interactions and during repeated cooperative interactions, adolescent males and females showed similar trusting behavior. Younger males showed stronger responses to unfairness by others. Gender-by-age interactions in specific ROIs suggest differential development in mentalizing and reward related cognitive processes. In conjunction with previous research, our findings suggest the presence of subtle gender and age-related changes in trust and cooperation that are only detectable using larger age windows.

Keywords: trust, reciprocity, gender, development, mentalizing, reward, fMRI

INTRODUCTION

Adolescence is a period of marked changes in social orientation, shifting from a family focus toward peer relations (Steinberg and Sheffield Morris, 2001; Brown, B.B., 2004; Nelson et al., 2005; Crone and Dahl, 2012). This development is supported by ongoing maturation of social (cognitive) skills. A crucial skill is the ability to trust and recognize trustworthiness in others. Trust is essential to initiate, establish, and maintain social relationships, by making relationships more cooperative and satisfactory, and strengthening norms that favor cooperation and/or increase group outcome (Balliet and Van Lange, 2013). Trust is associated with expectations, predictability, and confidence in others' behavior, with an emphasis on the benevolent motives of others in situations that involve a conflict between own interests and the interest of others (Balliet and Van Lange, 2013). The shift from a family focus toward peer relations in adolescence also encompasses a change from unconditional trust in close relatives to learning to trust people outside the family circle. Learning to trust others occurs in a process of repeated interactions that make it possible to build a mental model of the behavior of the other person. To initiate positive, cooperative interactions, trust in the positive reciprocity of the other is essential. For the maintenance of these interactions and for building social relationships, reciprocation of the initial trust is necessary (van den Bos et al., 2010). Initial distrust may be overcome by positive reciprocity, indicating that trust may grow in response to reciprocal behavior. Motivations to trust may vary (e.g., intrinsic, altruistic vs. extrinsic strategic), and both cognitive and affective processes play a role (Evans and Krueger, 2011; Balliet and Van Lange, 2013; Cutler and Campbell-Meiklejohn, 2019). In this study, trust is operationalized by means of the height of investments in the trust game (Berg et al., 1995).

In the trust game participants share a part of a given amount of money with an unknown person. The amount is tripled and the second person may return a certain amount to the investor, or keep it all. Trust in this paradigm is defined and operationalized as sending an endowment, so that the trustee can choose to honor trust, or not (Berg et al., 1995). The trust game allows to investigate baseline trust (i.e., the first investment given to an unknown person), as an index of a person's general inclination to trust. Additionally, in a multi-round trust game a context is created, in which trust can emerge as the outcome of a sustained social relationship (Cochard et al., 2004). In repeated interactions, the investor responds to the social feedback, adjusting the levels of trust accordingly (Tzieropoulos, 2013). Investigating trust in an experimental manner involves making commitments for real amounts of money, therefore resembling daily life situations more than questionnaires (Cochard et al., 2004). Experiments also allow for the systematic manipulation of context (response patterns of the trustee), yielding comparable data due to identical settings for all participants and added measures, such as neural data during task performance, acquired with functional Magnetic Resonance Imaging (fMRI).

Previous research yielded important insights into the development of trust and social mechanisms, such as reciprocity and cooperation (Eisenberg et al., 2002, 2005; Cochard et al., 2004; Steinberg, 2005; van den Bos et al., 2010, 2011, 2012; Smith et al., 2013; Fett et al., 2014b), and into gender differences in trust (Croson and Buchan, 1999; Balliet et al., 2011; Chaudhuri and Sbai, 2011; Chaudhuri et al., 2013; Van den Akker, 2018). People become more inclined to trust and to establish cooperation from childhood and early adolescence until middle adulthood (Sutter and Kocher, 2007; van den Bos et al., 2010, 2012; Evans et al., 2013; Fett et al., 2014a). Sutter and Kocher (2007) found that trust increases linearly (age 8–60^{+}) until 22 years of age, showing stability in adulthood and a slight decrease thereafter; Van den Bos and colleagues reported increasing trust from childhood to mid-adolescence and a slight decrease toward early adulthood (age 9–25) (van den Bos et al., 2010), as well as increased first investments and enhanced learning over trials with age (van den Bos et al., 2012). In very young children (age 4–5 and 9–10), trust was found to increase by 6-fold between kindergarten and elementary school, even when controlling for altruism (Evans et al., 2013). In contrast to the aforementioned studies, where different age groups were compared, research within a smaller age-range has shown a decrease of trust in adolescents aged 14–16.5 (Derks et al., 2014), or stable levels of trust between 12 and 18 years (van de Groep et al., 2018). These findings suggest that trust may develop until the early twenties, thereafter stabilizing or slightly decreasing, but the findings are contradictory about the exact time window of development.

Trust not only differs between developmental stages, but also between genders. During repeated interactions, males have been found to display more trust than females (Croson and Gneezy, 2009; Balliet et al., 2011). However, in negative, unfair interactions where trust is not reciprocated, females are more likely to stay trusting and to restore trust (Haselhuhn et al., 2015). Similarly, trust in unknown others differs between the genders, both in adolescents (Derks et al., 2014; van de Groep et al., 2018), and in adults (Buchan et al., 2008; Croson and Gneezy, 2009; Van den Akker, 2018), showing that men are more trusting than women. Only few studies have investigated gender differences and development of trust experimentally. In young children, age 4–5 girls trusted more often than boys, but a few years later (age 9–10), the reverse was found, resembling adult data (Evans et al., 2013). In a previous study, we have shown that during late-adolescence and early adulthood, males displayed higher baseline trust than females, and males reduced their trust more drastically with increasing age than females in interactions in which trust is not reciprocated (Lemmers-Jansen et al., 2017).

At the neural level, the motivation to cooperate is proposed to be modulated by the cognitive control system (centered on the dlPFC), regions of the social brain including the temporo-parietal junction (TPJ), the medial prefrontal cortex (mPFC), and the amygdala (Declerck et al., 2013), the anterior insula (Bellucci et al., 2016; Cutler and Campbell-Meiklejohn, 2019), and reward predicting areas, such as the caudate (Rilling et al., 2002; King-Casas et al., 2005; Tabibnia and Lieberman, 2007; Krill and Platek, 2012; Bellucci et al., 2016). Gender differences in neural activation during the trust game have shown increased activation of the TPJ in males compared to females, and increased activation of the caudate in females in a sample of late-adolescents and young adults (Lemmers-Jansen et al., 2017). Investigating trust in e-Bay offers in adults (30–35 years), females activated more striatal, whereas males activated more prefrontal areas (Riedl et al., 2010). Many of these regions are still developing during adolescence (Nelson et al., 2005; Blakemore, 2012; Crone and Dahl, 2012; Harenski et al., 2012). In the trust game, age-related increases of activation were found in the TPJ, posterior cingulate, right dorsolateral prefrontal cortex (dlPFC), right caudate, and precuneus (Fett et al., 2014b; Lemmers-Jansen et al., 2017). Age-related reductions in activation were also reported in the orbitofrontal cortex and caudate during interactions with a trustworthy, cooperative partner (Fett et al., 2014b), and in the anterior medial prefrontal cortex (amPFC) (van den Bos et al., 2011). In sum, previous findings suggest that differential neural activation patterns in brain areas involved in mentalizing, reward learning and cognitive control are associated with gender differences and age-related changes in trust and reciprocity toward others.

The Current Study

This study set out to investigate gender differences in the development and the underlying neural mechanisms of trust and reciprocity in adolescents (age 13–19). Participants played two repeated trust games, one with a cooperative partner, always returning the invested amount or more, and one with an unfair partner, who always returned less than invested. In our older adolescent-early adult sample, gender differences were present in baseline trust and males reacted with a steeper decline in investment to unfair treatment by the other than females. This effect became more pronounced with age (Lemmers-Jansen et al., 2017). However, overall, we found relatively stable patterns of trust, with neural activation that did not change with age (e.g., suggesting maturity). Possibly, changes in trust occur earlier in development. In an attempt to pinpoint the possible time window, the current study extends findings of our previous study to a younger sample of adolescents, who are in the middle of this process of social reorientation. Due to differential developmental speed, the development of trust and reciprocity may differ between boys and girls (Lenroot and Giedd, 2010; Blakemore, 2012; Crone and Dahl, 2012). Furthermore, social demands may differ between boys and girls, resulting in differential socialization processes, which lead to increasing gender differences in trust over time (Rose and Rudolph, 2006). In the current study we investigate differences in development of social behavior over repeated social interactions in an experimental setting, using a neuroeconomic trust game. Analogous to our previous study, we used two multi-round trust games, one with a pre-programmed cooperative and one with an unfair partner. Participants played the role of the investor and could make continuous investments. We investigated gender differences in baseline trust (i.e., first investments) and in the modulation of trust in response to reciprocated trust (i.e., cooperation) and in interactions where trust was not reciprocated (i.e., unfairness). Based on the previously discussed literature in adults and older adolescents, we hypothesized gender differences in baseline trust, with higher trust in males than in females. Additionally, we explored the association between age and first investment (i.e., baseline trust). Over a larger age range increases of baseline trust have been reported (Fett et al., 2014a), however, this was not found in adolescent samples (Derks et al., 2014; van de Groep et al., 2018). Furthermore, based on the literature and our previous study, we hypothesized that males and females would show similar investments during cooperative interactions, but that males would show more reduction of investments during unfair interactions than females. In addition, we expected that with age, trust would increase during cooperative interactions, and decrease during unfair interactions, and that gender differences would become more pronounced. At the neural level we tested gender differences and associations with age in nine predefined regions of interest (ROI), associated with mentalizing, reward, cognitive control, and conflict processing. Finally, we explored in the ROIs whether gender and age effects differed between cooperative and unfair interactions.

MATERIALS AND METHODS

Participants

Hundred healthy, right-handed adolescents, 51 female and 49 male, aged 13–19 (mean age = 16.5; *SD* = 1.57) participated in the behavioral part of this study. A subset of 24 males and 20 females also participated in fMRI. Part of the larger sample was previously described as the healthy comparison group for an early psychosis sample (Fett et al., 2016) and data of the males who took part in fMRI has previously been reported in a study that examined age effects in trust from adolescence to late adulthood (Fett et al., 2014a). For participant characteristics of this sub-sample, please see the **Supplementary Material** (**Table S1**). Participants were recruited at local schools in London, via colleagues and recruitment circulars at the Institute of Psychiatry, Psychology and Neuroscience. All participants had a good command of the English language. Participants had no history of neurological disorder, no psychiatric diagnosis, or psychotropic medication. Written informed consent was obtained from all participants and when under the age of 16 also from their parents/guardians. This study was approved by the research ethics committee London-Surrey Borders (10/H0806/38).

Measures

WASI Vocabulary Scale

The vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence (WASI) was used as indicator of general cognitive ability [13–18 years (Wechsler, 1999)], to investigate for possible confounding. T-scores were scaled for age.

Trust Game

Participants played the role of investor in two multi-round trust games. They were told that their two anonymous counterparts, the trustees, were connected to them via the Internet. In reality, they played against a computer, with two algorithms programmed to respond always in a cooperative and always in an unfair way. The algorithm was programmed in a probabilistic way: In the cooperative condition, with each increase in trust from the investor, the chance of a repayment of 200% increased with 10%. In the unfair condition, increases in trust from the investor increased the chance of a repayment of 50% (Gromann et al., 2013; Fett et al., 2014a, 2016). The two games were presented in counterbalanced order. Each game consisted of 20 experimental and 20 control trials. At the beginning of each experimental trial, participants started with £10. Any amount between £0 and £10 could be invested. The invested money was tripled and the trustee (i.e., computer) then made a repayment. Control trials were included as baseline condition for the fMRI analysis. The design and duration of the control trials were equal to the experimental trials, but without the element of investment. In the control trials participants had to move the cursor to a number between 0 and 10, which was indicated by a red arrow. Every trial started with an investment cue (2 s), followed by the investment period where participants made their choice (4 s, regardless of reaction times); the invested amount was shown (2 s), followed by a waiting period (jittered, 2–4 s), and a fixation cross (500 ms). Finally, the returned amount (3 s) and the final totals of both players (jittered, 2.5–4.5 s) were displayed, followed by a fixation cross (500 ms). Every trial lasted 18.5 s in total. For a graphical representation of the set-up of the trust game, see **Figure 1**. After the trust game, participants completed a short questionnaire that asked if at some point they had doubts that their counterpart was a real person (outcome represented in **Table 1**). About one third of the participants reported doubts. Therefore we report sensitivity analyses, comparing results of the participants with and without doubts. Additionally, all analyses were run including only participants without doubts that the trustee was real.

Procedure

After signing the consent form, participants were assessed with the WASI Vocabulary subtest. Other measures were administered, which are unrelated to the current topic. Before scanning participants completed 10 trust game practice rounds on a laptop. Participants were told that they were connected with their game partners via the Internet and that they would receive the earnings from one randomly selected round of the trust game. During scanning, two different runs of the trust game were administered, one with a cooperative and one with an unfair interaction partner, and structural scans were acquired. The complete scanning session lasted approximately one hour. After scanning the participants answered a short questionnaire, which examined their individual perceptions of the trust game and their game partners. Participants were given a fixed payment for participation, and for fairness reasons, all participants received £5 extra, as earnings from the trust game.

Data Analysis

Analyses of Behavioral Data

We analyzed the behavioral data using StataSE LAB 14 (StataCorp, 2015). We analyzed the effect of the condition on the amounts of the investments to check if the participants responded to the differences in response patterns of their interaction partners, with the investment as the dependent variable, using multilevel random regression analyses (XTREG), to account for multiple observations [investments (level 1); within participants (level 2)]. To test our hypotheses regarding changes of trust, we used the same multilevel regression analyses, including gender, age, and trial number, and their interaction as predictors. Trial number indicates the changes over time during the game, the development of trust in response to social feedback. The WASI score was added as covariate, to control for possible confounding of verbal cognitive ability. Analyses were run separately for the cooperative and unfair condition. Additionally, the effects of gender and age on first investment (e.g., baseline trust) were investigated. Results were considered significant when $p < 0.05$.

fMRI Image Acquisition and Analyses

Imaging data were acquired using a 3 Tesla GE Signa Neuro-optimized MR System. A quadrature birdcage head coil was used for radio frequency transmission and reception. For each game, 370 T2*-weighted whole-brain echo-planar images depicting the blood oxygen level-dependent (BOLD) contrast were acquired with the following parameters: slice thickness = 2.4 mm; inter-slice gap = 1 mm; TR = 2000 ms; TE = 25 ms; flip angle = 75°; in-plane voxel dimension = 3.4 mm; number of slices = 38; dummy acquisitions = 4 and matrix = 64 × 64. For anatomical reference, a whole-brain high-resolution gradient-echo image of 43 slices was acquired with the following parameters: slice thickness = 3 mm; inter-slice gap = 0.3 mm; TR = 3000 ms; TE = 30 ms; flip angle = 90°; in-plane voxel size = 1.9 mm and matrix = 128 × 128. Participants were placed head first in the scanner. Foam padding was placed around the head in the coil to minimize head movement and the participants were provided with ear protectors. The participants looked at the screen through a mirror. Participants were equipped with a button box in their right hand. One button was used to increase the investment, one to decrease the investment.

Data were analyzed with SPM12[1]. All images were corrected for head-motion using iterative rigid body realignment with six motion-parameters to minimize the residual sum of squares between the images. The functional images of each subject were co-registered to that subject's structural scan. The functional images were spatially normalized ("old normalized") using

[1] http://www.fil.ion.ucl.ac.uk/spm/

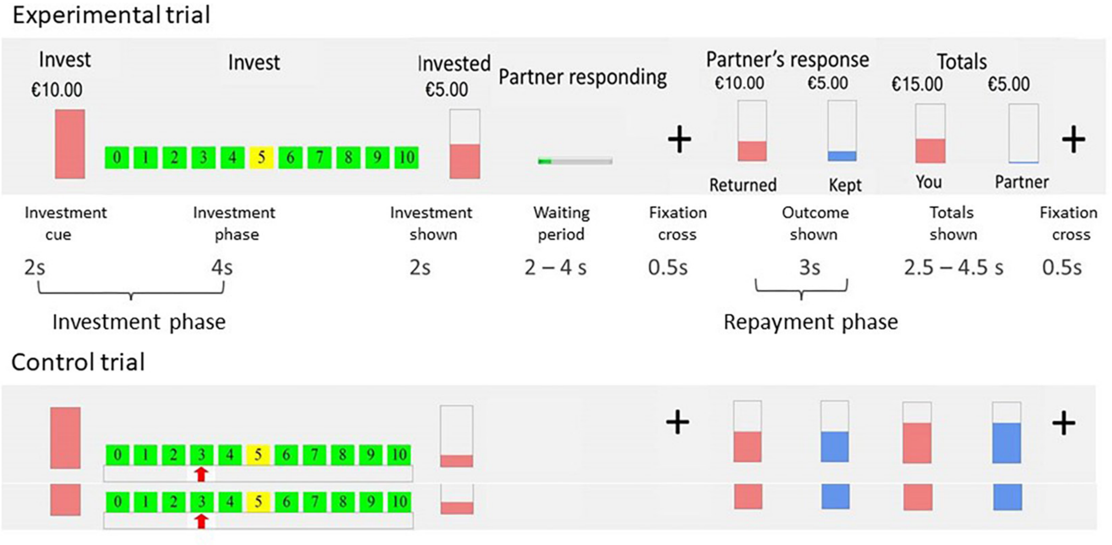


FIGURE 1 | Graphical overview of the trust game. *Note:* Top row represents the visual stimuli in the game trials; middle row are the separate phases including durations of the trust game; bottom row represents the visual stimuli in the control trials. Taken with permission from Lemmers-Jansen et al. (2017).

the Montreal Neurological Institute (MNI) 152 T1 template (voxel size = 3.5 × 3.5 × 3.5), and spatially smoothed using an 8-mm full-width, half-maximum Gaussian kernel, to allow for group-analyses. Per subject 370 scans were acquired per condition.

At first-level, fMRI time-series data were modeled by a series of events convolved with a canonical hemodynamic response function (HRF). The investment phase was modeled as an event lasting from the start of the investment phase until the moment the participant pressed the button to make the investment, or to choose the indicated number in the control condition (mean reaction time 3.7 s, *SD* = 0.93 s). The repayment phase was the period during which the response of the trustee was shown, lasting 3 s (see **Figure 1**). Game trials were contrasted with

TABLE 1 | Participant characteristics, trust game behavior and beliefs.

	Male *N* = 49	Female *N* = 51	Statistics		Overall *N* = 100
Measures	**Mean (*SD*)**	**Mean (*SD*)**	***Beta***	***p***	**Mean (*SD*)**
Age	16.35 (1.65)	16.58 (1.5)	0.08	0.45	16.47 (1.57)
WASI t score	55.31 (11.89)	49.55 (9.25)*	0.26	0.008	52.37 (11.02)
First investment, baseline trust	6.29 (2.10)	5.61 (2.30)	0.12	0.24	5.94 (2.22)
Mean investment Cooperative partner	6.84 (2.89)	6.33 (2.76)	0.02	0.35	6.58 (2.83)
Mean investment Unfair partner	3.73 (3.25)	4.08 (2.88)	−0.04	0.13	3.91 (3.07)
	***N* (%)**	***N* (%)**	χ^2	***p***	***N* (%)**
After trust game questionnaire#					
Manipulation doubt?	15 (32%)	13 (33%)	0.02	0.89	28 (32%)
Strategy:			2.52	0.64	
-responding to partner	21 (47%)	13 (31%)			34 (40%)
- maximize profit	9 (20%)	8 (20%)			17 (20%)
- no strategy	11 (24%)	15 (37%)			26 (30%)
- other	4 (9%)	5 (12%)			9 (10%)

**Significant difference at p < 0.01. #data of nine participants missing in the manipulation questionnaire, and 14 missing for the strategy questions. Note: WASI vocabulary = Wechsler Abbreviated Scale of Intelligence, vocabulary subscale, scaled for age. After the trust game participants were asked if at any time they had doubts whether their counterpart was real. If they responded personalizing (saying "he"), it was coded as believing the counterpart was real. If in two conditions they reported probabilistic answers, predictable or unreal, then it was coded as having doubts. Then participants were asked about the strategy used for investment. If the answer included the behavior of the counterpart, it was coded as "responding to partner"; if the answer contained "great," "maximum," "profit," or "more than the other" it was coded as "maximize profit"; some participants answered the did not use a strategy; and other comprises "random," "gambling," or always trying the same amount.*

the corresponding phases of the control trials. Six movement parameters were included in the model.

Analogous to our previous study (Lemmers-Jansen et al., 2017), ROI analyses were conducted on the right TPJ (MNI coordinates: 45, −43, 32), right dlPFC (51, 18, 30), right insula (36, 24, 0) and the ACC (−3, 27, 33), complemented with the left TPJ (−44, −46, 29), ventral striatum (VS; 14, 12, −5), amPFC (0, 42, 6), and bilateral caudate ROI's (right: 6, 11, 5; left: −7, 12, −4). ROIs were defined as a 10 mm sphere around the given coordinates, except for the caudate, where a 5 mm sphere was used. Analyses were conducted in SPM12, using Marsbar-0.44[2] to generate the ROIs. We used an event related, factorial design with gender as contrast and age as covariate. All ROI analyses were conducted separately for the investment and repayment phase, in the cooperative and unfair conditions.

Additionally, exploratory whole-brain analyses were performed to examine group wise differences in regions outside the a priori defined ROIs. The results are presented in the **Supplementary Material** (**Supplementary Tables S2–S4**).

RESULTS

Participant Characteristics

Participant characteristics are described in **Table 1**. There were no group differences between males and females in age. However, WASI vocabulary scores differed significantly between males and females, with males scoring on average 6 points higher than females. There was no significant correlation between WASI scores and investment, suggesting that any gender differences in investment were unlikely influenced by systematic differences in general cognitive ability. One third of the participants indicated doubts in response to the question if they believed they were interacting with a real partner. However, no differences in investments and ROI activation were found between the participants with and without doubts ($p > 0.7$ and $p > 0.4$, respectively), and analyses without those who had doubts that the trustee was real yielded similar outcomes as the results presented below. Several strategies were used during investments (see **Table 1**), but these did not differ significantly between genders.

Behavioral Results

The investments in the trust game are shown in **Table 1**. The effect of condition on investment was investigated as a manipulation check. Results showed significant differences between conditions (see **Figure 2**), indicating that the task conditions (cooperative vs. unfair) worked as intended ($b = 2.72$, $p < 0.001$, 95%CI = −2.89/−2.55).

For baseline trust there were no gender-by-age interaction ($\beta = -0.23$, $p = 0.83$) or significant main effects of gender or age ($\beta = 0.13$, $p = 0.22$ and $\beta = 0.08$, $p = 0.42$, respectively).

In the *cooperative condition*, no gender-by-age-by trial number interaction was found. After the three-way interaction was removed from the model, a gender-by-age interaction at trend level ($b = -0.40$, $p = 0.09$, 95%CI = −0.85/0.06) was observed. There was a significant age-by-trial number interaction ($b = -0.11$, $p = 0.032$, 95%CI = −0.02/−0.001), showing that younger participants increased their investments more than older participants (younger: $b = 0.05$, $p < 0.001$, 95%CI = 0.32/0.08; older: $b = 0.04$, $p < 0.001$, 95%CI = 0.01/0.06), based on a median age split (age 16.9; see **Figure 3**).

In the *unfair condition*, there was a significant gender-by-age-by-trial number interaction ($b = 0.03$, $p < 0.03$, 95%CI = 0.003/0.05). Analyses by gender showed a significant interaction between age and trial number on investment in males ($b = 0.02$, $p < 0.05$, 95%CI = 0.001/0.04), but not females ($b = -0.01$, $p = 0.32$, 95%CI = −0.03/0.01). *Post hoc* analyses with a median split for age showed that younger males decreased their investments more strongly toward the unfair other than older males (see **Figure 4**). In females there was a significant main effect of age ($b = 0.34$ $p < 0.05$, 95%CI = 0.02/0.66), showing that younger females invested less in the unfair partner than older females (see **Figure 4**), but there was no significant main effect of trial number.

fMRI ROI Results

Cooperative Interactions

ROI analysis revealed gender-by-age interactions in the cooperative investment phase, in the left TPJ and the right caudate (see **Figure 5**). During the cooperative repayment phase, a gender-by-age interaction was found with a significance level just bordering the threshold adjusted for multiple comparisons in the right TPJ (see **Figure 5**). All areas showed greater increase of activation with age in males compared to females. Main effects of gender, bordering significance, became apparent in the cooperative repayment phase (see **Table 2**), with males activating the TPJ more, and females activating the caudate more. There was no main effect of age.

Unfair Interactions

During the repayment phase, a gender-by-age interaction was found in the left TPJ, with greater increase of activation with age in females compared to males (see **Figure 5**). There were no significant main effects of gender. In the ACC and dlPFC, a non-significant trend-level effect of age was found, showing increased activation in older participants during investments.

DISCUSSION

This study set out to investigate the development of trust in adolescent boys and girls. Using two multi-round trust games, we found gender-by-age interactions on investment behavior during unfair interactions, with younger males reacting more strongly to unfair partner feedback. During cooperative interactions there was a significant age-by-trial number interaction, showing that younger participants increased their investments slightly more than older participants. At the neural level, significant gender-by-age interactions and main effects of gender bordering significance were found in the TPJ and caudate, suggesting differential cognitive mechanisms underlying trust between genders that

[2] http://marsbar.sourceforge.net/

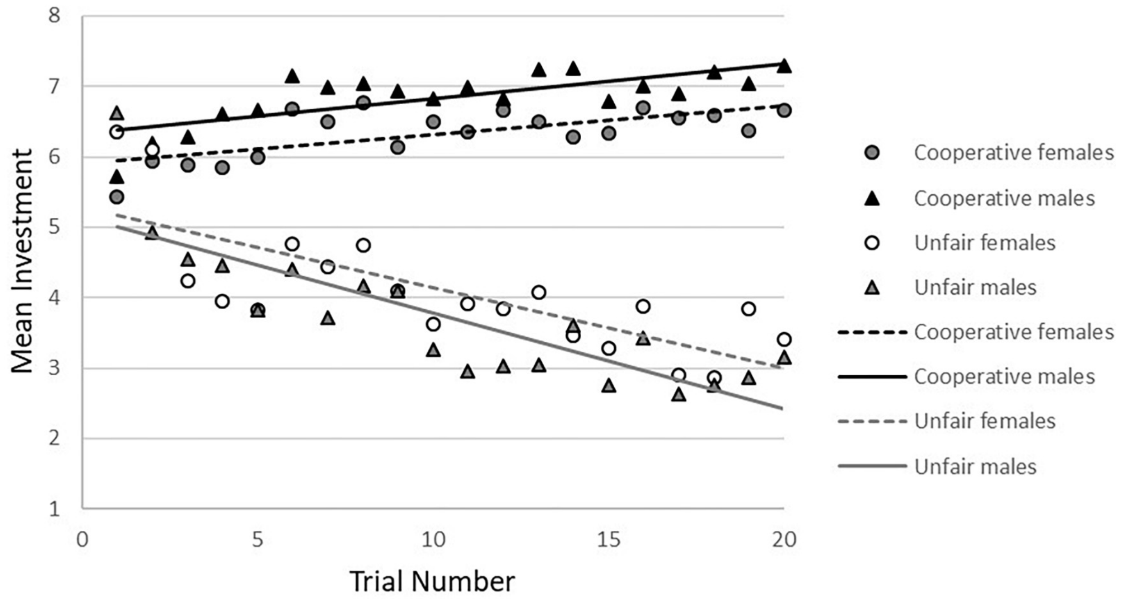


FIGURE 2 | Mean investment over trials by gender and condition of the trust game.

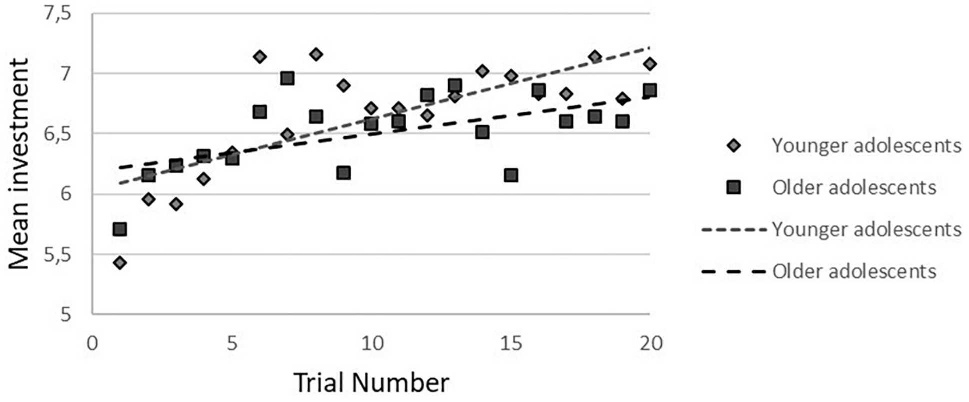


FIGURE 3 | Age-by-trial number interaction in younger and older adolescents during cooperation. To visualize the effect, a median split for age was performed.

change during this phase of development. Age-related increases of activation in cognitive control areas were found at trend level and only in unfair interactions.

Behavioral Findings

Baseline Trust

Contrary to our hypothesis and previous results, baseline trust did not differ significantly between genders in this adolescent sample. Adult males tend to trust more than females (Sutter and Kocher, 2007; Buchan et al., 2008; Croson and Gneezy, 2009; Van den Akker, 2018). This pattern was also found in our older adolescent sample (Lemmers-Jansen et al., 2017), and in a mid-adolescence sample (14–16.5 years) using a repeated one-shot trust game (Derks et al., 2014). These findings are contradictory, especially with Derks et al. (2014). This could be due to differences in the experimental set-up and needs to be investigated further.

No age-related changes in baseline trust were found, suggesting that baseline trust does not increase substantially from early to late adolescence. Possibly, age-related changes in baseline trust during adolescence are small, with variability throughout this phase of development, and thus are only detectable when looking at a larger time window [see also van den Bos et al. (2010, 2011, 2012); Fett et al. (2014a)].

Repeated Interactions

Changes in trust in response to cooperative feedback only showed a trend-level gender-by-age interaction, and no main effects of gender. Younger adolescents, however, showed a steeper increase of investments than older adolescents. The finding of absent

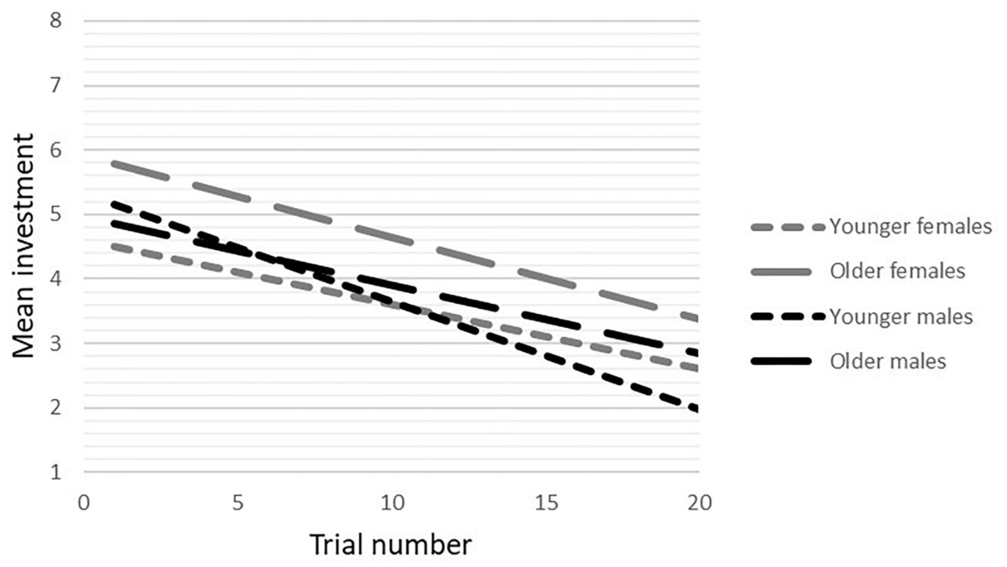


FIGURE 4 | Gender-by-age interaction on investments over trials in the unfair condition.

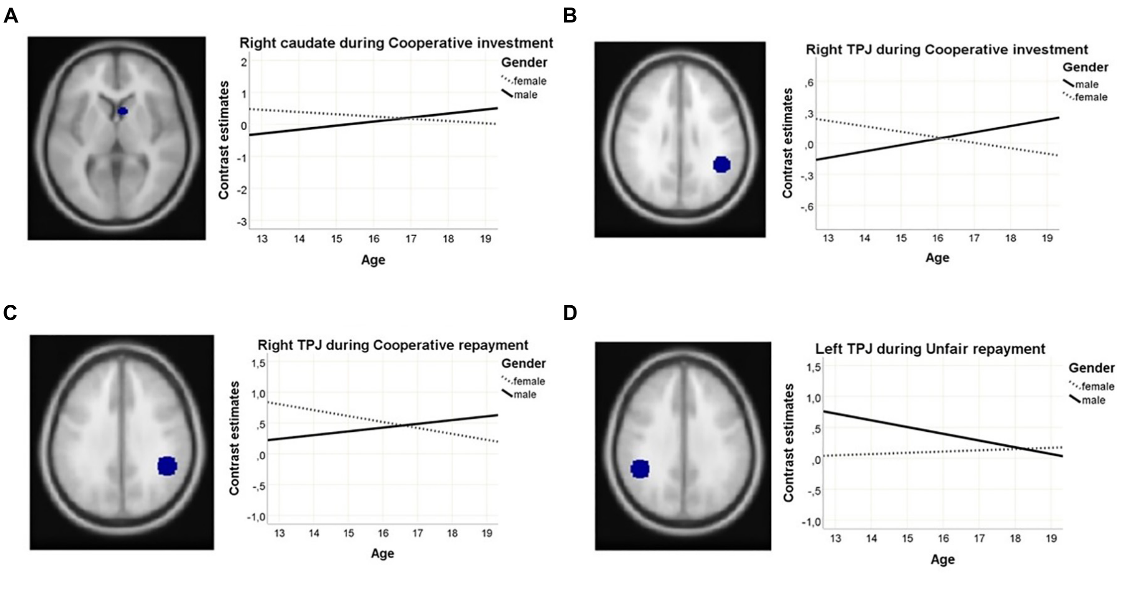


FIGURE 5 | Gender-by-age interactions in ROI activation showing **(A)** the right caudate during cooperative investments; **(B)** the left TPJ during cooperative investments; **(C)** the right TPJ during cooperative repayments; and **(D)** the left TPJ during unfair repayments.

gender effects during adolescence are in line with our previous study in slightly older adolescents and young adults (Lemmers-Jansen et al., 2017). Gender differences in repeated trust games have rarely been investigated. In males it has been found that investments increase with age from adolescence to mid adulthood (Fett et al., 2014a), thus it might be likely that gender differences also emerge later during development when gender roles become more established or specific cognitive abilities more refined. The current results do not support earlier work by van den Bos et al. (2010, 2012), who found age-related increases in reciprocity during development (age 9–25, and mean age 11, 16, and 19, respectively), using a two-choice trust game. It is possible that the development of trust and reciprocity follows different developmental trajectories.

During unfair interactions a gender-by-age-by-trial number interaction on levels of trust was found. The direction of the interaction, however, did not correspond with our hypothesis. All age groups adjusted levels of trust in response to unfair feedback, reflected in lower investments over time. Overall, younger individuals showed lower trust. Contrary to our expectation, younger males showed a steeper decline of investment than older males. This result contradicts our previous findings, where

TABLE 2 | ROI analyses outcome, by condition of the trust game.

Condition Association	ROI	*p*	*t*
Cooperative investment*			
Interaction age and gender:			
Age males > age females	Left TPJ	0.019	2.14
	Right caudate	0.015	2.26
Cooperative repayment**			
Interaction age and gender:			
Age males > age females	Right TPJ	0.037#	1.84
Main effect of gender:			
Males > females	Left TPJ	0.036#	1.85
Females > males	Left caudate	0.034#	1.87
Unfair investment***			
Increasing with age	ACC	0.003	3.13
	dlPFC	0.04#	1.79
Unfair repayment**			
Interaction age and gender: Age females > age males	Left TPJ	0.031	1.93

*Note: All ROIs were defined as a 10 mm sphere (except right caudate: 5 mm) around the following MNI coordinates: right temporo-parietal junction (TPJ): 45, −43, 32; left temporo-parietal junction: −44, −46, 29; right caudate: 6, 11, 5; left caudate −7, 12, −4; dorso-lateral prefrontal cortex (dlPFC): 51, 18, 30; anterior cingulate cortex (ACC): −3, 27, 33. *adjusted threshold for the cooperative investment phase: p = 0.037. **adjusted threshold for the cooperative and unfair repayment phase: p = 0.032. ***adjusted threshold for the unfair investment phase: p = 0.036. #bordering significance of the adjusted p-value.*

similar behavior was found in older, and not in younger males (Lemmers-Jansen et al., 2017). Younger females made lower investments than older females.

The current results suggest that the response to cooperation in females develops in a linear way, and that the development of trust in males might level off towards adulthood. However, responses to breaches of trust in males showed a different development, with a less drastic response to unfair partner behavior during later adolescence. Females seem to follow a more linear pathway in the development of trust in response to unfair feedback, with slightly higher investments in older females. Lower investments might reflect greater weariness of the unfair partner or less attempts to establish cooperation.

Neural Results

In contrast to the behavioral findings in the cooperative condition that showed similar levels of trust in males and females, at the neural level several gender-by-age interactions were found in the TPJ and caudate. These areas have been consistently found in the trust game, and have been linked to the mentalizing and reward learning components of the trust game, respectively (King-Casas et al., 2005; Lee, 2008; van den Bos et al., 2011; Lemmers-Jansen et al., 2017). In the investment phase, the activation in females in the left TPJ and right caudate decreased with age, whereas in males activation increased with age. The same pattern, however, at trend level, was found in the right TPJ during repayments. In the trust game and other social cognitive tasks, gender differences in TPJ activation have been reported in young adults, with higher activation in males compared to females (Schulte-Rüther et al., 2008; Luo et al., 2015; Lemmers-Jansen et al., 2017), but the reverse has also been reported (Chan, 2016; Zhang et al., 2017). Gender differences in activation of the caudate in response to emotional stimuli have been reported, showing higher activation in females [for a meta-analysis, see Stevens and Hamann (2012)]. In absence of behavioral differences, these results could suggest that males and females have different motivations or strategies for the same behavior, or adopt different cognitive strategies in response to processing social feedback (Cahill, 2006). These gender differences in strategies or motivations change with age. Apart from different strategies and motivations, these gender-by-age interactions may also point toward gender differentiated development in the given areas, which were not observed in the older sample (Lemmers-Jansen et al., 2017).

During unfair interactions, a gender-by-age interaction was observed in the right TPJ, with females showing slightly increasing activation with age, and males showing reduced activation with increasing age. In combination with the behavioral findings, this suggests that younger males respond stronger to negative feedback than older males, indicating increased efforts to mentalize about the other's behavior and a stronger tendency to retaliate untrustworthy behavior.

Only under unfair treatment by the other player, age-related changes in neural activity became apparent, in the ACC and at trend level in the dlPFC. The age-related changes in the ACC during unfair interactions are in line with findings in an overlapping sample (Fett et al., 2014a), which included a much larger age range. Increasing activation with age in ACC and dlPFC have been also reported by van den Bos et al. (2011), however, in decisions to trust compared to no trust decisions. These regions are associated with conflict processing and the cognitive control network (MacDonald et al., 2000; Botvinick et al., 2004; Pochon et al., 2008), which is still developing during adolescence (Fair et al., 2007; Kelly et al., 2009; Crone and Dahl, 2012; Steinbeis et al., 2014; Crone and Steinbeis, 2017). With increasing age, increasing activation of the dlPFC was found in the unfair condition, suggesting that cognitive control areas are more engaged during decisions to (dis)trust across adolescent development. The current findings are also in line with Van

Duijvenvoorde et al. (2008) and Van Den Bos et al. (2009). Using non-social rule selection and probabilistic learning tasks, based on learning from positive and negative feedback in a similar developmental sample, they found that both ACC and dlPFC were activated more with increasing age during negative feedback processing.

Limitations and Future Directions

The current findings need to be interpreted in the light of several limitations. Firstly, developmental changes in trust seem to be subtle. In order to investigate gender specific development, a larger age-range might have to be included as changes may be most obvious during transition to adulthood (Fett et al., 2014a). Secondly, participants played against a computerized algorithm, rather than a human counterpart. One third of the participants said that they doubted that the other person was real. Analyses comparing the behavior of the individuals who expressed or did not express doubts did not yield significant differences in terms of investments. In addition, higher investments in the cooperative condition and lower investments in the unfair condition showed that overall the experimental manipulation of the counterpart was effective. Moreover, we informed participants that they were paid upon performance in the trust game, aiming to increase task engagement. Additionally, despite the advantages of experimental investigations, they face the problem of generalizability to other, real world settings. The findings therefore should be considered with caution, when making generalizations to other contexts.

Different motives may underlie trust game behavior. For example, a reduced adjustment to unfair behavior of the partner may be associated with perspective-taking ability (Fett et al., 2014b), but also with an inclination to restore trust. Future studies may shed further light on underlying motives by including detailed experimental and questionnaire measures such as social value orientation (Derks et al., 2014, 2015), and Machiavellianism (Bora et al., 2009; Čavojová et al., 2011), or by specific experimental manipulations of the game. Underlying mechanisms and motivations may be revealed with fMRI (i.e., through activation of particular areas that have typically been associated with particular functioning by other studies), however, these interpretations rely on reverse inference (Poldrack, 2006; Poldrack et al., 2016). No firm conclusions can be drawn from these data, but they provide a starting point for generating new hypotheses. These hypotheses in turn warrant further investigation and testing in future research.

In summary, we set out to investigate the neural mechanisms underlying gender and age effects on social interactions using a trust game during functional MRI. Results showed that there were no gender and age differences in baseline trust, and age differences in the increase of investments over trials during cooperative interactions, with younger adolescents showing a slightly steeper increase over repeated interaction. The findings suggest relatively stable processes of trust and cooperation between 13 and 19 years of age. During unfair interactions, younger males showed stronger sensitivity to unfairness, suggested by a stronger increase in distrust than older males. In females, age was associated with higher overall investments. The current study suggests that younger adolescents are more sensitive to their partner's trustworthiness. Differential patterns of neural activation may suggest different cognitive strategies underlying similar behavior in males and females. Specifically, mentalizing and reward-related areas were differentially activated in males and females, and also showed different age-specific trends. Future studies need to investigate these mechanisms further.

ETHICS STATEMENT

This study was approved by the local research ethics committee [London-Surrey Borders (10/H0806/38)].

AUTHOR CONTRIBUTIONS

LK and A-KF designed the experiments. SS and LK supervised the study. A-KF collected the data. MvK and IL-J analyzed the data. IL-J, A-KF, MvK, SS, and LK interpreted the results. IL-J wrote the manuscript. All the authors discussed the results, reviewed and contributed to the critical development of the manuscript, and approved the final manuscript.

FUNDING

This work was supported by funding of the Hersenstichting Nederland [KS2011(1)-75 to LK]; the Netherland Organization for Scientific Research (NWO) [#451-13-035 to A-KF and #452-07-007 and #453-11-005 to LK]; a NARSAD Young Investigator Award from the Brain and Behavior Foundation [#24138 to A-KF]; ERC Consolidator Grants #648082 to LK and #311686 to SS]. SS is also supported by the NIHR Mental Health Biomedical Research Center at the SLaM NHS Trust and King's College London. MvK is supported by a Marie Curie Individual Fellowship of the EU Horizon2020 Framework Program for Research and Innovation [#704506].

ACKNOWLEDGMENTS

We thank the MHR support team and clinical studies officers, Farah Yakub, Mirjam van Driel, and Max van den Broek for their support with the project and the local schools and pupils for their participation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2019.00257/full#supplementary-material

REFERENCES

Balliet, D., Li, N. P., Macfarlan, S. J., and Van Vugt, M. (2011). Sex differences in cooperation: a meta-analytic review of social dilemmas. *Psychol. Bull.* 137, 881–909. doi: 10.1037/a0025354

Balliet, D., and Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. *Psychol. Bull.* 139, 1090–1112. doi: 10.1037/a0030939

Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., and Krueger, F. (2016). Neural signatures of trust in reciprocity: a coordinate-based meta-analysis. *Hum. Brain Mapp.* 38, 1233–1248. doi: 10.1002/hbm.23451

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

Blakemore, S.-J. (2012). Imaging brain development: the adolescent brain. *Neuroimage* 61, 397–406. doi: 10.1016/j.neuroimage.2011.11.080

Bora, E., Yucel, M., and Allen, N. B. (2009). Neurobiology of human affiliative behaviour: implications for psychiatric disorders. *Curr. Opin. Psychiatr*. 22, 320–325. doi: 10.1097/YCO.0b013e328329e970

Botvinick, M. M., Cohen, J. D., and Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* 8, 539–546. doi: 10.1016/j.tics.2004.10.003

Brown, B.B. (2004). "Adolescents' relationships with peers, ". in eds R. M. Lerner and L. Steinberg,*Handbook of Adolescent Psychology*. 363–394. Hoboken, NJ: John Wiley. doi: 10.1002/9780471726746.ch12

Buchan, N. R., Croson, R. T., and Solnick, S. (2008). Trust and gender: an examination of behavior and beliefs in the investment game. *J. Econ. Behav. Organ.* 68, 466–476. doi: 10.1016/j.jebo.2007.10.006

Cahill, L. (2006). Why sex matters for neuroscience. *Nat. Rev. Neurosci.* 7, 477–484. doi: 10.1038/nrn1909

Čavojová, V., Belovičová, Z., and Sirota, M. (2011). Mindreading and empathy as predictors of prosocial behavior. *Studia Psychol.* 53, 351–362.

Chan, Y.-C. (2016). Neural correlates of sex/gender differences in humor processing for different joke types. *Front. psychol.* 7:536. doi: 10.3389/fpsyg.2016.00536

Chaudhuri, A., Paichayontvijit, T., and Shen, L. (2013). Gender differences in trust and trustworthiness: individuals, single sex and mixed sex groups. *J. Econ. Psychol.* 34, 181–194. doi: 10.1016/j.joep.2012.09.013

Chaudhuri, A., and Sbai, E. (2011). Gender differences in trust and reciprocity in repeated gift exchange games. *N.Z Econ. Pap.* 45, 81–95. doi: 10.1080/00779954.2011.556072

Cochard, F., Van, P. N., and Willinger, M. (2004). Trusting behavior in a repeated investment game. *J. Econ. Behav. Organ.* 55, 31–44. doi: 10.1016/j.jebo.2003.07.004

Crone, E. A., and Dahl, R. E. (2012). Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nat. Rev. Neurosci.* 13, 636–650. doi: 10.1038/nrn3313

Crone, E. A., and Steinbeis, N. (2017). Neural perspectives on cognitive control development during childhood and adolescence. *Trends Cogn. Sci.* 21, 205–215. doi: 10.1016/j.tics.2017.01.003

Croson, R., and Buchan, N. (1999). Gender and culture: international experimental evidence from trust games. *Am. Econ. Rev.* 89, 386–391. doi: 10.1257/aer.89.2.386

Croson, R., and Gneezy, U. (2009). Gender differences in preferences. *J. Econ. Lit.* 47, 448–474.

Cutler, J., and Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *Neuroimage* 184, 227–241. doi: 10.1016/j.neuroimage.2018.09.009

Declerck, C. H., Boone, C., and Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain Cogn.* 81, 95–117. doi: 10.1016/j.bandc.2012.09.009

Derks, J., Lee, N. C., and Krabbendam, L. (2014). Adolescent trust and trustworthiness: role of gender and social value orientation. *J. Adolesc.* 37, 1379–1386. doi: 10.1016/j.adolescence.2014.09.014

Derks, J., van Scheppingen, M. A., Lee, N. C., and Krabbendam, L. (2015). Trust and mindreading in adolescents: the moderating role of social value orientation. *Front. psychol.* 6:965. doi: 10.3389/fpsyg.2015.00965

Eisenberg, N., Cumberland, A., Guthrie, I. K., Murphy, B. C., and Shepard, S. A. (2005). Age changes in prosocial responding and moral reasoning in adolescence and early adulthood. *J. Res. Adolesc.* 15, 235–260. doi: 10.1111/j.1532-7795.2005.00095.x

Eisenberg, N., Guthrie, I. K., Cumberland, A., Murphy, B. C., Shepard, S. A., and Zhou, Q. (2002). Prosocial development in early adulthood: a longitudinal study. *J. Personal. Soc. Psychol.* 82, 993–1006.

Evans, A. M., Athenstaedt, U., and Krueger, J. I. (2013). The development of trust and altruism during childhood. *J. Econ. Psychol.* 36, 82–95. doi: 10.1002/14651858.CD010414.pub2

Evans, A. M., and Krueger, J. I. (2011). Elements of trust: risk and perspective-taking. *J. Exp. Soc. Psychol.* 47, 171–177. doi: 10.1093/scan/nsp009

Fair, D. A., Dosenbach, N. U., Church, J. A., Cohen, A. L., Brahmbhatt, S., Miezin, F. M., et al. (2007). Development of distinct control networks through segregation and integration. *Proc. Natl. Acad. Sci.* 104, 13507–13512. doi: 10.1073/pnas.0705843104

Fett, A.-K., Gromann, P., Giampietro, V., Shergill, S., and Krabbendam, L. (2014a). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Soc. Cogn. Affect. Neurosci.* 9, 395–402. doi: 10.1093/scan/nss144

Fett, A.-K., Shergill, S., Gromann, P., Dumontheil, I., Blakemore, S.-J., Yakub, F., et al. (2014b). Trust and social reciprocity in adolescence–a matter of perspective-taking. *J. Adolesc.* 37, 175–184. doi: 10.1016/j.adolescence.2013.11.011

Fett, A.-K., Shergill, S., Korver-Nieberg, N., Yakub, F., Gromann, P., and Krabbendam, L. (2016). Learning to trust: trust and attachment in early psychosis. *Psychol. Med.* 46, 1437–1447. doi: 10.1017/S0033291716000015

Gromann, P., Heslenfeld, D., Fett, A.-K., Joyce, D., Shergill, S., and Krabbendam, L. (2013). Trust versus paranoia: abnormal response to social reward in psychotic illness. *Brain* 136(Pt 6), 1968–1975. doi: 10.1093/brain/awt076

Harenski, C. L., Harenski, K. A., Shane, M. S., and Kiehl, K. A. (2012). Neural development of mentalizing in moral judgment from adolescence to adulthood. *Devel. Cogn. Neurosci.* 2, 162–173. doi: 10.1016/j.dcn.2011.09.002

Haselhuhn, M. P., Kennedy, J. A., Kray, L. J., Van Zant, A. B., and Schweitzer, M. E. (2015). Gender differences in trust dynamics: women trust more than men following a trust violation. *J. Exp. Soc. Psychol.* 56, 104–109. doi: 10.1016/j.jesp.2014.09.007

Kelly, A. C., Di Martino, A., Uddin, L. Q., Shehzad, Z., Gee, D. G., Reiss, P. T., et al. (2009). Development of anterior cingulate functional connectivity from late childhood to early adulthood. *Cereb. Cortex* 19, 640–657. doi: 10.1093/cercor/bhn117

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

Krill, A. L., and Platek, S. M. (2012). Working together may be better: activation of reward centers during a cooperative maze task. *PloS One* 7:e30613. doi: 10.1371/journal.pone.0030613

Lee, D. (2008). Game theory and neural basis of social decision making. *Nat. Neurosci.* 11, 404–409. doi: 10.1038/nn2065

Lemmers-Jansen, I. L., Krabbendam, L., Veltman, D. J., and Fett, A.-K. J. (2017). Boys vs. girls: gender differences in the neural development of trust and reciprocity depend on social context. *Devel. Cogn. Neurosci.* 25, 235–245. doi: 10.1016/j.dcn.2017.02.001

Lenroot, R. K., and Giedd, J. N. (2010). Sex differences in the adolescent brain. *Brain Cogn.* 72, 46–55. doi: 10.1016/j.bandc.2009.10.008

Luo, P., Wang, J., Jin, Y., Huang, S., Xie, M., Deng, L., et al. (2015). Gender differences in affective sharing and self–other distinction during empathic neural responses to others' sadness. *Brain Imaging Behav.* 9, 312–322. doi: 10.1007/s11682-014-9308-x

MacDonald, A. W., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288, 1835–1838. doi: 10.1126/science.288.5472.1835

Nelson, E. E., Leibenluft, E., McClure, E. B., and Pine, D. S. (2005). The social re-orientation of adolescence: a neuroscience perspective on the process and its relation to psychopathology. *Psychol. Med.* 35, 163–174. doi: 10.1017/s0033291704003915

Pochon, J.-B., Riis, J., Sanfey, A. G., Nystrom, L. E., and Cohen, J. D. (2008). Functional imaging of decision conflict. *J. Neurosci.* 28, 3468–3473. doi: 10.1523/JNEUROSCI.4195-07.2008

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63. doi: 10.1016/j.tics.2005.12.004

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K., Matthews, P. M., Munafo, M., et al. (2016). Scanning the Horizon: future challenges for neuroimaging research. *bioRxiv* 059188.

Riedl, R., Hubert, M., and Kenning, P. (2010). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Q.* 34, 397–428.

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

Rose, A. J., and Rudolph, K. D. (2006). A review of sex differences in peer relationship processes: potential trade-offs for the emotional and behavioral development of girls and boys. *Psychol. Bull.* 132, 98–131. doi: 10.1037/0033-2909.132.1.98

Schulte-Rüther, M., Markowitsch, H. J., Shah, N. J., Fink, G. R., and Piefke, M. (2008). Gender differences in brain networks supporting empathy. *Neuroimage* 42, 393–403. doi: 10.1016/j.neuroimage.2008.04.180

Smith, A. R., Chein, J., and Steinberg, L. (2013). Impact of socio-emotional context, brain development, and pubertal maturation on adolescent risk-taking. *Horm. Behav.* 64, 323–332. doi: 10.1016/j.yhbeh.2013.03.006

StataCorp. (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp .

Steinbeis, N., Singer, T., Fehr, E., and Haushofer, J. (2014). Development of behavioral control and associated vmPFC–DLPFC connectivity explains children's increased resistance to temptation in intertemporal choice. *Cereb. Cortex* 26, 32–42. doi: 10.1093/cercor/bhu167

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends Cogn. Sci.* 9, 69–74. doi: 10.1016/j.tics.2004.12.005

Steinberg, L., and Sheffield Morris, A. (2001). Adolescent development. *Annu. Rev. Psychol.* 52, 83–110.

Stevens, J. S., and Hamann, S. (2012). Sex differences in brain activation to emotional stimuli: a meta-analysis of neuroimaging studies. *Neuropsychologia* 50, 1578–1593. doi: 10.1016/j.neuropsychologia.2012.03.011

Sutter, M., and Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games Econ. Behav.* 59, 364–382. doi: 10.1016/j.geb.2006.07.006

Tabibnia, G., and Lieberman, M. D. (2007). Fairness and cooperation are rewarding. *Ann. N. Y. Acad. Sci.* 1118, 90–101. doi: 10.1196/annals.1412.001

Tzieropoulos, H. (2013). The trust game in neuroscience: a short review. *Soc. Neurosci.* 8, 407–416. doi: 10.1080/17470919.2013.832375

van de Groep, S., Meuwese, R., Zanolie, K., Güroğlu, B., and Crone, E. A. (2018). Developmental changes and individual differences in trust and reciprocity in adolescence. *J. Res. Adolesc.* doi: 10.1111/jora.12459 [Epub ahead of print] .

Van den Akker, O. (2018). Sex differences in trust and trustworthiness-a meta-analysis of the trust game and the gift-exchange game. *PsyArXiv* doi: 10.31234/osf.io/5zbja

Van Den Bos, W., Güroğlu, B., Van Den Bulk, B., Rombouts, S., and Crone, E. (2009). Better than expected or as bad as you thought? The neurocognitive development of probabilistic feedback processing. *Front. Hum. Neurosci.* 3:52. doi: 10.3389/neuro.09.052.2009

van den Bos, W., van Dijk, E., and Crone, E. A. (2012). Learning whom to trust in repeated social interactions: a developmental perspective. *Group Process. Intergroup Relat.* 15, 243–256. doi: 10.1177/1368430211418698

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A., and Crone, E. A. (2011). Changing brains, changing perspectives the neurocognitive development of reciprocity. *Psychol. Sci.* 22, 60–70. doi: 10.1177/0956797610391102

van den Bos, W., Westenberg, M., van Dijk, E., and Crone, E. A. (2010). Development of trust and reciprocity in adolescence. *Cogn. Devel.* 25, 90–102. doi: 10.1016/j.cogdev.2009.07.004

Van Duijvenvoorde, A. C., Zanolie, K., Rombouts, S. A., Raijmakers, M. E., and Crone, E. A. (2008). Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. *J. Neurosci.* 28, 9495–9503. doi: 10.1523/jneurosci.1485-08.2008

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence. The Psychological Corporation* New York, NY: Harcourt Brace & Company.

Zhang, M., Liu, T., Pelowski, M., Jia, H., and Yu, D. (2017). Social risky decision-making reveals gender differences in the TPJ: a hyperscanning study using functional near-infrared spectroscopy. *Brain Cogn.* 119, 54–63. doi: 10.1016/j.bandc.2017.08.008

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

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 Race bias In/out-group bias Prosocial orientation	Prosocial orientation Personality traits Pupil dilation
Explicit	Experienced reciprocity Experienced trust violation Moral character Reputational priors	Bestowed trust Risk associated with trust Prior experience of reciprocity 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.

frontiers in Neuroscience

REVIEW
published: 10 September 2019
doi: 10.3389/fnins.2019.00887

Trust Games and Beyond

Carlos Alós-Ferrer[*†] *and Federica Farolfi*[†]

Department of Economics, Zurich Center for Neuroeconomics, University of Zurich, Zurich, Switzerland

Trust is fundamental for the stability of human society. A large part of the experimental literature relies on the Trust Game as the workhorse to measure individual differences in trust and trustworthiness. In this review we highlight the difficulties and limitations of this popular paradigm, as well as the relations to alternative instruments ranging from survey measures to neurochemical manipulations and neuroimaging.

Keywords: trust, trustworthiness, reciprocity, survey measures, social preferences, oxytocin, theory of mind, social neuroscience

OPEN ACCESS

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

Reviewed by:
Jun Tanimoto,
Kyushu University, Japan
Ananish Chaudhuri,
The University of Auckland,
New Zealand

****Correspondence:***
Carlos Alós-Ferrer
carlos.alos-ferrer@econ.uzh.ch

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Decision Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *30 April 2019*
Accepted: *07 August 2019*
Published: *10 September 2019*

Citation:
Alós-Ferrer C and Farolfi F (2019) Trust Games and Beyond. Front. Neurosci. 13:887. doi: 10.3389/fnins.2019.00887

1. INTRODUCTION

Trust is an essential ingredient of economic life. We implicitly or explicitly trust our financial institutions, employers, coworkers, and fellow citizens on a daily basis. Without trust, nobody would accept intrinsically-valueless bills and coins (or electronic transfers) in exchange for goods or services, or show up to work in exchange for the promise of later compensation. Yet, in spite of its fundamental importance, trust is an elusive concept which remains hard to quantify. How can we measure heterogeneity in trust or trustworthiness? Is there a quantifiable, measurable way to show that certain institutions foster (increase) trust?

Experimental economics has developed a small family of stylized paradigms used for precisely this purpose. They build on a bare-bones conceptualization of trust, which, in our view, is as follows. Trust is revealed when an agent performs an initial *sacrifice*, that is, an action which, depending on the reaction of another agent, might be detrimental to the first agent's own interests. *You put yourself in somebody else's hands.* Trust is repaid, and the second agent is revealed to be trustworthy, if his or her reaction offsets and compensates the first agent's sacrifice. Obviously, for such a situation to reflect trust and trustworthiness, the interaction must be isolated and free of any extraneous elements as might arise from strategic concerns due to repetition, coercion, etc. For this purpose, experimental economics has relied on paradigms which can be seen as *games* in the stringent sense of *game theory* (von Neumann and Morgenstern, 1944): complete descriptions of interpersonal, strategic problems. Since the very idea of trust requires a temporal structure, one ends with an *extensive form game* (e.g., Alós-Ferrer and Ritzberger, 2016), where some agent can observe and react to previous actions of another agent, creating the need for the latter to predict and forecast the reactions of the former.

The essence of the *Trust Game*, extensively used in economics as an experimental, incentivized measure of trust, is as follows. A first agent, called the *trustor*, is given a monetary endowment X, and can choose which fraction p of it (zero being an option) will be sent to the second agent, called the *trustee*. The transfer $p \cdot X$ is then gone, and there is nothing the trustor can do to ensure a return of any kind. Before the transfer arrives into the trustee's hands, the transfer is magnified by a factor $K > 1$ (e.g., doubled or tripled). That is, the trustor might send, say, \$5 but the trustee receives \$10 or more. The trustee is free to keep the whole amount without repercussion. Crucially, however, the trustee has the option to send a fraction q of the received transfer back to the trustor, hence honoring the trustor's initial sacrifice. Since p and q can in principle be any proportion, this is an infinite game, although in practice experimental implementations discretize the decisions, for instance requiring transfers to be integers. In the laboratory, roles are assigned randomly, the trustor-trustee matching is equally random, and interactions are computerized, one-shot, and anonymous, with the aim of isolating the essence of trust and trustworthiness.

The game described above is universally referred to as the *Trust Game* nowadays, and it is in this sense that we will use this name. However, the game was originally called the *Investment Game* by Berg et al. (1995), who used an endowment of $X = \$10$ and tripled the transfer, $K = 3$. Further, the name *Trust Game* was used for an earlier and simpler game by Kreps (1990). In that version, the trustor has the binary choice to trust the trustee or not, with payoffs of $0 for both players if no trust is shown. In case the trustor decides to trust, the trustee faces a binary choice to either honor it, leading to equal payoffs of $10 for each player, or abuse the demonstrated trust, resulting in a payoff of $15 for the untrustworthy trustee and a negative payoff of −$5 for the unhappy trustor. **Figure 1** presents standard game-theoretic depictions of Kreps's (1990) game (a), Berg et al.'s (1995) continuous game (b), an example of a discretized version where only integer transfers are allowed (c), and, for later reference, a mini-Trust Game with binary choices and general payoffs (d). In the latter, the structure of the Trust Game is preserved if $G > S > B$ and $C > H > T$, since in this case the ordinal preferences among outcomes is preserved: The trustor would prefer to trust if trust is repaid, but not if trust is abused, the trustee's payoffs are maximized by betraying trust, but there is an option where trust is repaid and both players are better off than if no trust is shown. A binary version of the game of Berg et al. (1995), where the amount transferred by the trustor is multiplied by $K > 1$ and the trustee decides on a split of the transfer, obtains if, additionally, $B + C = G + H = K \cdot S$.

While the structure of Kreps's (1990) game is different (and does not correspond to a Trust Game as currently understood), both games share four crucial features, which were put forward by Coleman (1990) to define a trust situation. First, the trustor's decision to trust is voluntary. Second, there is a time lag between the trustor's and the trustee's choices. Third, the possibility for the trustee to abuse or honor the demonstrated trust occurs if and only if the trustor does indeed show trust. And last, in case the trustee decides to (fully) abuse the demonstrated trust, the trustor will be left worse off than if no trust had been shown; that is, the trustor becomes vulnerable by exercising trust (Fehr, 2009). We would like, however, to add a fifth element to the list for a trust situation to become economically interesting: from the point of view of economic efficiency, trust should be optimal, at least in the sense of maximizing the sum of payoffs [in Kreps's (1990) version, efficiency further requires that trust is repaid; this is not the case in Berg et al.'s (1995) game, as seen most clearly comparing (a) and (d) in **Figure 1** for the multiplier case, $B + C = G + H = K \cdot S$].

As a consequence of these elements, rational but selfish agents fare poorly in these and similar games. A selfish trustee will never send any money back ($q = 0$), and, anticipating this (as built into the game-theoretic solution of *subgame-perfect equilibrium*), a selfish trustor will never make any transfer ($p = 0$). Needless to say, actual human beings are far more trusting and more trustworthy in the laboratory than selfishness would imply. The actual transfers of trustors can then be construed as a measure of trust, and the reactions of the trustees as a measure of trustworthiness [both becoming continuous measures in Berg et al.'s (1995) version].

The Trust Game was put forward at a time when the experimental literature was developing paradigms to measure not only trust, but also many other related constructs such as fairness or reciprocity. As a result, cross-fertilization or simply convergence of ideas is often apparent in experimental designs in behavioral economics. For example, the *Dishonest Salesman Game* (Dasgupta, 1988) framed the interaction as the purchase of a car at price $\beta \in (0, 1)$, where the salesman can hand over a reliable car (for a utility of 1 for the buyer and $\alpha > 0$ to himself) or a lemon (for a utility of zero for the buyer and $\gamma > \alpha$ for himself). A transformation of payoffs shows that this game is ordinally identical to Kreps's (1990). In the *Trading Game* of Lyons and Mehta (1997), after a previous, non-binding agreement, a Supplier decides how much to invest (say, effort or capital) and then a Buyer decides whether to pay as agreed or delay (unilaterally renegotiate the terms down). Other prominent examples have embedded trust-based interactions in more complex paradigms. For instance, in the basic building block of the *Gift-Exchange Game* (Fehr et al., 1993), employers make wage offers which employees can repay with appropriate effort levels. Employees have no incentive to provide any effort above the minimum level, which, if anticipated by the employers, leads to minimal wages. However, both employer and employee are better off if the employer trusts the employee by offering a wage above the minimum and the employee pays back that trust by exerting a higher effort. The *Lending Game* (Camerer and Weigelt, 1998) studies reputation formation in an incomplete-information setting where a borrower (whose type is unknown) interacts with several lenders, but each bilateral interaction displays the elements of a trust situation. Among all these and other games, however, it is Berg et al.'s (1995) game, and the label *Trust Game*, which has established itself as the most prominent instrument to measure trust in the laboratory, resulting in a large number of experimental replications and variations (see, e.g., Glaeser et al., 2000).

We would like to emphasize that the Trust Game and all the variants mentioned above arose from the discipline of *game theory*, and hence it might be worth providing some additional context at this point. This extensive, highly-developed, interdisciplinary field covers the formal and empirical study of interpersonal, strategic relations among multiple agents (see, e.g., Fudenberg and Tirole, 1991). For instance, *normal form games* are strategic situations where all involved agents act simultaneously. Starting with binary-action games (where a number of players have just two different actions each, the simplest case being two-player, 2 × 2 games), their study allows research in many relevant social issues, as coordination in efficient technologies. A case of particular interest is the study of cooperation in society using famous paradigms as the *Prisoner's Dilemma* (see Poundstone, 1993, for a detailed overview) or public good games (e.g., Fehr and Gächter, 2000). Indeed, large parts of the literature in social sciences beyond economics has often focused on 2 × 2 games as the Prisoner's Dilemma (e.g., Axelrod, 1984, 1997).

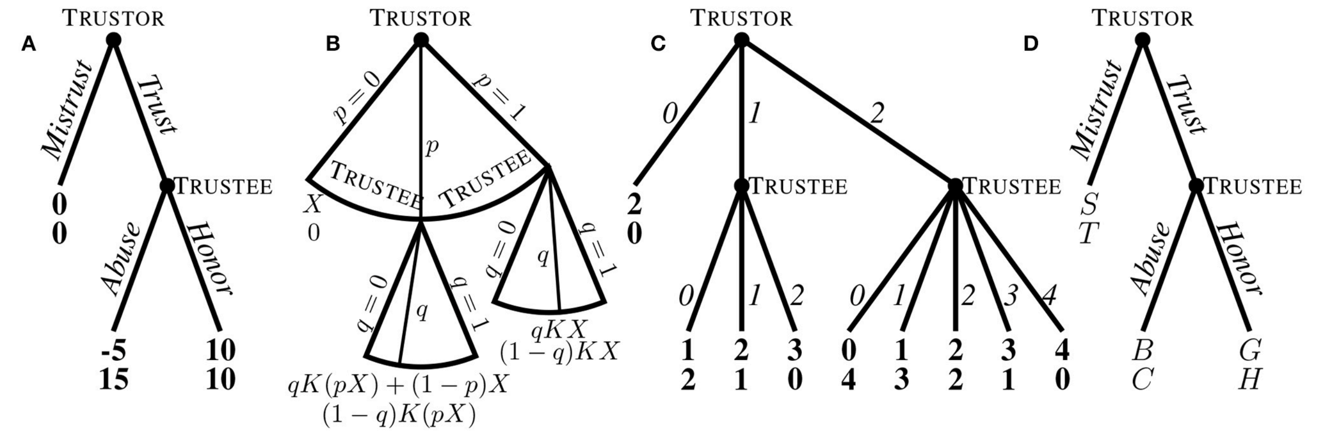


FIGURE 1 | (A) The original Trust Game of Kreps (1990), where both trustor and trustee have just binary choices. **(B)** A qualitative depiction of the infinite game representing the continuous version of the Trust Game of Berg et al. (1995), as envisioned in experimental economics. The trustor is endowed with $X and can send any proportion $p \in [0, 1]$. The transfer pX is multiplied by a factor $K > 1$. The trustee receives $K(pX)$ and can send back any proportion thereof, $q \in [0, 1]$. The trustee has a continuum of possible, alternative decision nodes, corresponding to all possible transfers by the trustor. **(C)** A discretized version of the Trust Game. The trustor is endowed with $X = 2$ and transfers are doubled ($K = 2$). Both trustor and trustee can only send integer amounts. **(D)** A heavily-discretized "mini-Trust Game," as used, e.g., in Bohnet and Zeckhauser (2004). In all games, the upper payoff is the trustor's and the lower payoff is the trustee's.

The Trust Game, however, is an *extensive form game*. This class of games allows to incorporate non-simultaneous play, and in particular reactions to previous actions of other agents (see Alós-Ferrer and Ritzberger, 2016, for a detailed formal treatment). The simplest examples, where all actions are observable, include paradigms which have been intensively used in *behavioral game theory* (see, e.g., Camerer, 2003) to investigate prosocial behavior, i.e., deviations from selfishness. For instance, in the Ultimatum Game (Güth et al., 1982) a proposer can offer a split of an endowment among two players, and a responder can then either accept it or destroy the entire endowment. On the basis of purely monetary payoffs, the normative *subgame-perfect equilibrium* predicts that the responder will accept any positive amount and, anticipating this, the proposer will offer as little as physically possible. In contrast, laboratory experiments show that human proposers make substantial offers and human responders reject small but positive offers. This does not, however, constitute a demonstration of prosocial behavior, since proposers might make positive offers strategically, to avoid rejections. For this reason, in the Dictator Game (Forsythe et al., 1994), which we will refer to in the following sections, the responder is passive and the proposer's decision is dictatorially implemented. Still, in this game human dictators typically grant positive amounts to the other player, in a striking deviation from selfishness.

Evolutionary game theory (see, e.g., Weibull, 1995), has focused on stylized games played in populations of agents to study the long-run evolution of fundamental features of society. Those include the evolution of cooperation (Nowak and Sigmund, 1993) and social preferences (Binmore et al., 1995; Miyaji et al., 2013), but the study of trust is so far underrepresented in this field. Although this subdiscipline has developed in economics and mathematical biology, it has recently been the subject of increased attention in other disciplines (e.g., Tanimoto, 2015, 2019).

In this review, we examine the difficulties and confounds inherent in the Trust Game, which include social preferences (section 2), attitudes to interpersonal risk (section 3), and other factors leading to a lack of stability of the paradigm (section 4). We also emphasize the differences in (measurements of) trust and trustworthiness (section 5), and conclude by exploring the relations to alternative instruments, in the form of survey questions (section 6), neurochemical manipulations and neuroimaging (section 7).

2. SOCIAL PREFERENCES AS A CONFOUND

In spite of its widespread use to measure trust and trustworthiness in the lab, the Trust Game is not exempt of critiques. An important one is the possible presence of motivational confounds, very especially in the form of other-regarding preferences (e.g., Fehr et al., 1993; Fehr and Schmidt, 1999). If a trustor is selfish, the decision to trust should be motivated exclusively by the belief that the trustee will reciprocate. However, since the amount transferred is magnified by a factor $K > 1$, an altruistic trustor might decide to transfer resources even if he or she does not expect any transfer back, since what the trustor sacrifices is far less than what the trustee receives. Additionally, since in Berg et al.'s (1995) game efficiency does not require that trust is repaid, even an efficiency-motivated trustor (Bolton and Ockenfels, 2000; Charness and Rabin, 2002) might be willing to make a transfer even when not expecting a return. A similar critique applies to the trustee's transfer as a measure of trustworthiness. It is unclear whether a trustee's decision to transfer back arises exclusively from the desire to reciprocate (which is what is usually understood as trustworthiness) or from unconditional other-regarding preferences (e.g., inequity aversion).

To address these possible confounds, Cox (2004) conducted a between-subjects experiment, with one of the treatments being a standard Trust Game ($K = 3$). Confounds in trustor motivation were addressed through a second treatment implementing a Dictator Game (Forsythe et al., 1994), where "trustors" make the same decision as in the Trust Game (in particular, the amount sent is tripled) but "trustees" are passive and cannot reciprocate. A majority of trustors in the Dictator treatment made positive transfers, but transfers were significantly larger in the Trust treatment (on average, \$5.97 vs. \$3.63). Confounds in trustee motivation were addressed through a third treatment where trustors were passive, receiving an endowment equivalent to the fraction kept by trustors in the Trust treatment (crucially, not framed as a transfer). Trustees received three times the remaining fraction (plus a fixed endowment) and could send an amount to the passive trustors as in a Dictator Game. Roughly one third of the "trustees" in this treatment sent a transfer to the other player, even though reciprocity was not a factor. However, trustee average transfers were significantly larger in the Trust treatment (\$4.94 vs. \$2.06). This suggests that trust and reciprocity are indeed present as a motivation in the trustors' and trustees' decisions, respectively, but they are not isolated by the paradigm. On the contrary, a non-negligible part of the transfers of both types of agents might be motivated by prosocial preferences. For the case of trustors, this was confirmed by Chaudhuri and Gangadharan (2007), who ran an experiment including a Trust Game and a Dictator Game (see section 5 below). Again, transfers were significantly larger in the Trust Game (on average, \$4.33 vs. \$1.345). Further, the difference between the amount sent in the Trust Game and the amount sent in the Dictator Game was predicted by the elicited expectation for a back transfer from the receiver.

In conclusion, it might be unwarranted to use behavior in the Trust Game as a pure indicator of trust or trustworthiness. Due to the confound with social preferences, this game might be overestimating both dimensions of human behavior. To remove the confound, one should rely not on transfers in the Trust Game, but (whenever possible) on the within-subjects differences between those transfers and transfers in control games inspired on the Dictator Game. This might, of course, create difficulties of its own. For instance, Ashraf et al. (2006) report order effects depending on whether players played the Trust Game or a Dictator Game first.

3. TRUST AND INTERPERSONAL RISK

Trusting someone puts you in a vulnerable position. By definition, the decision to trust implies assuming a risk. Hence, it is natural to ask whether attitudes toward risk influence the willingness to trust and hence whether there is another confound when measuring trust through the Trust Game. For instance, Houser et al. (2010) investigated the relation between trust and risk, measuring attitudes toward risk through the standard procedure of Holt and Laury (2002). However, the study did not find any systematic relation between trust decisions and risk attitudes. In contrast, risk attitudes did explain behavior in "risk games" where the trustee's decision was replaced by a known distribution. In a related study, Fairley et al. (2016) used a *risky Trust Game* as follows. Trustees' binary decisions to either keep the transfer or return half (independently of which proportion of the endowment was transferred) were elicited in advance. Trustors were told they would be matched with one out of four pre-determined trustees, and asked to provide their decisions conditional on how many of those trustees had decided to make a reciprocal transfer (Conditional Information Lottery design; Bardsley, 2000); hence, they provided five separate decisions. In practice, the trustor's decision was equivalent to a lottery. Behavior was compared to that in a standard Trust Game with no information on the trustee. Behavior in the risky Trust Game was used to estimate risk attitudes, and the resulting values do predict behavior in the Trust Game, although a standard measure obtained using Holt and Laury's (2002) procedure did not.

The lack of relation between risk attitudes and behavior in the Trust Game might be due to the fact that the risk involved in the Trust Game is of two qualitatively different kinds. On the one hand, there is the purely financial, dispassionate one, i.e., the risk to lose the money invested. On the other hand, there is a more psychological but not less-real risk, namely the risk to be betrayed by the trustee. More generally, attitudes toward risk in social and non-social situations might differ. To investigate this question, a number of studies have compared behavior in the Trust Game with behavior in risky situations where the social component is eliminated, but are otherwise equivalent to the Trust Game (in terms of outcomes) from the purely individual point of view.

Bohnet and Zeckhauser (2004) coined the term "betrayal aversion" to refer to the social aspect of risk in the Trust Game. In their study, they examined the question of whether the decision to trust a stranger is equivalent to taking a risky bet, or, on the contrary, the possibility of being betrayed by another human being represents an actual cost. For this purpose, they considered a mini-Trust Game as in **Figure 1D**, with $S = T = 10$, $B = 8$, $C = 22$, and $G = H = 15$ (and an implied $K = 3$). A distribution of responses (conditional on being trusted) was previously elicited from a population of trustees, resulting in an actual proportion of trustworthy players, p (unknown by trustors). Then, instead of a binary decision, a Minimum Acceptable Probability (MAP) was elicited from each trustor, with the explicit meaning that the trustor would actually trust if and only if the expressed MAP was larger than p. That is, if the MAP was larger than p, the trustor's decision would be implemented as "trust," and the game's payoffs would be (G, H) with probability p or (B, C) with probability $1 - p$. If the MAP was smaller than or equal to p, the trustor's decision would be implemented as "mistrust." The actual implementation varied across three variants of the game. In the actual Trust Game, the implementation was done by actually matching the trustor with a random trustee from the distribution, so that the outcome depended on the actual decision of the selected trustee. In a risky Decision Problem (framed as such), the outcome was implemented through a lottery with probabilities $(p, 1 - p)$, and, independently of the outcome, nobody received the trustee's payoffs; that is, in this case the participants chose between the safe payoff S and a lottery paying G with probability p and B with probability $1 - p$. The third variant was a Dictator Game which was identical to the risky Decision Problem, with the only difference that the trustee's payoffs corresponding to the actual outcome were received by an uninvolved, passive player.

The results of Bohnet and Zeckhauser (2004) showed a larger MAP for trusting in the Trust Game than for taking the risky option in the other games (but there were no differences among the latter two). Hence, participants revealed an aversion to experience betrayal in the Trust Game, separate from the non-social component of risk attitudes. This is an important result for the understanding of what the trustor decision actually measures in the Trust Game. Together with the results discussed in section 2, the results of Bohnet and Zeckhauser (2004) indicate that the decision to trust can be decomposed into a purely prosocial motivation and the willingness to assume risks of an interpersonal, social nature.

The relevance of betrayal aversion has been established in other studies. For instance, in a study using functional Magnetic Resonance Imaging (fMRI), Aimone et al. (2014) studied trusting behavior while controlling for social preferences. In their study, trustors played binary mini-trust games for 41 trials against human trustees whose decisions had been previously elicited, and for 41 further trials against random computer-generated decisions. Crucially, however, in the latter case another human being actually received the corresponding trustee payoff. Hence, the within-subject comparison controls for social preferences, but the trials with a computerized opponent should remove betrayal aversion. In line with Bohnet and Zeckhauser (2004), trust was observed significantly more often when playing against the computer (63%), compared to the trials with a human opponent (49%), although the effect was driven by male trustors.

Evidence to the contrary was presented by Fetchenhauer and Dunning (2012), who confronted participants with a binary-choice mini-Trust Game (as in **Figure 1D** with $S = 5$, $T = 0$, $B = 0$, $C = 20$, and $G = H = 10$, with an implied $K = 4$) and the choice to play an equivalent lottery. Trustee decisions (conditional on trust being shown) were collected in advance, and two different pools of decisions were created, containing 80 and 46% of trustworthy answers, respectively (High Chance and Low Chance conditions). Trustors were informed that the trustee's answer to their own decision would be extracted from the corresponding pool. Trustors also made an equivalent lottery choice decision, namely either to receive $S = \$5$ for sure or a lottery paying $G = \$10$ with either 80 or 46%, and zero otherwise. While there were no significant differences in the High Chance condition, in the Low Chance condition there were large differences (28.6% gambling in the lottery vs. 54.3% risking to trust in the Trust Game). That is, when the chances of winning were moderate, the decision to take the risk was made more often if the risk had a social component.

The results of Fetchenhauer and Dunning (2012) are in striking contrast to those of Bohnet and Zeckhauser (2004). The latter found that trust was increased when betrayal by a human being was not possible, i.e., there was less trust (as implied by a higher MAP) in the Trust Game compared with the risky version. Fetchenhauer and Dunning (2012) found that trust was reduced in the risky version of the game compared to the Trust Game, as revealed by the binary decisions on whether to trust or not. Those authors argue that the difference accrues from the elicitation methods. The MAP might elicit (abstract) betrayal aversion, but people are reluctant to openly signal distrust within the actual game. In light of these findings, further research should concentrate on clarifying the effects of betrayal aversion under different elicitation methods. There are, however, other differences in the designs which prevent a direct comparison. As also pointed out by the authors, in the design of Fetchenhauer and Dunning (2012) a trustor's mistrust decision gives the trustee zero payoffs and yields an unequal outcome ($S = 5, T = 0$), while in Bohnet and Zeckhauser (2004) in this case both players receive the same, positive payoff ($S = T = 10$). Hence, trustors might simply be reluctant to take the responsibility to leave the trustee empty-handed. A more complex argument would point out that the decision to trust in Fetchenhauer and Dunning (2012) "saves" the trustee from zero payoffs and might elicit stronger reciprocity motives than in Bohnet and Zeckhauser (2004), which in turn might be anticipated by the trustors. This is, however, at odds with the fact that trustors knew both that trustees' conditional decisions had been collected in advance *and* the percentage of trustworthy answers. We remark also that the sign of the difference found by Fetchenhauer and Dunning (2012) is consistent with the social-preferences confound described in section 2: since there were no trustees in the risky version, one could speculate that the higher transfers in the Trust Game of Fetchenhauer and Dunning (2012) might have been due to altruistic motivations which would naturally be absent in the risky version, where no trustee was present. However, this is inconsistent with the fact that there were no differences in High Chance condition.

In conclusion, betrayal aversion might be one of the main motivations behind the decision (not) to trust. Since this reflects a particular kind of *social* risk, researchers should be aware that standard measures of risk attitudes might not be well-suited to the study of trust. At the same time, this observation shows an essential defining characteristic of experimental paradigms measuring trust, in the sense that potential game "variants" which remove or weaken the social aspect of the trusting decision might very well end up measuring unrelated characteristics.

4. LACK OF STABILITY OF THE PARADIGM

In this section, we discuss current evidence showing that minor changes in the parameters, implementation, and description of the Trust Game might sometimes induce large changes in players' responses. This is problematic, as it suggests that the paradigm might not be as stable as would be desirable for an instrument measuring an aspect of human motivation.

A first example is the size of the multiplier, which was set to $K = 3$ in the original version of Berg et al. (1995). Lenton and Mosley (2011) found evidence that increasing the multiplier (from 2 to 3 or 4) increases the fraction of the endowment sent by the trustor. Some studies have shown that increases in the multiplier also increase the fraction returned by the trustee (reciprocity), comparing e.g., 3 vs. 6 (Ackert et al., 2011) and 2 vs. 4 (Mislin et al., 2015). However, the meta-analysis of Johnson and Mislin (2011), which included trust games with many implementation variations, reached the puzzling conclusion that increasing the multiplier from 2 to 3 *decreases* the trustee's transfers, but it does not affect the trustor's transfers. This finding coexists with the observation that trustees respond to larger fractions transferred by the trustors by transferring back larger fractions of the income they receive.

A second factor which might affect behavior in the Trust Game, specifically trustees' transfers, is whether answers are elicited through the strategy method (Selten, 1967) or as answers to the actual trustor decision. In the former, trustees are asked what their return transfer would be conditional on each possible transfer of the trustor (before the actual one is revealed), and trustor-trustee decisions are paired afterwards. In the latter, trustees are confronted with the actual trustor decision and asked to react to it (and only to it). It has been argued that the strategy method might in general induce more deliberative, "cold" thinking in experiments (Brandts and Charness, 2000), and in particular Casari and Cason (2009), argue that this method might reduce transfers of trustees in the Trust Game. However, Brandts and Charness (2011) found no difference.

Another example is framing. Burnham et al. (2000) consider an extensive form game where each player has multiple decisions but the first two decisions roughly correspond to a trust situation. They show that players in the role corresponding to a trustor trusted more if the instructions called the other players "partners" rather than "opponents." However, in many implementations of the Trust Game, the word "Trust" is not mentioned at all, hereby avoiding framing effects. In an EEG experiment, Sun et al. (2019) framed a (repeated) Trust Game either literally as a "Trust Game," or alternatively as a "Power Game," and found that earnings (and hence trusting behavior) were larger in the first case.

A subtler issue related to framing is that the way the instructions of a game are spelled out might influence how participants interpret the situation, and also whether a shared interpretation arises. As pointed out by Ermisch and Gambetta (2006), even the attempt to keep the game frame-free raises the concern that trustees might develop other, alternative interpretations. A demonstration of the effects of a shared interpretation was given by Cronk (2007), who conducted trust games among Maasai natives, with $K = 3$ and the initial trustor endowment X corresponding to about a day's wage at the time. Half the games were kept unframed, and the rest were explicitly called *osotua* games. This word describes a strong, culture-specific concept where a request of a gift or favor arises out of genuine need, and granting it creates a strong, long-lasting bond. Both trustor's transfers and trustee's returns were smaller in the *osotua* treatment. Crucially, there was a *negative* correlation between the trustor's transfer and the trustee's return share, which was absent in the unframed treatment. This is natural given that *osotua* refers to freely-given gifts in case of genuine need, but in the absence of this culture-specific information, the evidence might be misinterpreted as reduced trust and trustworthiness.

Chaudhuri et al. (2016) conducted several laboratory treatments using trust games ($X = 10$, $K = 3$) with incremental differences in the instructions. Some of the treatments provided context by explicitly spelling out the conflict underlying the trustor's decision. Specifically, the instructions contained two paragraphs which described the subgame-perfect outcome (where the trustor "sends" zero) and the fact that both players would be better off if the trustor sent the entire endowment of 10 units and the trustee returned more than 10 units back. The trustors' transfers were significantly larger in the context-rich treatments, compared to a control. Importantly, also the trustees' return shares were larger in the context-rich treatments, i.e., trust did pay off. At the same time, there were no significant differences across the context-rich treatments even though two of them, but not the third, included the words "trust" and "trustworthy." That is, the framing effect was not due to the use of particular words but rather to the fact that the conflict between self-interest and the maximization of social surplus was made evident. At a conceptual level, this study and Cronk (2007) suggest that players in the Trust Game (and, more generally, in many experimental paradigms) might often face some uncertainty on whether all involved participants share a common view of the game. Both cultural labels (shared by all participants) and explicit information on the nature of the conflict underlying player decisions help reduce such uncertainty.

In conclusion, mixed evidence on several fronts has cast doubt on the stability of the measures derived from the Trust Game. Further, systematic research is needed to clarify to what extent those measures reflect stable personality traits or rather situation-specific reactions. Comparability to the literature can only be guaranteed by relying on designs as close as possible to those used in the relevant, previous contributions. The issue of framing is particularly worrying, and simply striving to keep a neutral framing might not always be enough to ensure that the subjects' interpretation of the game coincides with that of the experimenter.

5. TRUST VS. TRUSTWORTHINESS

Trust and trustworthiness go hand-in-glove, as one cannot exist without the other. However, no matter how interrelated they might be, they are clearly different concepts. For instance, while monetary concerns might partially explain the decision to trust (accepting interpersonal risk in order to obtain a higher return), they cannot explain the decision to repay trust, as the selfish, profit-maximizing decision is always to keep the entire transfer received. Hence, one should expect that trust and trustworthiness are explained by partially different determinants. A key contribution on this front is Ashraf et al. (2006), who confronted $N = 359$ college students from different countries with a Trust Game, two dictator games, and a number of questionnaires (including measurement of risk attitudes). Half of the participants played the Trust Game ($K = 3$) as trustors, and the rest as trustees. Trustors were also asked what they expected to get back, as a measure of their expectation of trustworthiness. The dictator games were a regular one and a

"Triple Dictator Game" identical to the Trust Game (so the amount sent by the first player was tripled) except that the second player was passive and could not return anything back (as in the Dictator treatment of Cox, 2004). The amount sent in the Triple Dictator Game is used as a proxy of "unconditional kindness" or prosocial behavior for the trustors, while the amount sent in the regular Dictator Game plays the same role for the trustees.

The first observation is that, out of the 159 trustors who sent a positive amount, only 36% expected back more than what they sent. This suggests that most people exhibit trusting behavior due to motivations other than purely monetary ones. As in Cox (2004) (recall section 2), trustor behavior is partially explained by prosocial behavior (the amount sent in the Triple Dictator Game). However, a regression analysis shows that most of the variance in trustor transfers is actually explained by the expectations of trustworthiness. On the other hand, trustees' return transfers are explained both by trust shown (the amount sent by trustors) and prosocial behavior (the amount sent in the regular Dictator Game). The latter is a restatement of the observation that, as trust, trustworthiness as measured in the Trust Game is confounded with prosocial behavior (recall again section 2). As for the former, the relation between the trustor's transfer and the trustee's return (elicited here through the strategy method) is commonly taken as a demonstration of reciprocity. Ashraf et al. (2006) further argue that trustees' transfers are better explained by prosocial motivation than by reciprocity, because, in a regression analysis, the amount sent in the Dictator Game explains most of variance in trustees' transfers, compared to a model where the trustor's transfer is also included as a regressor. Interestingly, the authors also consider a different measure of other-regarding preferences, namely the "predicted distributional preference." This is the amount that the trustee would need to return to the trustor to create the same payoff ratio as the trustee created in the regular Dictator Game. Hence, it is a function of both the trustor's transfer and the amount sent by the trustee in the Dictator Game. When included in a regression, this variable captures almost the same variance as that explained in a model including both of the latter variables.

Since the determinants of trust and trustworthiness are different, it is natural to ask whether, at the individual level, being more trusting implies that one is also more trustworthy, and vice versa. The Trust Game measures trust and trustworthiness as the behavior of the two different players, trustor and trustee. Hence, the relation between trust and trustworthiness at the individual level can be tackled by relying on experimental designs where participants play both roles (in different trust games).

Chaudhuri et al. (2003) let 76 participants play both roles in a bargaining game with a structure akin to a mini-Trust Game, but with the option for the trustee to (costly) punish if no trust was shown (an option that nobody used). Of the 39 participants who trusted their counterparts and were themselves shown trust when in the trustee role, 18 did not reciprocate, suggesting that people who trust are not necessarily trustworthy. This hypothesis was confirmed in Chaudhuri and Gangadharan (2007), who collected data from participants who played both roles in trust games ($X = \$10$, $K = 3$). Dividing participants into trusting and non-trusting depending on whether they transferred 50% or more ($N = 42$) or less than 50% ($N = 58$), respectively, they found no significant differences in their average return transfers (16 and 18%, respectively). On the other hand, dividing participants into trustworthy and less trustworthy depending on whether they returned one third or more ($N = 27$) or less than one third ($N = 55$) of the amount actually offered to them ($N = 18$ received zero), they found that the trustworthy participants sent significantly more as trustors than the less trustworthy ones ($5.33 vs. $3.82 on average). That is, while trustworthy participants were found to be generally trusting, there was no evidence that more trusting individuals are also necessarily more trustworthy.

Chaudhuri and Gangadharan (2007) argue that what has been interpreted as trust in many studies could be decomposed in two components. One essentially corresponds to the predisposition to accept a social risk that we discussed in section 3, and that obviously plays a role for trustors but not trustees in the Trust Game. The other is a general prosocial orientation, related to the social-preferences confound that we discuss in section 2, and that could be considered a "social virtue" in the sense of Fukuyama (1995). For instance, participants in the experiment of Chaudhuri and Gangadharan (2007) also played as dictators in a Dictator Game, and trustworthy participants transferred significantly more in the latter game than less trustworthy ones ($1.89 vs. $0.83, respectively). Chaudhuri and Gangadharan (2007) conclude that trustworthiness as measured in the Trust Game might be more relevant than trust for the study of social capital and its relation to economic growth.

In conclusion, trust and trustworthiness are interrelated but different concepts, influenced by different individual characteristics and factors. There is evidence suggesting that trustworthy individuals might be generally more trusting, but the converse is in general not true.

6. SURVEYS AND THE TRUST GAME

Besides experimental paradigms as the Trust Game, the most common method for measuring trust is the use of generalized trust questions in surveys. The most prominent example is the General Social Survey (GSS) of the U.S. National Opinion Research Center (http://gss.norc.org), which has collected evidence on trust and social capital since 1972. The specific question used to measure trust was adapted from Rosenberg's (1956) misanthropy index (see Uslaner, 2012) and reads as follows. "*Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?*" The question is a binary-choice one with two possible answers, "*Most people can be trusted*" or "*Can't be too careful*" (plus "I don't know"). This question is used in many other surveys, as, e.g., the European Values Survey, the World Values Surveys (WVS), the British Household Panel Study (BHPS), and American National Election Studies (ANES).

The question, however, is not exempt of criticism. Obviously, responses to those depend on the respondents' interpretation of the questions and their personal experiences. The question has been criticized for being too generic (Ermisch et al., 2009) and for reducing a presumably-continuous characteristic to a dichotomous answer (Lundmark et al., 2015), although Uslaner (2012) points out that binary answers avoid "clumping" in intermediate options. Most importantly, Miller and Mitamura (2003) have pointed out that the two alternative answers refer to an assessment of other people's trustworthiness and an assessment of one's own willingness to take risks, respectively. That is, respondents are forced to choose between trust and caution, although these options are not opposed (recall section 3).

Glaeser et al. (2000) tested whether attitudinal trust questions from surveys predict actual, incentivized trusting behavior in the Trust Game. Maybe unsurprisingly, there was no relation between the answers in survey questions and trustor's behavior. There was, however, a correlation between the answer to the GSS question and trustworthiness as measured by trustees' behavior in the Trust Game. Trustor behavior correlated with answers to different questions on placing trust on strangers. Similar results have been obtained by Ashraf et al. (2006), Lazzarini et al. (2005), and Ermisch et al. (2009) (using a representative sample from the BHPS). In contrast, the representative-sample studies of Fehr et al. (2003) (using the German Socio-Economic Panel, SOEP) and Bellemare and Kröger (2007) (using the Dutch CentERpanel) found that survey questions about people's trust and especially past trusting behavior are predictors of trusting behavior. In summary, evidence is mixed, and the relation between the two measures of trust is unclear at this point. A possible preliminary conclusion in view of the evidence is that the Trust Game tests a very specific, strategic situation and trustor behavior might not be a good indicator of the generalized form of trust captured by the GSS question or related ones.

Since our understanding of the relation of various survey measures to behavioral measures of trust is still limited, it might be worth exploring survey questions in populations comparable to those of standard laboratory experiments. For instance, Chaudhuri et al. (2016) replicated one of the context-rich treatments of their experiment (discussed in section 4) and found a significant positive correlation between trustors' transfers and their responses to a five-item questionnaire by Yamagishi (1986), but the questionnaire was not related to the decision to reciprocate. Ben-Ner and Halldorsson (2010) examined both survey measures and behavior in the Trust Game with a sample of university students ($N = 204$). They found a weak but significant correlation between the amount sent by trustors in the Trust Game and the answer to the GSS question. By examining the impact of a battery of measures (including both factors determined at birth and factors determined by attitudes, views, and social preferences) on behavioral and survey measures of trust and trustworthiness, they suggest that trustor behavior in the Trust Game and survey measures of trust might capture different facets of a richer (but unspecified) construct. Trustee behavior might capture trustworthiness for investment situations, but certain survey measures, for instance the Machiavellian scale (Christie and Geis, 1970), appear to reflect different facets of trustworthiness.

Recently, Falk et al. (2018) conducted the *Global Preference Survey (GPS)* eliciting a variety of preference reports from 80,000 people across 76 countries. Trust was measured through one self-report item asking respondents simply whether they assumed that other people have only the best intentions (Likert scale, 0–10). The scale correlated significantly at 0.49 with the GSS question, which could be interpreted as positive evidence on the validity of the latter. The main message, however, is that, as for many other indicators of individual preferences, there is a large heterogeneity in trust, with variation arising from both individual and aggregate characteristics (cultural and geographical). Regarding individual characteristics, unsurprisingly, trust as measured in the GPS correlates with measures of altruism and positive reciprocity, but it also correlates positively with measures of patience and negative reciprocity. In almost all countries, trust increased with cognitive ability. At the purely demographic level, older individuals tended to be more trusting as measured by the GPS, but gender effects where less clear, with differences (in favor of women being more trusting) being significant in only one third of all countries. Interestingly, trust increased in the presence of domestic animals. Concerning aggregate characteristics, trust correlated with latitude, with trust levels being particularly high in the USA, Canada, and Australia. Last, trust was a positive predictor of economic development (as proxied by income per capita), in a country-level regression, but the relation became non-significant when controlling for patience.

In view of the country differences found by Falk et al. (2018), it is interesting to recall the results of Yamagishi and Yamagishi (1994), which called into question previous findings showing higher levels of trust, as measured by survey scales, in the USA than in Japan. This is in contrast with the conventional image of Japanese society, where mutual trust and stable long-term relationships (both social and economic) are highly appreciated. In stark contrast to single-item measures, Yamagishi and Yamagishi (1994) used an extensive (86-item) questionnaire in both countries to examine differences in the concept of trust. They proposed to distinguish a concept of *general trust* from a more specific concept of *assurance.* The former refers to the evaluation of potential partners in the presence of incomplete information and social uncertainty. The latter refers to a need for the reduction of social uncertainty, mostly through the formation of mutual-commitment relations (which might lead to foregoing new opportunities). While Americans scored higher than Japanese in general trust, the opposite was true for assurance. That is, Japanese place a higher value on the long-term aspect of trust, which emphasizes forming lasting, stable relationships. This study serves as a word of caution on cultural differences in the concept of trust.

In conclusion, survey measures appear to be only weakly related (if at all) to behavior in the Trust Game. A possible interpretation of the state of the literature is simply that the various behavioral and survey measures capture different facets of generalized, abstract notions of trust and trustworthiness. Hence, researchers should not assume that any particular behavioral or survey measure available at this point suffices to cover all aspects of our intuitive notions of human trust and trustworthiness.

7. THE NEUROSCIENCE OF TRUST

As illustrated in the previous sections, behavioral (choice) studies have made abundantly clear that trust is a multifaceted concept which interacts with many other aspects of social behavior (which are potential confounds). As a consequence, behavioral and self-report measures might be too simple to capture trust at the individual level in a stable and reliable way. It is natural to ask whether more objective, biological correlates exist. Recent advances in neuroscience point at two natural avenues of research. On the one hand, the hormone oxytocin has been shown to be related to trusting behavior. On the other hand, brain scanning studies are shedding light on the neural correlates of trust.

7.1. Oxytocin and Trust

The neuropeptide hormone oxytocin (OT), synthesized in the hypothalamus, is known to modulate social behavior both in humans (IsHak et al., 2011) and non-human animals (Donaldson and Young, 2008), enabling pair bonding (Young and Wang, 2004) and maternal attachment (Insel and Young, 2001). Indeed, OT has been popularly labeled as "love hormone" or "liquid trust" (Nave et al., 2015).

The literature can be usefully divided into correlational studies, which exploit the endogenous variation in OT levels in blood (Zak et al., 2005), saliva (Tops et al., 2013), or urine (Ebert et al., 2013)[1] and causal studies, which produce exogenous variation by administering OT, via either intranasal (Kosfeld et al., 2005; Baumgartner et al., 2008) or intravenous routes (Hollander et al., 2007; Lee et al., 2018).

Both approaches have been used to link OT levels and behavior in the Trust Game. Zak et al. (2005) was the pioneering correlational study, measuring natural variation of OT levels in blood samples immediately after subject decisions in a Trust Game with $X = \$10$ and $K = 3$ (as in Berg et al., 1995). This was compared to a Random Draw condition where the trustor's transfer was determined as a random integer from 0 to 10. Trustee's OT levels in the Trust Game were 41% higher than in the Random Draw condition. In the Trust Game, the amount sent back by the trustees was a function of the amount sent by the trustor and the log of OT levels (the logarithm form is expected to capture saturation). However, log OT was not statistically significant in the Random Draw condition. Hence, OT levels can be seen as a correlate of reciprocity (trustworthiness), in the sense that they correlate with return transfers but only if those respond to an actual, intentional transfer. However, OT levels did not predict trustors' transfers, that is, the evidence of Zak et al. (2005) refers exclusively to trustee behavior. The basic message of this and other studies is that being treated well (e.g., shown trust) results in OT production, which in turn increases reciprocity. This has motivated a recent "neuromanagement" view (Zak, 2017, 2018) which tries to spell out the possible changes in organizational culture which can (presumably through induced oxytocin release) promote trust and prosocial behavior within the organization.

Studies exploiting endogenous OT variation, though, cannot establish causality. Kosfeld et al. (2005) was the first causal study testing the hypothesis that OT increases trusting behavior in humans using the Trust Game. A single dose of either OT or a placebo was administered intranasally to each study participant. Then they played a Trust Game with $X = 12$ monetary units and $K = 3$, with trustor transfers constrained to being multiples of four (trustee transfers were unconstrained; additionally, trustees had a supplementary endowment of 12 units). The results showed significantly larger trustor transfers for participants who received an OT dose, compared to those who received placebo. In contrast, there were no significant differences in the levels of reciprocity (return transfers from the trustees) between the OT and the placebo group.

Further, the Trust Game in Kosfeld et al. (2005) was contrasted with a later Risk experiment, where the transfer was framed as an individual, risky investment and the returns to the investors were determined by a random device (there was no second player). Specifically, the random device reproduced the distribution of decisions from the trustees in the previous Trust Game experiment, conditional on each trustor transfer level (recall section 3). In this game behavior did not differ between OT participants and placebo ones. Hence, the results suggest that OT causally increased trust (and not simply risk-taking behavior). However, since there was no second player in the risk experiment, one may ask whether OT simply increases prosociality in general. This appears unlikely, since trustees' behavior was unaffected by OT administration. This suggests that OT administration differentially influences trust, but not reciprocity or general prosocial behavior. Further, trustors were asked about their beliefs on the trustee's transfers, and again the OT group and the placebo one did not differ. Thus, OT did not alter trustor's beliefs (e.g., making them more optimistic). Hence, the natural conclusion is that the mechanism by which OT administration increases trusting behavior is a reduction of betrayal aversion (recall section 3).

The latter conclusion was strengthened by the work of Baumgartner et al. (2008) (see also Fehr, 2009), who administered OT or placebo intranasally to participants who took the trustor role in the Trust Game while undergoing functional magnetic resonance imaging (fMRI). The Trust Game was implemented as in Kosfeld et al. (2005), with the only difference that trustees had just the binary option to either betray the trustor by keeping the whole transfer or honor trust by making a return transfer which equalized payoffs (this was made possible by the trustee's endowment of 12 units). The experiment also included a Risk Game with equivalent binary return transfers. In both conditions, return transfers occurred 50% of the time (ensured by using previously-elicited answers from trustees in a pilot experiment). The focus of Baumgartner et al. (2008), however, was on the reaction of trustors to feedback, implemented as follows. Trustors first played 6 Trust Games against different trustees and 6 risk games, in random order. Trustors were then informed that half

[1] See McCullough et al. (2013) for a discussion of the accuracy of the different methods.

of the time there had been no return transfers (with some randomness added if trust or investment happened an odd number of times). After this feedback, they again played 6 Trust Games and 6 risk games. In the post-feedback phase, trust levels (as measured by transfers) decreased in the placebo group, but *not* in the OT group. That is, experiencing previous betrayals did not reduce trust when OT had been administered.

As the discussion above illustrates, there is a fundamental inconsistency between the results arising from the two methods to study the relation between OT and trust. The study of Zak et al. (2005) finds that naturally-occurring levels of OT predict trustee behavior (reciprocity) but not trustor behavior (trust), venturing the explanation that being shown trust increases OT levels, and those lead to reciprocity. However, the latter link is absent in the data of Kosfeld et al. (2005), who found that OT administration increases trustors' transfers (trust) but does not influence trustee behavior. More recently, Nave et al. (2015) cast doubts on previous results. Pooling the data of seven different experiments where OT had been administered intranasally to participants in the Trust Game, the authors found no robust results in the aggregate. However, it is too early to draw conclusions. As cautioned by Nave et al. (2015), failed replications might be linked to technical difficulties with intranasal OT administration, and the effectiveness of the method itself is still not fully established because, at this point, it is not entirely clear how OT reaches the human brain after administration.

The relation between OT and trust or trustworthiness might simply be more complex than initially assumed. Zhong et al. (2012) measured blood OT levels in 1,158 Chinese undergraduates and found evidence for a non-monotonic (U-shaped) relationship. Subjects in the top *and* bottom 20% of the OT distribution made significantly larger transfers both as trustors and as trustees (respectively, 15.6 and 8.3%) than those in the middle 20% of the distribution. However, participants in the study played both roles, which could have led to spillovers. Also, the relation might be context-dependent. Mikolajczak et al. (2010) provided trustors with stereotypical descriptions of trustees which emphasized reliability or lack thereof. Intranasally-administered OT increased trust compared to a placebo group, but only when the trustee was described as reliable.

Other studies have explored the relation between OT and other dimensions of social behavior, beyond the specific concepts of trust and trustworthiness captured by the Trust Game. Theodoridou et al. (2009) find that OT administration increases judgments of trustworthiness and ratings of attractiveness of pictures of faces. Shamay-Tsoory et al. (2009) found that OT increased envy and *Schadenfreude* when observing own payoffs and the (purported) payoffs of another participant, which is, at least conceptually, in contrast with the results reviewed above. Domes et al. (2007) showed that OT improves the ability to infer the mental states of others from pictures of the eyes regions ("Reading the Mind in the Eye Test," Baron-Cohen et al., 2001).

Recently, Marsh et al. (2017) asked German male students to make (costly) donations for either refugees (outgroup) or locals (ingroup) in need using (truthful) vignette-based descriptions. OT administration resulted in higher donations toward both groups, compared to placebo. Participants completed a version of the xenophobia scale of Schweitzer et al. (2005), and those scoring lower (according to a median split) more than doubled their contributions to both ingroup and outgroup under OT, but the peptide had no effect for those scoring higher in the xenophobia scale. Also, those scoring lower donated 31% more to the outgroup than to the ingroup (irrespective of the OT treatment), but again there was no difference for those scoring higher. Importantly, in a second donations round social norms were manipulated by (truthfully) reporting the average donations per vignette from a previous experiment where donations to the outgroup were 19% higher than those to the ingroup (this difference was achieved using reputation pressure). Strikingly, under OT, subjects who scored high in the xenophobia scale donated 74% more to the outgroup when the norm was manipulated, compared to the absence of a communicated norm. That is, neither the norm nor OT administration alone did influence outgroup donations for the high-scores group, but the conjunction of both manipulations was successful. This study effectively illustrates the main takeaways of the recent literature on OT. First, OT modulates a more general aspect of trust than the dimensions studied in the Trust Game. Second, it does so in a nuanced, context-dependent way which interacts with individual differences.

7.2. Neural Indicators of Trust

The decision to trust entails an evaluation of the expected actions that a different person will take in response. That is, one needs to anticipate the reactions of another decision maker, which requires the set of social-cognitive functions known as *Theory of Mind* (ToM; see, e.g., Singer and Tusche, 2014; Alós-Ferrer, 2018). The brain network underlying Theory of Mind is known to be built along a frontal-temporoparietal link, in particular including key areas as the medial prefrontal cortex (mPFC) and the temporoparietal junction (TPJ). Accordingly, the first fMRI study on the Trust Game, McCabe et al. (2001), targeted the mPFC. Participants played a series of binary-action games including mini-Trust Games, either with a human partner (outside the scanner) or against a computer that made stochastic choices following a given distribution. For the players who consistently made more "cooperative" (trusting) decisions, the study found increased mPFC activity when playing against human partners, compared to playing against the computer. In the Trust Game participants played both as trustor and as trustee (in different trials), but brain activity was analyzed in the time window corresponding to either the end of a decision period of a trustor or the end of the waiting period of a trustee (where presumably the participant was thinking about the trustor's action). Hence, it can be assumed that mPFC activity was linked to the decision whether to trust or not.

Krueger et al. (2007) pointed out that the role of the mPFC might be complemented by other, different brain regions depending on the strategies followed to establish trust in repeated interactions. Participants played a repeated, non-anonymous binary mini-Trust Game while alternating their roles as trustor and trustee. Consistently with McCabe et al. (2001) and the need for Theory of Mind, decisions to trust resulted in differential activation of the mPFC (paracingulate cortex), compared to decisions in a control "game" which involved no interpersonal interaction. Also, in alignment with the results discussed in the last subsection, the contrast also revealed differential activation of the septal area (and the adjacent hypothalamus), which contains oxytocin receptors and is involved in the releases of that peptide. Krueger et al. (2007) divided the experiment in two phases, assuming that the earlier and later one would correspond more to partnership building and maintenance, respectively. Participants were classified as defectors and non-defectors, depending on whether their groups experienced some or no defections during play, respectively. In the building-partnership stage (first-mover), non-defectors showed higher mPFC activation than defectors. However, mPFC activity decreased for non-defectors and increased for defectors over the course of the experiment. In the maintaining-partnership stage, non-defectors showed higher activation of the septal area than defectors, while the latter showed higher activation of the Ventral Tegmental Area (VTA) than the former. The VTA is part of the (dopaminergic) reward valuation network of the brain (see, e.g., Daw and Tobler, 2014). The interpretation is that defector and non-defector groups used different strategies. Defectors relied less on Theory of Mind in the building stage, resulting in comparatively lower levels of trust and lower payoffs. This negative reinforcement, through VTA involvement, resulted in later attempts to repair trust, adding up to a *conditional trust* strategy. Non-defectors, in contrast, employed an *unconditional trust* strategy which led to increased social attachment as reflected by activity in the septal area.

The role of the ToM network for trust is by now firmly established, to the point that the effect of various factors influencing trust might be better understood in terms of their effects on this network and the connectivity between its nodes and other brain areas. For example, Engelmann et al. (2019) showed that aversive affect, induced through prolonged periods of threat of shock, reduced trusting behavior in the Trust Game. The study also provides insights on the likely neural mechanisms underlying this result. Aversive affect reduced activity in the TPJ and also reduced functional connectivity between this area and the amygdala, which plays a key role in emotional processing.

It needs to be remarked that a further brain region, the anterior insula, may also play an important role for the decision to trust. This was shown by Aimone et al. (2014), who investigated the neural foundations of betrayal aversion (recall section 3). Participants in the trustor role showed significantly higher anterior insula activation when deciding to trust a human partner, compared to a computerized one, even though in the latter case the trustee payoffs were also received by a human being. In contrast, there was no difference when deciding not to trust. Hence, activity in the anterior insula might be crucial indicator of betrayal aversion. This is in agreement with data showing that the insula plays an important role in social emotions (Singer and Lamm, 2009).

The decision to reciprocate trust, as discussed in previous sections, presents major differences with the decision to trust. Consequently, it would not be justified to assume that the same neural processes underlie both decisions. An example of the specificities of reciprocity is given by King-Casas et al. (2005). In their study, a trustor and a trustee, with fixed roles, played a repeated Trust Game ($X = 20$, $K = 3$) consisting of ten consecutive rounds. Hence, even the trustor's decision involves a reciprocity component, as it can repay or betray the previous trustee decision. When playing in the trustee role, there was increased activity in the striatum (caudate head) when the trustor behaved generously (sending more in response to a previous trustee defection), compared to when the trustor defected (repaying the trustee's previous reciprocity with a decreased transfer). This might be reflecting a signal on the expectation of reciprocal behavior, consistent with current interpretations on the role of the reward prediction error for human decision making (Daw and Tobler, 2014). More generally, while trust might be motivated by the expectation of future reciprocity, reciprocal behavior will be influenced by the experience of trust, and specifically deviations from expectations.

However, the human decision to reciprocate clearly depends on the intentionality of the received transfer. The latter presupposes Theory of Mind. Hence, it would also be surprising to find no overlap between the neural substrates of trust and reciprocity. Van Den Bos et al. (2009) showed that key nodes of the ToM network also play an important role in reciprocity in the Trust Game. Specifically, reciprocity might reflect the interaction of anterior mPFC and the TPJ. In their study, participants played as trustees in a binary mini-Trust Game. Higher anterior mPFC activation was found when participants defected compared to when they reciprocated. In contrast, activity in the TPJ, bilateral insula, and anterior cingulate cortex (ACC) was modulated by individual differences in social preferences as captured by the Social Value Orientation incentivized scale (SVO; Murphy et al., 2011).

In addition to deepening our understanding of the processes underlying trust and reciprocity, recent research in neuroscience also might suggest a possible way of developing more reliable versions of individual heterogeneity in the underlying predispositions. Bellucci et al. (2019) show that (task-free) resting-state functional connectivity (RSFC) predict individual differences in both trust and reciprocity in a one-shot Trust Game. This is less surprising than it might seem at first glance, because the RSFC reflects the activity in the Default Mode Network (DMN), which displays a large overlap with the ToM network, including, e.g., the mPFC and the TPJ (see, e.g., Alós-Ferrer, 2018).

8. CONCLUSION

Trust and reciprocity are complex behavioral phenomena which interact with many other, different aspects of human social behavior. There might be multiple (but not necessarily mutually exclusive) definitions of trust, reflecting cultural, situational, individual, and neural differences. Quite possibly, there might even be disciplinary differences across the social sciences. The Trust Game is an ingenious but highly-stylized experimental paradigm, which has delivered important insights and remains an important benchmark. It is, however, too stylized to provide a complete picture of the nuances behind trust and reciprocity by itself.

The limitations of the Trust Game might be overcome by carefully controlling for known confounds, as prosocial motivations or social risk. Additionally, a number of complementary measures are readily available, even if none of them seems ready to become the new golden standard. Survey measures have their own problems, but are easy to administer and help acquire longitudinal data which are typically beyond reach when using laboratory-based methods. Neurochemical measurements (chiefly oxytocin) offer a different perspective which might open the door to causal interventions, although, in view of mixed results, caution should be advised at this point. Brain imaging studies allow us to identify direct, neural correlates with the potential to ultimately open the black box of why and how trust takes place.

In view of the literature, there is no doubt that the Trust Game will remain an important instrument in the social scientist's toolbox for many years to come. At the same time, that toolbox, and in particular the part used to measure trust and reciprocity, has grown significantly in the recent years, and it is not necessary to arbitrarily restrict attention to a particular instrument.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Ackert, L. F., Church, B. K., and Davis, S. (2011). An experimental examination of the effect of potential revelation of identity on satisfying obligations. *New Zealand Econ. Papers* 45, 69–80. doi: 10.1080/00779954.2011.556071

Aimone, J. A., Houser, D., and Weber, B. (2014). Neural signatures of betrayal aversion: an fMRI study of trust. *Proc. R. Soc. B Biol. Sci.* 281, 1–6. doi: 10.1098/rspb.2013.2127

Alós-Ferrer, C. (2018). A review essay on social neuroscience: can research on the social brain and economics inform each other? *J. Econ. Literat.* 56, 1–31. doi: 10.1257/jel.20171370

Alós-Ferrer, C., and Ritzberger, K. (2016). *The Theory of Extensive Form Games*. Berlin/Heidelberg/New York, NY: Monographs of the Game Theory Society; Springer-Verlag.

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing trust and trustworthiness. *Exp. Econom.* 9, 193–208. doi: 10.1007/s10683-006-9122-4

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY: Basic Books.

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton: Princeton University Press.

Bardsley, N. (2000). Control without deception: individual behaviour in free-riding experiments revisited. *Exp. Econ.* 3, 215–240. doi: 10.1007/BF01669773

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiat. Allied Discipl.* 42, 241–251. doi: 10.1111/1469-7610.00715

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., and Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58, 639–650. doi: 10.1016/j.neuron.2008.04.009

Bellemare, C., and Kröger, S. (2007). On representative social capital. *Eur. Econ. Rev.* 51, 183–202. doi: 10.1016/j.euroecorev.2006.03.006

Bellucci, G., Hahn, T., Deshpande, G., and Krueger, F. (2019). Functional connectivity of specific resting-state networks predicts trust and reciprocity in the trust game. *Cogn. Affect. Behav. Neurosci.* 19, 165–176. doi: 10.3758/s13415-018-00654-3

Ben-Ner, A., and Halldorsson, F. (2010). Trusting and trustworthiness: what are they, how to measure them, and what affects them. *J. Econ. Psychol.* 31, 64–79. doi: 10.1016/j.joep.2009.10.001

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav.* 10, 122–142.

Binmore, K., Gale, J., and Samuelson, L. (1995). Learning to be imperfect: the ultimatum game. *Games Econ. Behav.* 8, 56–90.

Bohnet, I., and Zeckhauser, R. (2004). Trust, risk and betrayal. *J. Econ. Behav. Organ.* 55, 467–484. doi: 10.1016/j.jebo.2003.11.004

Bolton, G. E., and Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. *Am. Econ. Rev.* 90, 166–193. doi: 10.1257/aer.90.1.166

Brandts, J., and Charness, G. (2000). Hot vs. cold: sequential responses and preference stability in experimental games. *Exp. Econ.* 2, 227–238. doi: 10.1007/BF01669197

Brandts, J., and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Exp. Econ.* 14, 375–398. doi: 10.1007/s10683-011-9272-x

Burnham, T., McCabe, K., and Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *J. Econ. Behav. Organ.* 43, 57–73. doi: 10.1016/S0167-2681(00)00108-6

Camerer, C., and Weigelt, K. (1998). Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56, 1–36.

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

Casari, M., and Cason, T. N. (2009). The strategy method lowers measured trustworthy behavior. *Econ. Lett.* 103, 157–159. doi: 10.1016/j.econlet.2009.03.012

Charness, G., and Rabin, M. (2002). Understanding social preferences with simple tests. *Quart. J. Econ.* 117, 817–869. doi: 10.1162/003355302760193904

Chaudhuri, A., Ali Khan, S., Lakshmiratan, A., Py, A.-L., and Shah, L. (2003). Trust and trustworthiness in a sequential bargaining game. *J. Behav. Dec. Mak.* 16, 331–340. doi: 10.1002/bdm.449

Chaudhuri, A., and Gangadharan, L. (2007). An experimental analysis of trust and trustworthiness. *South. Econ. J.* 73, 959–985. doi: 10.2307/20111937

Chaudhuri, A., Li, Y., and Paichayontvijit, T. (2016). What's in a frame? Goal framing, trust and reciprocity. *J. Econ. Psychol.* 57, 117–135. doi: 10.1016/j.joep.2016.09.005

Christie, R., and Geis, F. L. (1970). *Studies in Machiavellianism*. New York, NY: Academic Press.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games Econ. Behav.* 46, 260–281. doi: 10.1016/S0899-8256(03)00119-2

Cronk, L. (2007). The influence of cultural framing on play in the trust game: a Maasai example. *Evol. Hum. Behav.* 28, 352–358. doi: 10.1016/j.evolhumbehav.2007.05.006

Dasgupta, P. (1988). "Trust as a commodity," in *Trust: Making and Breaking Cooperative Relations*, ed D. Gambetta (Oxford; Cambridge, MA: Blackwell), 49–72.

Daw, N. D., and Tobler, P. (2014). "Value learning through reinforcement: the basics of dopamine and reinforcement learning," in *Neuroeconomics: Decision Making and the Brain, 2nd Edn*, eds P. W. Glimcher and E. Fehr (London: Academic Press), 283–298.

Domes, G., Heinrichs, M., Michel, A., Berger, C., and Herpertz, S. C. (2007). Oxytocin Improves "Mind-Reading" in Humans. *Biological Psychiatry*, 61(6):731–733.

Donaldson, Z. R., and Young, L. J. (2008). Oxytocin, Vasopressin, and the Neurogenetics of Sociality. *Science*, 322(5903):900–904.

Ebert, A., Kolb, M., Heller, J., Edel, M.-A., Roser, P., and Brüne, M. (2013). Modulation of Interpersonal Trust in Borderline Personality Disorder by Intranasal Oxytocin and Childhood Trauma. *Social Neuroscience*, 8(4):305–313.

Engelmann, J. B., Meyer, F., Ruff, C. C., and Fehr, E. (2019). The neural circuitry of affect-induced distortions of trust. *Sci. Adv.* 5, 1–16. doi: 10.1126/sciadv.aau3413

Ermisch, J., and Gambetta, D. (2006). *People's Trust: The Design of a Survey-Based Experiment*. IZA Discussion Papers 2216, Institute for the Study of Labor.

Ermisch, J., Gambetta, D., Laurie, H., Siedler, T., and Uhrig, S. C. N. (2009). Measuring people's trust. *J. R. Stat. Soc. Ser. A* 172, 749–769. doi: 10.1111/j.1467-985X.2009.00591.x

Fairley, K., Sanfey, A., Vyrastekova, J., and Weitzel, U. (2016). Trust and risk revisited. *J. Econ. Psychol.* 57, 74–85. doi: 10.1016/j.joep.2016.10.001

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *Quart. J. Econ.* 133, 1645–1692. doi: 10.1093/qje/qjy013

Fehr, E. (2009). On the economics and biology of trust. *J. Eur. Econ. Assoc.* 7, 235–266. doi: 10.1162/JEEA.2009.7.2-3.235

Fehr, E., Fischbacher, U., Von Rosenbladt, B., Schupp, J., and Wagner, G. G. (2003). *A Nation-Wide Laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Surveys.* CESifo Working Paper Series.

Fehr, E., and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90, 980–994. doi: 10.1257/aer.90.4.980

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quart. J. Econ.* 108, 437–459.

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quart. J. Econ.* 114, 817–868.

Fetchenhauer, D., and Dunning, D. (2012). Betrayal aversion versus principled trustfulness – how to explain risk avoidance and risky choices in trust games. *J. Econ. Behav. Organ.* 81, 534–541. doi: 10.1016/j.jebo.2011.07.017

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games Econ. Behav.* 6, 347–369.

Fudenberg, D., and Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.

Fukuyama, F. (1995). *Trust*. New York, NY: Free Press Paperbacks.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *Quart. J. Econ.* 115, 811–846. doi: 10.1162/003355300554926

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* 3, 367–388.

Hollander, E., Bartz, J., Chaplin, W., Phillips, A., Sumner, J., Soorya, L., et al. (2007). Oxytocin Increases Retention of Social Cognition in Autism. *Biol. Psychiatry* 61, 498–503. doi: 10.1016/j.biopsych.2006.05.030

Holt, C. A., and Laury, S. K. (2002). Risk aversion and incentive effects. *Am. Econ. Rev.* 92, 1644–1655. doi: 10.1257/000282802762024700

Houser, D., Schunk, D., and Winter, J. (2010). Distinguishing trust from risk: an anatomy of the investment game. *J. Econ. Behav. Organ.* 74, 72–81. doi: 10.1016/j.jebo.2010.01.002

Insel, T. R., and Young, L. J. (2001). The neurobiology of attachment. *Nat. Rev. Neurosci.* 2, 129–136. doi: 10.1038/35053579

IsHak, W. W., Kahloon, M., and Fakhry, H. (2011). Oxytocin role in enhancing well-being: a literature review. *J. Affect. Disorders* 130, 1–9. doi: 10.1016/j.jad.2010.06.001

Johnson, N. D., and Mislin, A. A. (2011). Trust games: a meta-analysis. *J. Econ. Psychol.* 32, 865–889. doi: 10.1016/j.joep.2011.05.007

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

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

Kreps, D. M. (1990). "Corporate culture and economic theory," in *Perspectives on Positive Political Economy*, eds J. E. Alt and K. A. Shepsle (Cambridge: Cambridge University Press), 90–143.

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

Lazzarini, S. G., Madalozzo, R., Artes, R., and de Oliveira Siqueira, J. (2005). Measuring trust: an experiment in Brazil. *Brazil. J. Appl. Econ.* 9, 153–169.

Lee, M. R., Scheidweiler, K. B., Diao, X., Akhlaghi, F., Cummins, A., Huestis, M. A., et al. (2018). Oxytocin by intranasal and intravenous routes reaches the cerebrospinal fluid in rhesus macaques: determination using a novel oxytocin assay. *Mol. Psychiatry* 23, 115–122. doi: 10.1038/mp.2017.27

Lenton, P., and Mosley, P. (2011). Incentivising trust. *J. Econ. Psychol.* 32, 890–897. doi: 10.1016/j.joep.2011.07.005

Lundmark, S., Gilljam, M., and Dahlberg, S. (2015). Measuring generalized trust: an examination of question wording and the number of scale points. *Public Opin. Quart.* 80, 26–43. doi: 10.1093/poq/nfv042

Lyons, B. R., and Mehta, J. (1997). Contracts, opportunism and trust: self-interest and social orientation. *Cambridge J. Econ.* 21, 239–257.

Marsh, N., Scheele, D., Feinstein, J. S., Gerhardt, H., Strang, S., Maier, W., et al. (2017). Oxytocin-enforced norm compliance reduces xenophobic outgroup rejection. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9314–9319. doi: 10.1073/pnas.1705853114

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11832–11835. doi: 10.1073/pnas.211415698

McCullough, M. E., Churchland, P. S., and Mendez, A. J. (2013). Problems with measuring peripheral oxytocin: can the data on oxytocin and human behavior be trusted? *Neurosci. Biobehav. Rev.* 37, 1485–1492. doi: 10.1016/j.neubiorev.2013.04.018

Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., and Luminet, O. (2010). Oxytocin makes people trusting, not gullible. *Psychol. Sci.* 21, 1072–1074. doi: 10.1177/0956797610377343

Miller, A. S., and Mitamura, T. (2003). Are surveys on trust trustworthy? *Soc. Psychol. Quart.* 66, 62–70. doi: 10.2307/3090141

Mislin, A., Williams, L. V., and Shaughnessy, B. A. (2015). Motivating trust: can mood and incentives increase interpersonal trust? *J. Behav. Exp. Econ.* 58, 11–19. doi: 10.1016/j.socec.2015.06.001

Miyaji, K., Wang, Z., Tanimoto, J., Hagishima, A., and Kokubo, S. (2013). The evolution of fairness in the coevolutionary ultimatum games. *Chaos Sol. Fract.* 56, 13–18. doi: 10.1016/j.chaos.2013.05.007

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment Dec. Mak.* 6, 771–781. doi: 10.2139/ssrn.1804189

Nave, G., Camerer, C., and McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspect. Psychol. Sci.* 10, 772–789. doi: 10.1177/1745691615600138

Nowak, M., and Sigmund, K. (1993). Chaos and the evolution of cooperation. *Proc. Natl. Acad. Sci. U.S.A.* 90, 5091–5093.

Poundstone, W. (1993). *Prisoner's Dilemma*. New York, NY: Anchor Books.

Rosenberg, M. (1956). Misanthropy and political ideology. *Am. Soc. Rev.* 21, 690–695.

Schweitzer, R., Perkoulidis, S., Krome, S., Ludlow, C., and Ryan, M. (2005). Attitudes towards refugees: the dark side of prejudice in Australia. *Aust. J. Psychol.* 57, 170–179. doi: 10.1080/00049530500125199

Selten, R. (1967). "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments," in *Beiträge zur experimentellen Wirtschaftsforschung*, ed H. Sauermann [Tübingen: J. C. B. Mohr (Paul Siebeck)], 136–168.

Shamay-Tsoory, S. G., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N., and Levkovitz, Y. (2009). Intranasal administration of oxytocin increases envy and schadenfreude (gloating). *Biol. Psychiatry* 66, 864–870. doi: 10.1016/j.biopsych.2009.06.009

Singer, T., and Lamm, C. (2009). The social neuroscience of empathy. *Ann. NY Acad. Sci.* 1156, 81–96. doi: 10.1111/j.1749-6632.2009.04418.x

Singer, T., and Tusche, A. (2014). "Understanding others: brain mechanisms of theory of mind and empathy," in *Neuroeconomics: Decision Making and the Brain, 2nd Edn*, eds P. W. Glimcher and E. Fehr (London: Academic Press), 513–532.

Sun, H., Verbeke, W. J. M. I., Pozharliev, R., Bagozzi, R. P., Babiloni, F., and Wang, L. (2019). Framing a trust game as a power game greatly affects interbrain synchronicity between trustor and trustee. *Soc. Neurosci.* doi: 10.1080/17470919.2019.1566171. [Epub ahead of print].

Tanimoto, J. (2015). *Fundamentals of Evolutionary Game Theory and its Applications*. Tokyo: Springer.

Tanimoto, J. (2019). *Evolutionary Games With Sociophysics*. Tokyo: Springer.

Theodoridou, A., Rowe, A. C., Penton-Voak, I. S., and Rogers, P. J. (2009). "Oxytocin and social perception: oxytocin increases perceived facial trustworthiness and attractiveness. *Horm. Behav.* 56, 128–132. doi: 10.1016/j.yhbeh.2009.03.019

Tops, M., Huffmeijer, R., Linting, M., Grewen, K. M., Light, K. C., Koole, S. L., et al. (2013). The role of oxytocin in familiarization-habituation responses to social novelty. *Front. Psychol.* 4:761. doi: 10.3389/fpsyg.2013.00761

Uslaner, E. M. (2012). "Measuring generalized trust: in defense of the "Standard" question. in *Handbook of Research Methods on Trust*, eds F. Lyon, M. Guido, and S. Mark (Cheltenham; Northampton, MA: Edward Elgar Publishing), 72–82.

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

von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge; Massachusetts: The MIT Press.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *J. Personal. Soc. Psychol.* 51, 110–116.

Yamagishi, T., and Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Mot. Emot.* 18, 129–166.

Young, L. J., and Wang, Z. (2004). The neurobiology of pair bonding. *Nat. Neurosci,* 7, 1048–1054. doi: 10.1038/nn1327

Zak, P. J. (2017). The neuroscience of trust. *Harvard Business Rev.* 95, 84–90.

Zak, P. J. (2018). The neuroscience of high-trust organizations. *Consult. Psychol. J. Pract. Res.* 70, 45–58. doi: 10.1037/cpb0000076

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

Zhong, S., Monakhov, M., Mok, H. P., Tong, T., San Lai, P., Chew, S. H., et al. (2012). U-shaped relation between plasma oxytocin levels and behavior in the trust game. *PLoS ONE* 7:e51095. doi: 10.1371/journal.pone.0051095

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

frontiers
in Human Neuroscience

OPINION
published: 15 May 2020
doi: 10.3389/fnhum.2020.00176

The Critical Role of the Right Dorsal and Ventral Anterior Insula in Reciprocity: Evidence From the Trust and Ultimatum Games

Frank Krueger[1,2*†], *Gabriele Bellucci*[3†], *Pengfei Xu*[4,5,6] *and Chunliang Feng*[7*]

[1] Department of Psychology, George Mason University, Fairfax, VA, United States, [2] School of Systems Biology, George Mason University, Fairfax, VA, United States, [3] Max Planck Institute for Biological Cybernetics, Tübingen, Germany, [4] Shenzhen Key Laboratory of Affective and Social Neuroscience, Center for Brain Disorders and Cognitive Sciences, Shenzhen University, Shenzhen, China, [5] Center for Neuroimaging, Shenzhen Institute of Neuroscience, Shenzhen, China, [6] Great Bay Neuroscience and Technology Research Institute, Kwun Tong, Hong Kong, [7] Guangdong Provincial Key Laboratory of Mental Health and Cognitive Science, Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China

Keywords: trust, reciprocity, insula, decision-making, network

OPEN ACCESS

Edited by:
Jessica A. Turner,
Georgia State University,
United States

Reviewed by:
Khalil Thompson,
Georgia State University,
United States
Jan B. Engelmann,
University of Amsterdam, Netherlands

****Correspondence:***
Frank Krueger
FKrueger@gmu.edu
Chunliang Feng
Chunliang.Feng@m.scnu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to
Cognitive Neuroscience,
a section of the journal
Frontiers in Human Neuroscience

Received: *16 January 2020*
Accepted: *20 April 2020*
Published: *15 May 2020*

Citation:
Krueger F, Bellucci G, Xu P and Feng C (2020) The Critical Role of the Right Dorsal and Ventral Anterior Insula in Reciprocity: Evidence From the Trust and Ultimatum Games. Front. Hum. Neurosci. 14:176. doi: 10.3389/fnhum.2020.00176

Social norms represent a fundamental grammar of social interactions, as they refer to shared expectations about behaviors of one's social group members (Bicchieri, 1990, 2005; Santos et al., 2018). Based on these expectations, particularly accurate predictions about another person's future behavior are possible—establishing the preconditions for cooperative interactions. Overall, group prosperity is enhanced when all members comply with social norms (i.e., *norm compliance*). However, social norms need to be enforced by sanctioning violators (i.e., *norm enforcement*). For instance, expectations of compliance with a norm of reciprocity may help overcome the fear of being betrayed by a social partner. As cooperation allows for better collective solutions than those attained by self-interested individuals, social groups are interested in enforcing compliance with social norms by their members, and developing tools for successful recognition of norm violators (Fehr and Schurtenberger, 2018). Thus, a fragile balance between incentives for norm enforcement and deterrents for sanctions of violators is required for a well-functioning society.

Interactive economic games, such as the trust game (TG) (Berg et al., 1995) and the ultimatum game (UG) (Güth et al., 1982), provide reliable experimental settings for the investigation of motivational, affective, and socio-cognitive processes involved in social norm compliance and enforcement (Corradi-Dell'acqua et al., 2016; Feng et al., 2017; Engelmann et al., 2019; Krueger and Meyer-Lindenberg, 2019). Based on the learned and internalized social norms, an agent's *reciprocal* behavior is determined by the evaluation of the expected or experienced *kindness* of a partner by weighting the partner's *intentions* (i.e., the underlying motivation in performing an action) and the action *outcomes* (i.e., positive or negative consequences of an action for oneself and others) (Falk and Fischbacher, 2006).

Recent work has shown that individuals integrate this information into their beliefs about another person's character traits for reliable predictions of the other's most likely behavior in a new social interaction (Krueger et al., 2009; Bellucci et al., 2019b; Dorfman et al., 2019). Hence, reliably estimating the kindness/unkindness of a partner facilitates norm compliance (i.e., positive reciprocity) or norm enforcement (i.e., negative reciprocity) across contexts and time. Importantly, the ability to learn from feedback about a partner's intentions and action outcomes heavily hinges on the degree to which feedback information violates one's priors and expectations (Fouragnan et al., 2013; Dorfman et al., 2019; Bellucci and Park, 2020). The ability to detect expectancy violations might even counteract biases in belief updating about another person's benevolence or malevolence.

Integrating neuroimaging data from economic games across a plethora of neuroimaging studies via coordinate-based meta-analyses (Feng et al., 2015; Bellucci et al., 2017a)—in combination with task-based and task-free functional connectivity analyses (Gurevitch et al., 2018)—has revealed the right anterior insula (R AI) as a candidate brain region for detection of norm deviations in trusting (i.e., trust game) and fairness-related (i.e., ultimatum game) interactions (Krueger et al., 2008; Bellucci et al., 2018). Representing a posterior-to-anterior remapping of interoceptive signals within the insular cortex, the R AI takes a crucial role in salience detection across multiple domains, whereas the posterior insular cortex mediates sensorimotor processes (Craig, 2009). Being part of the salience network (SAN), two functionally distinct brain regions within the R AI—a dorsal AI (dAI) and ventral AI (vAI) cluster—have been identified Kelly et al., 2012; Chang et al., 2013; Wager and Barrett, 2017). Whereas the R dAI act as a switch that exerts direct influences on the central executive network (CEN, i.e., cognitive control system, including high-order executive functions; Seeley et al., 2007; Bressler and Menon, 2010; Menon, 2011; Sheffield et al., 2015 and the default-mode network (DMN, i.e., social cognition system, including autobiographical memory, self-monitoring, and theory of mind; Andrews-Hanna et al., 2010; Bressler and Menon, 2010; Menon, 2011), the R vAI exerts direct influence on limbic cortices (which mediate affective processes) (Sridharan et al., 2008; Goulden et al., 2014; Uddin et al., 2014). These AI subregions—encoding a *common currency of aversion*—were both found consistently activated for responses to unfair behavior but differently engaged by trust and trustworthiness behaviors (Bellucci et al., 2018). In particular, the dAI was preferentially engaged by trust behavior while the vAI by trustworthiness behavior (Bellucci et al., 2018). We propose that consistent recruitment of the AI during those social behaviors is a signature of their common neural processing related to expectancy violation in the form of deviations from social norms. In particular, social behaviors in the TG and UG, such as trust in unknown partners, trustworthiness during repeated interactions and rejection of unfair offers, imply violations of two fundamental social norms —fairness and reciprocation. With this respect, they require evaluations of intentions and outcomes of actions that are aligned with individual expectations in case of compliant behaviors but that deviate from individual expectations in case of violations.

When interacting with a stranger in a one-shot TG, in which the investor interacts only once with a trustee, investors feel compelled to comply with a fairness norm and share some fair amount with the trustee. However, the probability that the trustee, whose reputation and past social behavior are supposedly unknown, betrays trust in these circumstances is not negligible. Behavioral studies have repeatedly shown that individuals in these situations worry about a hypothetical, but not much unlikely, defection to occur (Mccabe et al., 1998; Bohnet and Zeckhauser, 2004; Ashraf et al., 2006; Bohnet et al., 2008; Aimone and Houser, 2011, 2013). Individuals might hence begin prospecting to decide whether to trust, for instance, by thinking about what would be most likely that the partner thinks about compliance with a reciprocity norm, and about the reasons for which the partner would consider convenient to violate this norm—processes that likely require the recruitment of the dAI. In iterative interactions, on the contrary, individuals are likely to base their trust decisions on what they have learned from the partner over multiple encounters, switching to a more automatic, knowledge-based decision-making process involving social affiliation regions (Krueger et al., 2007). This is further consistent with the absence of AI signaling during iterative trust decisions with the same partner (Bellucci et al., 2017a).

Reciprocation of trust requires similar evaluations of norm-deviant behaviors by the trustee in a multi-round TG. The concerns that investors have from a second-person perspective, trustees have those from a first-person perspective. In particular, trustees have to weigh the advantages and disadvantages of a cooperative and non-cooperative response to the investor's kind behavior. Also, as the amount of money entrusted by investors in the TG is multiplied by a predetermined factor (usually, tripled), trustees are in an advantageous situation in which defection lures with its convenience. However, defection also implies the violation of a reciprocation norm that will enforce inequality in the payoff distribution between investors and trustees. Hence, trustees might feel guilty of taking advantage of their situation and might fear of what the partner could think of them, especially in iterative interactions where future encounters loom and the importance of a good reputation is more pressing. These aversive feelings are likely encoded in the vAI. On the contrary, in circumstances of low external incentives, such as during reciprocal decisions in single interactions where concerns about what others might think and the pressure of social norm compliance are absent, cognitive control might be required to enact reciprocity. This nicely chimes with the recruitment of dorsolateral prefrontal regions during trustworthiness behavior in single and anonymous interactions (Knoch et al., 2006; Van Den Bos et al., 2011; Nihonsugi et al., 2015).

The receiver in the UG, who faces an unfair offer from the proposer, is in a situation that likely elicits similar psychological processes to those evoked by both investors' and trustees' concerns in the TG. On the one hand, the receiver is confronted with an actual violation of the fairness norm perpetrated by the proposer who sent an unfair offer. Unfair offers elicit negative feelings (e.g., increases in skin conductance activity) in receivers who respond by rejecting the offer. Since the unfair offer implies an actual inequal outcome in resource distributions (given that unfair offers are generally lower than one-third of the resources available to proposers), the receiver might be concerned about the inequality derived from the norm violation. Outcome inequality might hence evoke negative feelings in the receiver that support negative reciprocity via recruitment of the vAI. On the other hand, however, high rejection rates and increased skin conductance activity have been observed only for unfair offers proposed by a human partner, but not for unfair offers generated by computers (Sanfey et al., 2003; Van 'T Wout et al., 2006). These results suggest that the receiver in the UG is further concerned about the intentions of the proposer and is determined to forgo immediate benefits to enforce a fairness norm via a rejection of the offer, which likely recruits the dAI.

Hence, consistent activations of the AI in all these behaviors likely refer to general signaling of violations of expectations about actions that deviate from social norms. However, given the different activation patterns of the dAI and vAI, we here propose an overarching framework in which the R AI—part of the salience network (SAN)—recruits other large-scale brain networks to determine the appropriate reciprocal behavior (via the central-executive network, CEN) based on evaluations about the partner's kindness (via the default-mode network, DMN) (Krueger and Hoffman, 2016; Bellucci et al., 2019a). Hereby, the R AI subregions play a crucial role in signaling how a deviation has occurred, in particular, because of an intentional action (R dAI) or due to an action outcome (R vAI; **Figure 1**).

We propose that the SAN detects (vAI) and generates an aversive experience based on the salience of the social norm violation and provides an emotional signal (amygdala) encoding the severity of outcome related to the norm violation (Buckholtz et al., 2008). The DMN anchored in the medial prefrontal cortex (mPFC) integrates the outcome (via the ventromedial PFC's inter-network connectivity with SAN) and the intention (via the dorsomedial PFC's intra-network connectivity with the temporoparietal junction, TPJ) of the norm violation into an assessment of kindness (Krueger et al., 2009). The CEN anchored in lateral PFC (lPFC) converts the kindness signal from the DMN into an appropriate reciprocal behavior that fits the norm violation. Previous work has demonstrated that connectivity between the mPFC and lPFC was associated with evaluations of norm violations for appropriate punishment decisions (Bellucci et al., 2017b).

Therefore, in the social settings of the economic paradigms here considered, the vAI likely represents forms of *violations of an expected outcome* such as outcome inequalities (i.e., less- vs.

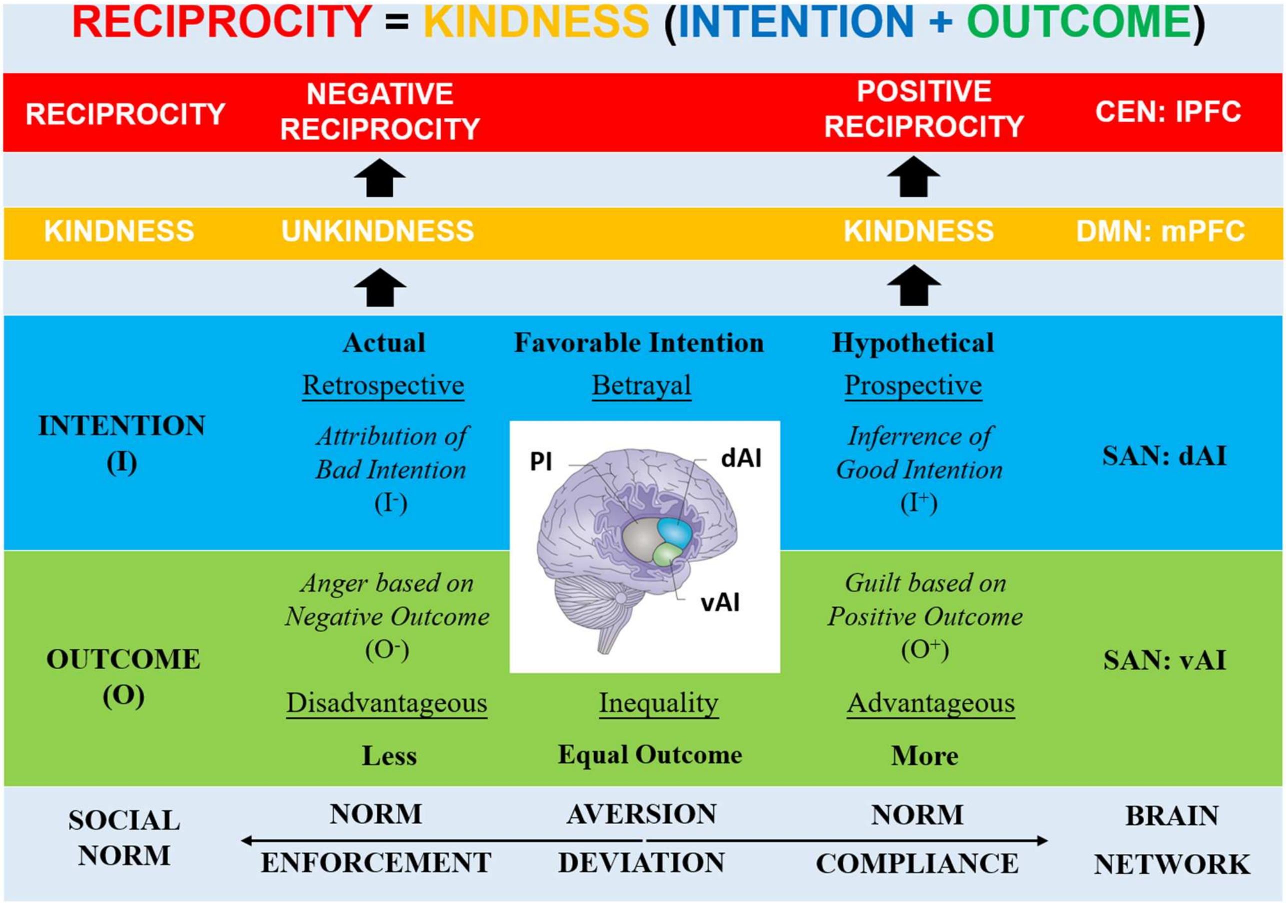


FIGURE 1 | Framework: Role of R dAI and R vAI in Reciprocity. Based on social norms, an agent's reciprocal behavior is determined by evaluating the expected or experienced kindness/unkindness of a partner's normative action: the intention as the underlying motivation and the outcome as the consequence of the action. The R AI (part of SAN) recruits other large-scale networks to determine the appropriate reciprocity (e.g., lPFC via CEN) based on kindness evaluations (e.g., mPFC via DMN). The R AI subregions play a crucial role in signaling deviations from expectations on outcomes (R vAI) and intentions (R dAI) of an action, facilitating norm compliance (positive reciprocity), and norm enforcement (negative reciprocity). The vAI signals violations of expected outcomes (disadvantageous vs. advantageous outcome inequality) that elicit aversive feelings (anger vs. guilt). The dAI signals violations of expected intentional behaviors (actual vs. hypothetical betrayal) that evoke social-cognitive processes (attribution vs. inference) [Note that brain image adopted from Uddin (2015)]. R, right; SAN, Salience Network; PI, Posterior Insula; AI, Anterior Insula; vAI, Ventral Anterior Insula; dAI, Dorsal Anterior Insula; DMN, Default-mode Network; mPFC, Medial Prefrontal Cortex; CEN, Central-executive Network, lPFC, Lateral Prefrontal Cortex; O+, Positive Outcome; O-, Negative Outcome; I-, Negative Intention; I+, Positive Intention.

more-than-equal) that elicit negative feelings via co-activation of the limbic network (e.g., amygdala). In particular, less-than-equal outcomes refer to a situation of *disadvantageous inequality* that triggers negative feelings such as anger and envy (due to a negative outcome for the self), which support norm enforcement in the form of negative reciprocity (e.g., punishment). On the contrary, more-than-equal outcomes refer to situations of *advantageous inequality* that likely triggers different negative feelings such as guilt (due to a positive outcome for the self), which compel to norm compliance in the form of positive reciprocity (e.g., cooperation).

The dAI, instead, likely represents forms of *violations of an expected intentional behavior* such as betrayal (both actual and hypothetical) that elicit social-cognitive processes via co-activation of the default-mode network. In particular, actual deviant behaviors prompt to retrospection on the intentionally perpetrated betrayal that triggers socio-cognitive processes such as attribution of bad intentions, thereby promoting norm enforcement in the form of negative reciprocity (e.g., punishment). On the contrary, hypothetical deviant behaviors prompt to prospection on a possible intentional betrayal that triggers socio-cognitive processes such as inferences on the other's intentions, thereby supporting norm compliance in the form of positive reciprocity (e.g., trust).

Given the proposed neuropsychological model, some predictions for other recently reported activation patterns associated with social normative behaviors are possible. For instance, social interactions in which some form of expectancy violation is involved might require recruitment of the AI. For the classical Prisoner's Dilemma game, where two players can decide to cooperate or betray each other, both parties—acting in their own self-interests—choose often to protect themselves at the expense of the other player; thereby, producing the worst outcome for both parties by non-reciprocation of cooperation (Peterson, 2015). A neuroimaging study employing an iterated version of the Prisoner's Dilemma game showed greater activation in R dAI during unreciprocated compared to reciprocated cooperation when both players were informed about the outcome of each trial game (but not during their decisions) (Rilling et al., 2008). Another study revealed that depressed compared to healthy individuals reported higher levels of negative feelings (i.e., betrayal, guilt) during this game. Across all players, the R vAI was more activated comparing outcomes, where one of the players cooperated and the other defected, with outcomes, where both players either cooperated or defected (Gradin et al., 2016).

Further, shame and embarrassment, which emerge from the recognition that one's behavior diverges from a group's expectancies, should elicit activations in the AI. Preliminary evidence aligns with this prediction and suggests that shame and embarrassment elicit activations particularly in the vAI, consistently with the fact that these negative feelings are based on violations caused by the consequences (and not the intentions) of one's behavior (Muller-Pinzler et al., 2015; Zhu et al., 2019). Similarly, punishment and blame, which rely on the recognition of another's deviant behavior, should recruit the AI as well. Previous evidence chimes with this prediction, pointing specifically to the dAI, consistently with the fact that punishment and blame require socio-cognitive processes for understanding reasons and motives of another's wrongdoing (Krueger and Hoffman, 2016; Patil et al., 2017; Bellucci et al., 2020). On the contrary, other social behaviors such as generosity or altruism should activate the AI only if they also involve expectancy violations. Previous work on these behaviors seems to confirm such prediction (Moll et al., 2006; Coll et al., 2017; Karns et al., 2017), showing AI activations only when a form of expectancy violation is involved such as when helping an offender or breaking a promise to cooperate (Baumgartner et al., 2009; David et al., 2017).

Taken together, the AI is an underestimated but essential brain region for understanding human social cognition and its pathophysiological forms in social brain disorders such as schizophrenia and autism (Namkung et al., 2017). Our framework provides a distinctive mapping of the R AI subdivisions that can be employed in future multimodal neuroimaging studies to test hypotheses on the AI functioning in reciprocity. For this reason, our neuropsychological framework contributes to a more comprehensive understanding of this region for basic and clinical neuroscience in which altered processing in AI subdivisions determine different aspects of prevalent brain disorders (e.g., psychosis, autism).

AUTHOR CONTRIBUTIONS

FK and CF prepared the first draft of the article. All authors contributed to the final version.

FUNDING

PX was supported by the National Natural Science Foundation of China (31871137), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2018), Guangdong Pearl River Talents Plan Innovative and Entrepreneurial Team grant (2016ZT06S220), Shenzhen Science and Technology Research Funding Program (JCYJ20180507183500566), and Shenzhen Peacock Program (827-000235, KQTD2015033016104926). CF was supported by the funding from the National Natural Science Foundation of China (31900757). FK was supported by the Air Force Office of Scientific Research (FA9550-18-1-0455).

REFERENCES

Aimone, J. A., and Houser, D. (2011). Beneficial betrayal aversion. *PLoS ONE* 6:e17725. doi: 10.1371/journal.pone.0017725

Aimone, J. A., and Houser, D. (2013). Harnessing the benefits of betrayal aversion. *J. Economic Behav. Organiz.* 89, 1–8. doi: 10.1016/j.jebo.2013.02.001

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron* 65, 550–562. doi: 10.1016/j.neuron.2010.02.005

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing trust and trustworthiness. *Exp. Econom.* 9, 193–208. doi: 10.1007/s10683-006-9122-4

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., and Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron* 64, 756–770. doi: 10.1016/j.neuron.2009.11.017

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., and Krueger, F. (2020). The emerging neuroscience of social punishment: meta-analytic evidence. *Neurosci. Biobehav. Rev*. 6:e1000097. doi: 10.1016/j.neubiorev.2020.04.011

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K. M., et al. (2017b). Effective connectivity of brain regions underlying third-party punishment: functional MRI and Granger causality evidence. *Soc. Neurosci.* 12, 124–134. doi: 10.1080/17470919.2016.1153518

Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., and Krueger, F. (2017a). Neural signatures of trust in reciprocity: a coordinate-based meta-analysis. *Hum Brain Mapp* 38, 1233–1248. doi: 10.1002/hbm.23451

Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S. B., and Krueger, F. (2018). The role of the anterior insula in social norm compliance and enforcement: evidence from coordinate-based and functional connectivity meta-analyses. *Neurosci. Biobehav. Rev.* 92, 378–389. doi: 10.1016/j.neubiorev.2018.06.024

Bellucci, G., Hahn, T., Deshpande, G., and Krueger, F. (2019a). Functional connectivity of specific resting-state networks predicts trust and reciprocity in the trust game. *Cogn. Affect. Behav. Neurosci.* 19, 165–176. doi: 10.3758/s13415-018-00654-3

Bellucci, G., Molter, F., and Park, S. Q. (2019b). Neural representations of honesty predict future trust behavior. *Nat. Commun.* 10:5184. doi: 10.1038/s41467-019-13261-8

Bellucci, G., and Park, S. Q. (2020). Honesty biases trustworthiness impressions. *J. Exp. Psychol. Gen*. doi: 10.1037/xge0000730. [Epub ahead of print].

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

Bicchieri, C. (1990). Norms of cooperation. *Ethics* 100, 838–861. doi: 10.1086/293237

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Bohnet, I., Greig, F., Herrmann, B., and Zeckhauser, R. (2008). Betrayal aversion: evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *Am. Economic Rev.* 98, 294–310. doi: 10.1257/aer.98.1.294

Bohnet, I., and Zeckhauser, R. (2004). Trust, risk and betrayal. *J. Economic Behav. Organization* 55, 467–484. doi: 10.1016/j.jebo.2003.11.004

Bressler, S. L., and Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* 14, 277–290. doi: 10.1016/j.tics.2010.04.004

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron* 60, 930–940. doi: 10.1016/j.neuron.2008.10.016

Chang, L. J., Yarkoni, T., Khaw, M. W., and Sanfey, A. G. (2013). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cereb. Cortex* 23, 739–749. doi: 10.1093/cercor/bhs065

Coll, M. P., Gregoire, M., Eugene, F., and Jackson, P. L. (2017). Neural correlates of prosocial behavior towards persons in pain in healthcare providers. *Biol. Psychol.* 128, 1–10. doi: 10.1016/j.biopsycho.2017.06.005

Corradi-Dell'acqua, C., Tusche, A., Vuilleumier, P., and Singer, T. (2016). Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nat. Commun.* 7:10904. doi: 10.1038/ncomms10904

Craig, A. D. (2009). How do you feel–now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

David, B., Hu, Y., Kruger, F., and Weber, B. (2017). Other-regarding attention focus modulates third-party altruistic choice: An fMRI study. *Sci. Rep.* 7:43024. doi: 10.1038/srep43024

Dorfman, H. M., Bhui, R., Hughes, B. L., and Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychol. Sci.* 30, 516–525. doi: 10.1177/0956797619828724

Engelmann, J. B., Meyer, F., Ruff, C. C., and Fehr, E. (2019). The neural circuitry of affect-induced distortions of trust. *Sci. Adv.* 5:eaau3413. doi: 10.1126/sciadv.aau3413

Falk, A., and Fischbacher, U. (2006). A theory of reciprocity. *Games Economic Behav.* 54, 293–315. doi: 10.1016/j.geb.2005.03.001

Fehr, E., and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nat. Hum. Behav.* 2, 458–468. doi: 10.1038/s41562-018-0385-5

Feng, C., Azarian, B., Ma, Y., Feng, X., Wang, L., Luo, Y. J., et al. (2017). Mortality salience reduces the discrimination between in-group and out-group interactions: a functional MRI investigation using multi-voxel pattern analysis. *Hum. Brain Mapp.* 38, 1281–1298. doi: 10.1002/hbm.23454

Feng, C., Luo, Y.-J., and Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. *Hum. Brain Mapping* 36, 591–602. doi: 10.1002/hbm.22649

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

Goulden, N., Khusnulina, A., Davis, N. J., Bracewell, R. M., Bokde, A. L., Mcnulty, J. P., et al. (2014). The salience network is responsible for switching between the default mode network and the central executive network: replication from DCM. *Neuroimage* 99, 180–190. doi: 10.1016/j.neuroimage.2014.05.052

Gradin, V. B., Perez, A., Macfarlane, J. A., Cavin, I., Waiter, G., Tone, E. B., et al. (2016). Neural correlates of social exchanges during the Prisoner's Dilemma game in depression. *Psychol. Med*. 46, 1289–1300. doi: 10.1017/S0033291715002834

Gurevitch, J., Koricheva, J., Nakagawa, S., and Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature* 555, 175–182. doi: 10.1038/nature25753

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *J. Economic Behav. Organiz.* 3, 367–388. doi: 10.1016/0167-2681(82)90011-7

Karns, C. M., Moore, W. E. III., and Mayr, U. (2017). The cultivation of pure altruism via gratitude: a functional MRI study of change with gratitude practice. *Front. Hum. Neurosci.* 11:599. doi: 10.3389/fnhum.2017.00599

Kelly, C., Toro, R., Di Martino, A., Cox, C. L., Bellec, P., Castellanos, F. X., et al. (2012). A convergent functional architecture of the insula emerges across imaging modalities. *Neuroimage* 61, 1129–1142. doi: 10.1016/j.neuroimage.2012.03.021

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314, 829–832. doi: 10.1126/science.1129156

Krueger, F., Barbey, A. K., and Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends Cognitive Sci.* 13, 103–109. doi: 10.1016/j.tics.2008.12.005

Krueger, F., Grafman, J., and Mccabe, K. (2008). Neural correlates of economic game playing. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 363, 3859–3874. doi: 10.1098/rstb.2008.0165

Krueger, F., and Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends Neurosci.* 39, 499–501. doi: 10.1016/j.tins.2016.06.004

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

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

Mccabe, K. A., Rassenti, S. J., and Smith, V. L. (1998). Reciprocity, trust, and payoff privacy in extensive form bargaining. *Games Econ. Behav.* 24, 10–24. doi: 10.1006/game.1998.0638

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* 15, 483–506. doi: 10.1016/j.tics.2011.08.003

Moll, J., Krueger, F., Zahn, R., Pardini, M., De Oliveira-Souza, R., and Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15623–15628. doi: 10.1073/pnas.0604475103

Muller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frassle, S., et al. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *Neuroimage* 119, 252–261. doi: 10.1016/j.neuroimage.2015.06.036

Namkung, H., Kim, S. H., and Sawa, A. (2017). The insula: an underestimated brain area in clinical neuroscience, psychiatry, and neurology. *Trends Neurosci.* 40, 200–207. doi: 10.1016/j.tins.2017.02.002

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

Patil, I., Calo, M., Fornasier, F., Cushman, F., and Silani, G. (2017). The behavioral and neural basis of empathic blame. *Sci. Rep.* 7:5200. doi: 10.1038/s41598-017-05299-9

Peterson, M. (2015). *The Prisoner's Dilemma.* Cambridge: Cambridge University Press.

Rilling, J. K., Goldsmith, D. R., Glenn, A. L., Jairam, M. R., Elfenbein, H. A., Dagenais, J. E., et al. (2008). The neural correlates of the affective response to unreciprocated cooperation. *Neuropsychologia* 46, 1256–1266. doi: 10.1016/j.neuropsychologia.2007.11.033

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science* 300, 1755–1758. doi: 10.1126/science.1082976

Santos, F. P., Santos, F. C., and Pacheco, J. M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature* 555, 242–245. doi: 10.1038/nature25763

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* 27, 2349–2356. doi: 10.1523/JNEUROSCI.5587-06.2007

Sheffield, J. M., Repovs, G., Harms, M. P., Carter, C. S., Gold, J. M., Macdonald, A. W. III, Daniel Ragland, J., et al. (2015). Fronto-parietal and cingulo-opercular network integrity and cognition in health and schizophrenia. *Neuropsychologia* 73, 82–93. doi: 10.1016/j.neuropsychologia.2015.05.006

Sridharan, D., Levitin, D. J., and Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12569–12574. doi: 10.1073/pnas.0800005105

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* 16, 55–61. doi: 10.1038/nrn3857

Uddin, L. Q., Kinnison, J., Pessoa, L., and Anderson, M. L. (2014). Beyond the tripartite cognition-emotion-interoception model of the human insular cortex. *J. Cogn. Neurosci.* 26, 16–27. doi: 10.1162/jocn_a_00462

Van Den Bos, W., Van Dijk, E., Westenberg, M., Rombouts, S. A., and Crone, E. A. (2011). Changing brains, changing perspectives: the neurocognitive development of reciprocity. *Psychol. Sci.* 22, 60–70. doi: 10.1177/0956797610391102

Van 'T Wout, M., Kahn, R. S., Sanfey, A. G., and Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Exp. Brain Res.* 169, 564–568. doi: 10.1007/s00221-006-0346-5

Wager, T. D., and Barrett, L. F. (2017). From affect to control: functional specialization of the insula in motivation and regulation. doi: 10.1101/102368

Zhu, R., Feng, C., Zhang, S., Mai, X., and Liu, C. (2019). Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses. *Neuroimage* 186, 476–486. doi: 10.1016/j.neuroimage.2018.11.012

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

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 16 June 2020
doi: 10.3389/fnhum.2020.00225

Neural Signatures of Gender Differences in Interpersonal Trust

Yan Wu[1,2], Alisha S. M. Hall[3], Sebastian Siehl[4,5], Jordan Grafman[6] and Frank Krueger[7,8]*

[1] Department of Psychology, College of Education, Hangzhou Normal University, Hangzhou, China, [2] Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments, Hangzhou Normal University, Hangzhou, China, [3] Department of Psychology, University of Mannheim, Mannheim, Germany, [4] Department of Cognitive and Clinical Neuroscience, Medical Faculty Mannheim, Central Institute of Mental Health, Ruprecht-Karls-University Heidelberg, Mannheim, Germany, [5] Graduate School of Economic and Social Sciences, University of Mannheim, Mannheim, Germany, [6] Shirley Ryan AbilityLab, Northwestern University, Chicago, IL, United States, [7] School of Systems Biology, George Mason University, Fairfax, VA, United States, [8] Department of Psychology, George Mason University, Fairfax, VA, United States

Edited by:
Carmen Moret-Tatay,
Catholic University of Valencia
San Vicente Mártir, Spain

Reviewed by:
Christoph W. Korn,
University Medical Center
Hamburg-Eppendorf, Germany
Adrian Alacreu-Crespo,
University of Valencia, Spain
José Salvador Blasco Magraner,
University of Valencia, Spain

***Correspondence:**
Frank Krueger
fkrueger@gmu.edu

Specialty section:
This article was submitted to
Cognitive Neuroscience,
a section of the journal
Frontiers in Human Neuroscience

Received: *16 January 2020*
Accepted: *19 May 2020*
Published: *16 June 2020*

Citation:
Wu Y, Hall ASM, Siehl S,
Grafman J and Krueger F (2020)
Neural Signatures of Gender
Differences in Interpersonal Trust.
Front. Hum. Neurosci. 14:225.
doi: 10.3389/fnhum.2020.00225

Trust plays a critical role in nearly every aspect of social life. Parental investment theory and social role theory predict that women trust less than men due to a higher sensitivity to risk and betrayal, while men trust more than women to maximize resources and to signal their willingness to lose something. However, the underlying neuropsychological underpinnings for this gender difference are still obscure. In this study, we used functional magnetic resonance imaging (fMRI) to investigate the neural signatures of gender differences in trust by simultaneously scanning 11 male and 11 female same-gender, fixed dyads who played a multi-round binary trust game with varying levels of payoff (low/moderate/high) as an indicator of social risk. Our results showed that men trusted more than women and payoff level moderated the effect of gender on trust. While men trusted the same at all payoff levels, women trusted less with higher payoff levels. This pattern was supported by our neuroimaging finding: men showed a higher activation in the left inferior frontal gyrus (ventrolateral prefrontal cortex) and right precuneus than women, indicating that men exert more effort to inhibit the information of payoff levels and to use self-referencing to infer the strategies of partners with the goal of maximizing profit. Furthermore, men showed equivalent activation in the subgenual anterior cingulate cortex across payoff levels, whereas women showed a decreased activation with increasing payoff level – indicating decreased group bonding with higher risk in women. In conclusion, our results imply that women are more sensitive to social risk while trusting, which has implications for financial interactions, interpersonal relationships, and social involvement.

Keywords: trust game, gender, risk, parental investment theory, social role theory, subgenual anterior cingulate cortex, inferior frontal gyrus, precuneus

INTRODUCTION

Trust is integral to relationships, cooperative behavior (Poppo et al., 2018), and a functioning society (Tov and Diener, 2009) and its dysfunction is a component of many mental disorders (King-Casas et al., 2008; Kéri et al., 2009; Sripada et al., 2009; Bell et al., 2019). Interpersonal trust is defined as the psychological state of a person (i.e., trustor) comprising the intention to accept vulnerability based upon positive expectations of behavior of another person (i.e., trustee) (Rousseau et al., 1998). The complexity and importance of interpersonal trust have motivated researchers to generate a wealth of experimental data over the years using standardized paradigms such as the trust game (Berg et al., 1995; Camerer, 2011).

In the standard version of the trust game, two players engage in a sequential, one-shot economic exchange. Player 1 (i.e., trustor) is given an endowment in monetary units (MU), of which the trustor can choose to send any amount (i.e., trust) or none (i.e., non-trust) to an anonymous player 2 (i.e., trustee). The MU sent by the trustor is then multiplied (e.g., doubled or tripled) by the experimenter. The trustee can choose to send back any amount of the received money (i.e., reciprocate) to the trustor or keep everything (i.e., betrayal). Although the subgame perfect Nash equilibrium for a rational player 1 – expecting that player 2 will not return any money – is to *not trust*, studies have shown that people typically send 50% of their initial endowment as player 1 and return about 37% of the money received as player 2 in one-shot trust game interactions (Johnson and Mislin, 2011).

A modification of the trust game – the experimental paradigm used in this study – is the binary multi-round trust game, where players alternate between the role of trustor and trustee with the same partner over multiple rounds and simply decide to *trust* or *not trust* and *reciprocate* or *betray* their partners' trust. The binary multi-round trust game has a higher ecological validity compared to the standard trust game because trust relationships in real life are reciprocal and go through phases of trust building and maintenance (and/or trust violation and recovery).

An important interpersonal factor in trust relationships is gender, which describes the norms, roles, and identity of women and men (and other culture-specific genders). Gender norms and roles, such as women being communal and caring and men being agentic and independent, are internalized through socialization, also known as the social role theory (Eagly and Wood, 2012). These internalized norms are known to guide and influence all types of behavior on an unconscious as well as conscious level, such as when individuals are anxious about conforming to negative stereotypes about their gender (also known as "stereotype threat"). For example, one study found that stereotype threat led to increased loss- and risk aversion behavior in women but not men when faced with a financial decision (Carr and Steele, 2010). Furthermore, when a socialized individual is met with a situation in which she/he must choose whether to make her/himself vulnerable to another person or not – especially when little else is known about the other person – gender is a readily available piece of information that can be used to predict how the other person will behave. For example, we have strong beliefs about each gender's trustworthiness (Slonim and Guillen, 2010; Zhao and Zhang, 2016), and the perceived trustworthiness of another person is known to be positively related to our decisions to trust (van 't Wout and Sanfey, 2008).

A recent meta-analysis of the one-shot trust game – encompassing 77 behavioral studies, 174 effect sizes, and 17,082 participants from 23 countries – found a robust effect (g = 0.22) of gender on trust, revealing that men send more money as player 1 than women (Van den Akker et al., 2018). In contrast, no overall effect of gender on trustworthiness was found: women and men were found to return about the same proportion of money received as player 2. One theory that has been proposed to explain the role that gender plays specifically in interpersonal trust is the parental investment theory (Trivers, 1972), which stems from Darwin's sexual selection theory (Darwin, 1871). The parental investment theory posits that (based on biological differences in the "cost" of producing and nurturing offspring between females and males) females evolved through natural selection to avoid physical and social risks to ensure their reproductive potential, while males evolved to take risks to signal their health, status, and resources to potential mates (Trivers, 1972). Males' reliance on using their resources to gain reproductive success, and females' reliance on holding resources to invest in parenting, leads to the evolution of risk-seeking behavior in males and risk-aversive behavior in females. These adaptations may have been selected specifically for their implications on decision-making under uncertainty (Dreber and Hoffman, 2010). When it is considered that the gender roles and norms for women and men developed from the biological roles of females and males, the evolutionary and sociocultural theories combined explain that women are less trusting because they have more to lose from social interactions and need to be more sensitive to treachery.

In contrast, men are more trusting because it signals that they can afford to lose something and provides an opportunity to gain resources and become a more attractive mate. In line with this, men have been found to take more risks (Fischer and Hills, 2012; Apicella et al., 2017). Furthermore, studies investigating specifically the propensity to take risks in different social settings find that this type of risk-taking is associated with trust behavior in the one-shot trust game (Karlan, 2005; Ben-Ner and Halldorsson, 2010; Lönnqvist et al., 2015). Although evidence about gender differences in trust behavior exists, its underlying neuropsychological mechanisms are still obscure. Previous neuroimaging studies regarding gender differences in trust found stronger activation in the temporal-parietal junction in males, whereas stronger activation in the caudate in females (Lemmers-Jansen et al., 2017, 2019). However, these studies only investigated trust gender differences playing against a cooperative and an unfair partner and did not vary social risks.

While trust (and the trust game) originate in economics research, trust has increasingly become a topic of research in psychology and, more recently, social neuroscience. The findings from these different scientific fields, with their diverse methods and perspectives, are complementary. A neuropsychoeconomic model of trust was recently proposed that aims to integrate findings from the fields of behavioral economics, psychology, and social neuroscience (Krueger and Meyer-Lindenberg, 2019). According to this model, trust arises through the interplay of trust components (i.e., treachery, reward, uncertainty, strategy, and trustworthiness) – linked to psychological systems (i.e., motivation, affect, and cognition) – that engage key brain regions anchored in domain-general large-scale brain networks. The anticipation of reward (motivational system, reward network including striatum, and ventromedial prefrontal cortex) contrasted with the risk of treachery (affective system, salience network including ACC, and anterior insula) creates uncertainty, which is associated with vulnerability of trusting another person. To remove uncertainty, trustors can adopt a context-based strategy (cognitive system, central-executive network including lateral prefrontal cortex, and posterior parietal cortex) to recap personal benefits (i.e., economic rationality) and/or evaluate the relationship-based trustworthiness (social cognition, default-mode network including medial prefrontal cortex, and posterior cingulate cortex) to contribute to the relationship's success (i.e., social rationality).

Interpersonal trust may evolve through repeated interactions from *calculus-based trust* (performing rational calculations of the costs and benefits of trust decisions driven mainly by the salience network), through *knowledge-based trust* (using knowledge about their context and the context of their partners to predict trustees' behavior and advance their trust relationships, driven mainly by the cognitive control network to adopt a strategy or default-mode network to evaluate trustworthiness), to *identification-based trust* (developing a rewarding identification with trustees, driven mainly by reward processing/pair bonding network) (Krueger and Meyer-Lindenberg, 2019).

The goal of this study was to test the predictions of the parental investment theory – namely, that men trust more to maximize resources whereas women trust less due to higher sensitivity to betrayal – in an experimental setting. To achieve this goal, we re-analyzed our previously published data (Krueger et al., 2007). Unlike previous studies, we measured trust behavior with a binary multi-round trust game closely imitating trust relationships in real life where dyads of participants switched between the role of trustor and trustee after each round. Further, we let dyads play rounds of the trust game at different payoff levels, with increasing MU's representing a higher social risk of betrayal not only in terms of material loss, but also with regard to the building and maintenance of the trust relationship over the course of the experiment. Our previously published study investigated the neural correlates of trust in partnership-building and maintenance stages in non-defector and defector groups, evidencing that conditional trust selectively activated the ventral tegmental area (reward system), whereas unconditional trust selectively activated the septal area (social attachment system) (Krueger et al., 2007). The present study focused on the gender differences of trust.

At the behavioral level, we hypothesized that dyads of both genders quickly develop a trust relationship from calculus-based over knowledge-based to identification-based trust due to the multi-round role-switching trust game format. However, we predicted that men trust more independently of payoff levels to maximize resources, whereas women trust less and adjust their trust based on payoff level to minimize their social risk. At the neural level, we predicted that to maximize resources, men utilize more brain regions associated with knowledge-based trust, implementing a context-based strategy to reap personal benefits (e.g., lateral prefrontal cortex) and evaluate the relationship-based trustworthiness to contribute to the relationship's success (e.g., precuneus and temporal-parietal junction). Finally, we hypothesized that men develop identification-based trust, activating brain regions associated with reward/bonding processes independently of payoff levels, whereas women decrease their activation in those brain regions with higher payoff levels.

MATERIALS AND METHODS

Participants

Forty-four (22 women and 22 men) healthy individuals were recruited via the subjects database of the National Institutes of Neurological Disorders and Stroke (NINDS) in Bethesda, Maryland, United States participated as same-gender dyads in an fMRI hyperscanning experiment (Montague et al., 2002) for financial compensation. We used same-gender dyads instead of opposite gender dyads to establish a valid start to test the interaction effect between social risk and gender. Dyads were matched by age ($M \pm SD$ = 28.4 ± 7.2 years, range = 21–51), education (17.3 ± 2.2 years, range = 12–23), and handedness (95.3 ± 8.7, range of 65–100, all right-handed) as assessed with the Edinburgh Handedness Inventory (Oldfield, 1971; **Supplementary Table S1**). The participants were native English speakers that had normal or corrected vision and were not taking any medication. Exclusion criteria were a history of medical, psychiatric, or neurological diagnoses and left-handedness. Prior to participating in the experiment, participants underwent a neurological examination as part of the screening procedure and provided informed consent in compliance with the standards of the NINDS Institutional Review Board. Note that the collected data from this study was used in a previous publication (Krueger et al., 2007).

Procedure

The experiment consisted of three phases. During the *pre-scanning phase*, two strangers of the same gender – playing the trust game as a pair – were first instructed separately in different rooms. The participants were briefly allowed to see each other (via webcam) and then asked to rate the perceived closeness, partnership, and leadership of the other person based on their first impression. Participants were asked to rate the perceived closeness to their partner (not at all vs. very close; 0–10) and rank themselves in comparison to the other player by leadership (leader vs. follower; 0–10), and partnership (competitor vs. partner; 0–10). Digital pictures were taken of the participants' neutral facial expression to be displayed to their partner on the screen during the trust game. Participants completed a practice run after being instructed about the experimental paradigm and private payment procedure at the end of the experiment.

During the *scanning phase*, dyads completed the fMRI experiment that consisted of 36 sequential trust games (i.e., experimental condition) and 16 control games (i.e., control condition). They played six blocks (i.e., six games per block) of the trust game and four blocks (i.e., four games per block) of the control game (**Supplementary Figure S1**). Six different payoff levels (p1–p6) were used during the experiment and each level was only used once per trust game block (**Figure 1A**). Payoff levels were categorized into three types of *social risk* for analysis: low (p1–p2), medium (p3–p4), and high (p5–p6). For p2, p4, and p6, the initial endowment was tripled (but for p1, p3, and p5, to the tripled amount 5 MUs were added) if the trustor chose to trust, so it could be evenly split if the trustee chose to reciprocate (i.e., to return 50% of the received amount). The order of blocks, payoff levels, and display layouts were counterbalanced across games, dyads, and participants.

Participants alternated between playing as player 1 (P1) and player 2 (P2) every round: the role assignment appeared on the screen at the beginning of each round along with the decision tree displaying the specific payoff outcomes of the game round (**Figure 1B**). For each trial of the binary trust game, P1 was given the choice to *trust* or *not trust* the other player, risking all or nothing of her/his initial endowment. If P1 chose to not trust, then the decision was shown to P2, and the round ended with both players keeping their initial endowment (both players received the same small amount of MU at the beginning of each round). If P1 chose to trust, then the amount was sent over, and P2 was given the choice to *reciprocate* or *betray* (keep everything). The game then ended after both players were shown the outcome of P2's decision. Participants were instructed to decide as quickly as possible: 100 MU were deducted from an individual player's cumulative total earnings every time they failed to decide within 6 s.

The control game (identical to the trust game) was used to control for the monetary, sensorimotor, and decision-making aspects of the task, but the participants played alone and were only tasked with choosing the action that would lead to the highest monetary payoff (**Supplementary Figure S2**). Cumulative totals for the MU earned in the trust game were not displayed during the experiment.

During the *post-scanning phase*, participants were once again asked to rate the closeness, partnership, and leadership felt toward their partner. Further, participants rated certain aspects of their own gameplay in the fMRI experiment with one-item measures: cooperation (competitive vs. cooperative; 0–10), trustfulness (suspicious vs. trusting; 0–10), hemisphere ("left-brained"/intuitive vs. "right-brained"/analytic; 0–10), and game strategy (always used the same strategy vs. used many strategies; 0–10). Moreover, empathy was assessed with the *Z*-score of the Balanced Emotional Empathy Scale (BEES), a 30-item self-report of one's tendency to experience other's emotional experiences (Mehrabian and Epstein, 1972). Finally, participants received their accumulated earnings from the experiment (between 0 and 25 USD) in addition to a fixed compensation for participating in the fMRI experiment.

Behavioral Data Analysis

Statistical analyses for the behavioral data was carried out with SPSS (Version 22.0.0.0, IBM Corporation 2013) applying a significance level of $\alpha < 0.05$ (two-tailed). Standardized effect sizes were calculated, including Cohen's d for t-tests and planned contrasts and $\eta_p{}^2$ (partial eta squared) for factors in analyses of variance (ANOVAs). The strength of effect sizes calculated using Cohen's d was interpreted using the cutoffs > 0.20 for small, > 0.50 for medium, and > 0.80 for large, and for effect sizes calculated using $\eta_p{}^2 > 0.01$ for small, > 0.06 for medium, and > 0.14 for large (Cohen, 1988; Miles and Shevlin, 2001).

Decisions and reaction times for the trust and trustworthiness behavior were calculated for different levels of payoff and phases of the trust game. Mixed three-way 3 × 2 × 2 ANOVAs were conducted for those measures with Payoff (low, medium, and high) and Phase [building (run 1) vs. maintenance (run 2)] as within-subjects factors and Gender (men vs. women) as a between-subjects factor. The direction of significant effects involving Payoff were investigated *post hoc* with Bonferroni-corrected Helmert contrasts of low payoff vs. higher payoffs and moderate vs. high payoff.

Further, the effect of gender on perceived changes in the trust relationship over time was invested employing mixed two-way 2 × 2 ANOVAs for partner ratings (i.e., closeness, partnership, and leadership) with Time (pre- vs. post-game) as within-subjects factor and Gender (men vs. women) as between-subjects factor. Finally, the self-ratings of cooperation, trustfulness, hemispheric use, and variability in game strategy as well as the assessed personality characteristics of empathy were compared between men and women with independent samples *t*-tests to identify potential control variables for behavior in the trust game.

Functional Image Acquisition and Preprocessing

Brain images were acquired on two 3-Tesla MRI whole-body scanners (General Electric) in the NMR Research Center at the National Institutes of Health. Head motion was restricted by using foam pads placed around the participant's head. Anatomical images were acquired using T1-weighted MP-RAGE sequence (TR = 9.7 ms, TE = 4.0 ms, flip angle = 12°, field of view = 240 mm, thickness = 1.2 mm, in-plane resolution = 0.8594*0.8594 mm^2), and T2*-weighted echo-planar images (EPI) optimized for BOLD contrast were collected (TR = 2 s, TE = 30 ms, flip angle = 90°, thickness = 6 mm, number of slices = 22, field of view = 240 mm, voxel dimensions = 3.75*3.75*6 mm). For each of the two functional runs, 291 volume images per run were taken parallel to AC–PC line. The first five volumes were discarded prior to analysis.

Image analyses were performed by using BrainVoyager QX (Version 20.6.2.3266 for Windows, Brain Innovation). Preprocessing steps included slice-time correction, linear trend removal, temporal high-pass filtering to remove low-frequency non-linear drifts of three or fewer cycles per time course, spatial smoothing (8 mm FWHM), and three-dimensional motion correction to detect and correct for small head movements by spatial alignment of all participants to the first volume by rigid body transformation. Estimated translation and rotation parameters were inspected and never exceeded 2 mm or 2°. To transform the functional data into Talairach space, the functional time series data of each subject were first co-registered with the subject's three-dimensional anatomical data set and resampled to 3 mm × 3 mm × 3 mm isotropic voxels, resulting in a normalized four-dimensional volume time course data.

Image Data Analysis

A general linear model (GLM) corrected for first-order serial correlation was applied (Friston et al., 1999). Random-effect

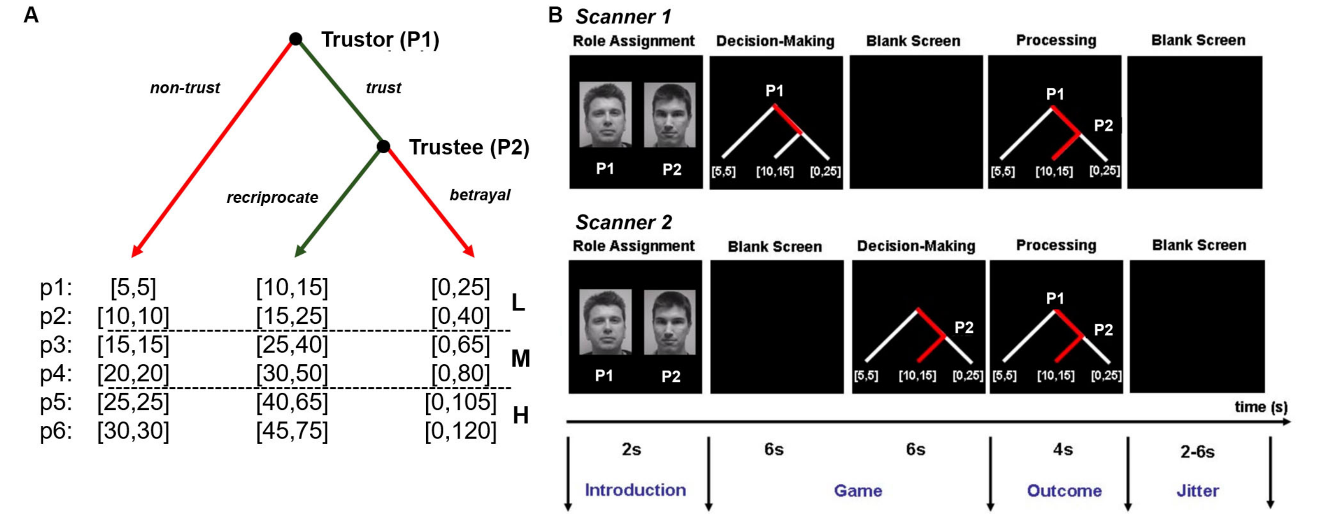


FIGURE 1 | Experimental design. **(A)** Binary trust game. Partners made sequential decisions as player 1 (P1) and player 2 (P2) for payoffs in cents presented in a binary decision tree [c: (cP1, cP2)]. P1 can either move left (i.e., non-trust) or quit the game with a small payoff for P1 and P2 [e.g., (5, 5)] or right (i.e., trust) to continue the game. Next, P2 can move left (i.e., reciprocate) giving both players a higher payoff [e.g., (10, 15)] or right (i.e., betrayal) resulting in an even larger payoff for P1 and a payoff of zero for P1 [e.g., (0, 25)]. Payoffs (p1–p6) were split into three level: L (low: p1–p2), M (medium: p3–p4, and H (high: p5–p6). Numbers inside brackets indicate the specific payoff outcomes of the game trial. **(B)** Game trial. An introduction screen (2 s) informed partners about the role (P1 or P2) that they were playing. P1 made a decision (i.e., non-trust or trust) (within 6 s) while seeing the game tree and waited afterward (6 s) for P2's decision while seeing a blank screen. After seeing a blank screen (6 s), P1 saw the game tree with P1's decision and made a decision (i.e., reciprocate or betrayal) (within 6 s). If P1 had chosen not to trust P2, the game was over, and P2 saw P1's decision (6 s). Partners saw the outcome of the game (4 s) followed by a blank screen with a jittered interstimulus interval (2–6 s). Adapted from Krueger et al. (2007), used with permission, Copyright (2007) National Academy of Sciences.

analyses were performed on the multisubject level (group data: n = 44) and group-level (subgroup data: n = 22 men; n = 22 women) to explore brain regions that are associated with decisions to trust. For each participant, regressors were created based on individual decisions as P1 and P2 in the trust games (TG) [P1: trust_low_payoff (p1–p2), trust_medium_payoff (p3–p4), trust_high_payoff (p5–p6), non-trust; blank_screen; P2: reciprocate, betrayal, blank_screen; P1 and P2: introduction, outcome_reciprocate, outcome_betrayal] and control games (CG) (P1: choice, blank_screen; P2: choice, blank_screen; P1 and P2: introduction, outcome_P1, outcome_P2) over both functional runs [i.e., building (run1) and maintenance (run2) phase of the trust relationship]. The regression model consisted of a set of 36 (2 × 18) regressors (11 TG and 7 CG regressors per phase). Regressor time courses were adjusted for the hemodynamic response delay by convolution with a double-gamma hemodynamic response function (Büchel et al., 1998). Multiple regression analyses were performed independently for the time course of each individual voxel.

A mixed three-way 3 × 2 × 2 ANOVA for parameter estimates of each voxel was conducted with Payoff (low, medium, and high) and Phase [building (run 1) vs. maintenance (run 2)] as within-subjects factors and Gender (women vs. men) as a between-subjects factor. Brain activations for the decision phase were reported after correcting for multiple comparisons using a cluster-level statistical threshold – employing the cluster-level statistical threshold estimator plugin in BrainVoyager QX. The thresholded map (p < 0.005) was used for a whole-brain correction criterion, which is based off an estimate of the map's spatial smoothness and on a Monte Carlo simulation (1,000 iterations). The minimum cluster size at a specified confidence level (α = 0.05) was then calculated (Forman et al., 1995; Goebel et al., 2006). The significant activation clusters were displayed in Talairach space on an average anatomical brain of all participants reversed left to right (i.e., radiological convention). Parameter estimates (mean weights) from activated brain regions were derived from the peak voxel activation and surrounding voxels encompassing 54 mm^3.

RESULTS

Behavioral Results

Participants trusted their partner 85% of the time (SD = 21.4, range = 22.2–100). The ANOVA for trust behavior showed a significant main effect of Payoff [$F_{(2, 84)}$ = 7.05, p = 0.001, η_p^2 = 0.14] (**Supplementary Table S2**), indicating that trust decreased with the level of payoff. Further, a significant main effect of Gender was found [$F_{(1, 42)}$ = 7.42, p = 0.009, η_p^2 = 0.15], showing that men trusted more (∼94%) than women (∼77%) (**Supplementary Table S11**). Finally, a significant Payoff × Gender interaction effect (but no Payoff × Gender × Phase interaction) was demonstrated [$F_{(2, 84)}$ = 5.17, p = 0.008, η_p^2 = 0.11], indicating that men showed the same levels of trust across all payoff levels, whereas women showed a decrease in trust with higher payoff levels independently of the phase for the trust relationship (**Supplementary Figure S3**). *Post hoc* tests revealed significant gender differences in trials with moderate

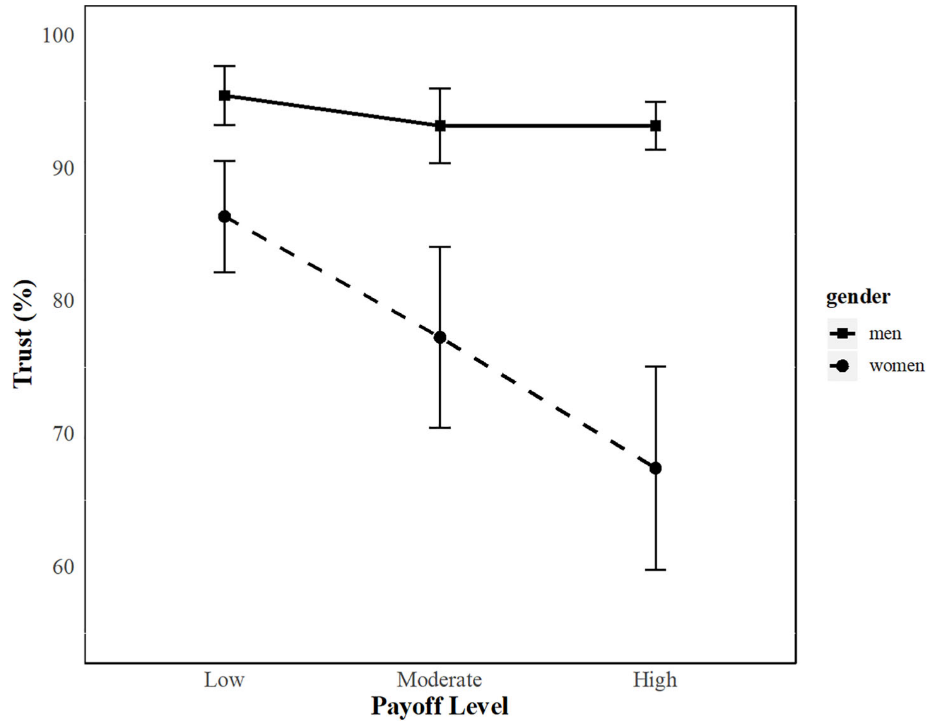


FIGURE 2 | Trust as a function of gender and payoff level. Trust (mean ± standard error) decreased for women but stayed the same for men across all payoff levels.

payoff (t_{42} = 2.16, p = 0.04, d = 0.65) as well as high payoff (t_{42} = 3.29, p = 0.003, d = 0.99), but marginal significance in trials with low payoff (t_{42} = 1.91, p = 0.065, d = 0.58) (**Figure 2**). Men and women chose to reciprocate their partner's trust about 85% of the time (*SD* = 25.8, range = 0–100). The ANOVA for trustworthiness behavior showed no significant main and interaction effects (**Supplementary Table S3**, Note that 6 women were subject to listwise exclusion because some women never trusted their partner for the high payoff level).

The ANOVA for response times showed only a significant Gender effect for trust [$F_{(1,\ 39)}$ = 4.62, p = 0.038, $\eta_p{}^2$ = 0.11] but not for trustworthiness (**Supplementary Tables S4, S5**). Women were on average 113 ms faster than the men trusting their partner (women: *M* = 1359 ms, *SD* = 587; men: *M* = 1472 ms, *SD* = 736). Further, significant main effects of Phase were observed, indicating that participants (independently of gender) became faster to trust [$F_{(1,\ 39)}$ = 7.23, p = 0.011, $\eta_p{}^2$ = 0.17] and reciprocate [$F_{(1,\ 36)}$ = 6.40, p = 0.017, $\eta_p{}^2$ = 0.17] their partners from the building (run 1) to the maintenance (run 2) phase of the trust relationship.

The ANOVAs for participants' changing beliefs about their partners over time showed no significant main effects of Gender on closeness [$F_{(1,\ 42)}$ = 2.15, p = 0.150, $\eta_p{}^2$ = 0.05], partnership [$F_{(1,\ 42)}$ = 0.87, p = 0.355, $\eta_p{}^2$ = 0.02], and leadership [$F_{(1,\ 42)}$ = 0.12, p = 0.737, $\eta_p{}^2$ = 0.01] (**Supplementary Tables S6–S8**). However, significant main effects of Time were observed for closeness [$F_{(1,\ 42)}$ = 14.5, p = 0.001, $\eta_p{}^2$ = 0.26] and partnership [$F_{(1,\ 42)}$ = 4.82, p = 0.034, $\eta_p{}^2$ = 0.10] but not for leadership [$F_{(1,\ 42)}$ = 2.90, p = 1.000, $\eta_p{}^2$ = 0.00]. Independently of gender, participants felt that they became closer to and developed a higher degree of partnership with their partner after playing the multi-round trust game.

Among the series of the assessed control and personality measures, women were using more strategies than men in the trust game ($t_{36\cdot 0}$ = 2.89, p = 0.006, d = 0.68) (**Supplementary Table S9**). Although women scored higher than men on empathy ($t_{31\cdot 0}$ = 2.86, p = 0.007, d = 0.65), the previously identified significant main effects of Payoff and Gender as well as the Payoff × Gender interaction effect remained significant after running a mixed three-way analysis of covariance (ANCOVA) for trust behavior with Payoff (low, medium, and high) and Phase (building vs. maintenance) as within-subjects factors, Gender (men vs. women) as a between-subjects factor, and Empathy as a covariate (**Supplementary Table S10**).

Neuroimaging Results

A mixed three-way 3 × 2 × 2 ANOVA was performed to identify brain activations during the decision phase of the trust game (corrected for multiple comparisons at the cluster level). A significant main effect of Gender was found, indicating a greater brain activation for men compared to women in the left inferior frontal gyrus (ventrolateral prefrontal cortex, VLPFC, BA 44; Tal −45, 5, 28) and right precuneus (PreC, BA 7; Tal 9, −63, 38) (α < 0.05, k = 21) (**Figures 3A,B**). Further, a significant interaction effect of Payoff × Gender (α < 0.05, k = 14) was found in which activation in the left subgenual anterior cingulate cortex (SgACC, BA 24; Tal −5, 26, 1) decreased with payoff levels in women compared to men (**Figure 3C**). Note that to avoid

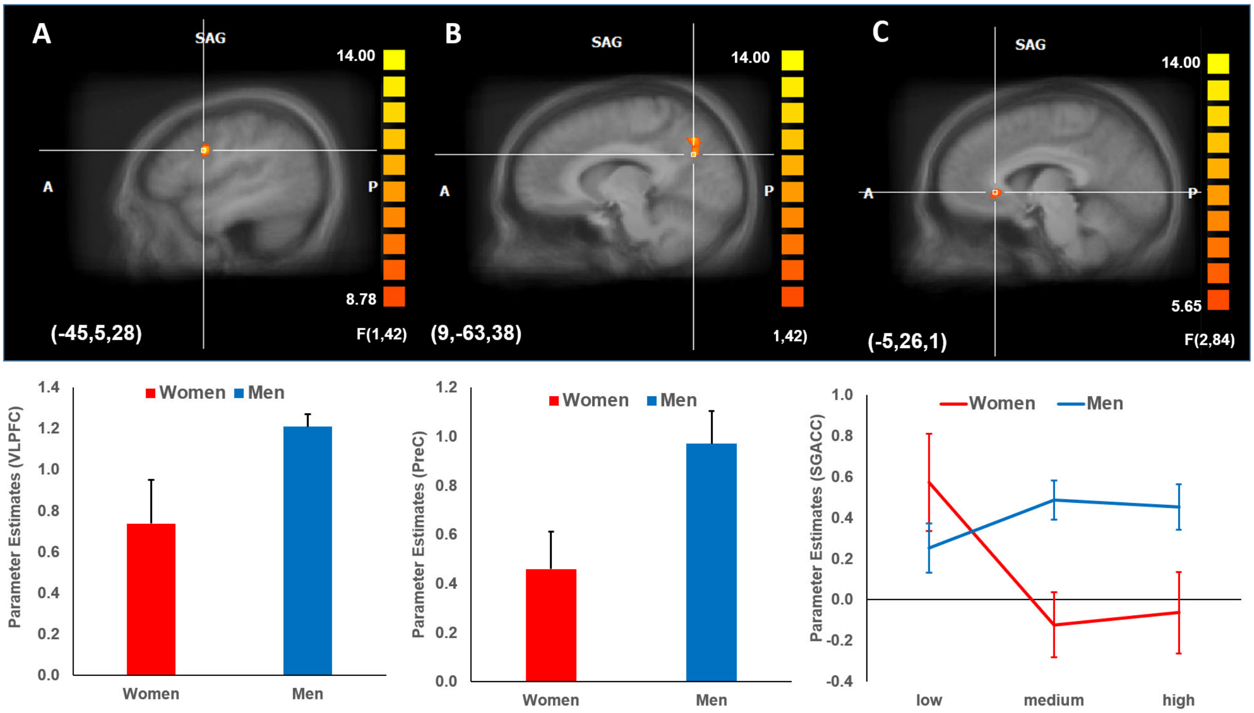


FIGURE 3 | Brain activations (mean parameter estimates ± standard error) during decision phase corrected for multiple comparisons at the cluster level. **(A,B)** Gender effect. Men showed higher activation in the left inferior frontal gyrus (VLPFC, BA 44) and right precuneus (PreC, BA 7) compared to women ($\alpha < 0.05$, $k = 21$). **(C)** Payoff x Gender interaction effect. Men activated consistently the left SgACC (BA 24) ($\alpha < 0.05$, $k = 14$), whereas activation decreased for women across payoff levels. Parameter estimates (mean weights ± standard error) from activated brain regions were derived from the peak voxel activation and surrounding voxels encompassing 54 mm^3. BA, Brodman area; VLPFC, ventrolateral prefrontal cortex; PreC, precuneus; SgACC, subgenual ACC; SAG, sagittal.

circularity (i.e., "double dipping"), no further statistical analyses were performed regarding this interaction effect (Kriegeskorte et al., 2009). No further brain activations for the remaining main and interactions effects were found.

DISCUSSION

The goal of this study was to investigate the underlying neuropsychological signatures of gender differences in trust by combining fMRI with a multi-round binary trust game that varied the social risk in the form of different payoff levels. On the behavioral level, we observed that women trusted less than men – decreasing their trust with the payoff level – but reciprocated the same. Further, we demonstrated on the neural level that men compared to women showed greater activation in the left VLPFC and right PreC, indicating a stronger recruitment of cognitive control and self-referencing strategies respectively in men for the goal of maximizing profit. Further, men exhibited similar activation across all payoff levels in the left SgACC – a region involved in group affiliation and reward processing – whereas women showed decreased activation with increasing payoff levels. Our results support the parental investment theory, which posits a gender-specific effect of social risk on trust behavior in women and higher trust among men to maximize resources.

In this study, both men and women trusted their partner about 85% of the time overall while playing a multi-round version of the binary trust game. Although similar in design to a previous study (Haselhuhn et al., 2015), our study was designed so that participants played repeatedly with the same partner while alternating roles as a trustor and trustee and, as such, were likely more motivated to develop a trust relationship. As a consequence, both men and women felt closer to their partners and reported a higher degree of partnership after completing the trust game. Further, they did not perceive a change in the degree of leadership shown by their partner, reflecting the reciprocal nature of trust relationships in which both partners are sometimes the trustor and sometimes the trustee. Moreover, to control for the possible confounding by discrimination against the opposite gender, participants were paired by same-gender, in which people may be involved in a parochial altruism situation (be altruistic to in-group members).

Despite the high amount of trust across genders, men trusted more than women overall independently of payoff levels, while women decreased their trust with increasing payoff levels as a measure of social risk. This remained the case after controlling for empathy, a personality trait on which women consistently score higher than men (Christov-Moore et al., 2014). Women were faster overall in deciding to trust and reported utilizing more strategies than men – pointing to adaptive behavior in the face of different levels of social risk.

Our results not only replicated the findings of a previous meta-analysis investigating gender differences in trust games (Van den Akker et al., 2018), but also provided empirical evidence for the parental investment theory. Women trusted less due to a higher sensitivity for social risk, supporting the assumption that gender norms for women and men are acquired through socialization "evolved" from a female biological imperative to avoid social risk and betrayal in relationships and the male motivation to maximize resources. These results are consistent with recent evidence showing that hormonal changes after competition predict sex-differentiated decision-making, i.e., women make more conservative decisions and men more riskier decisions after experiencing a competitive social context (Alacreu-Crespo et al., 2019).

Mirroring our findings on the behavioral level that men trust more than women, men showed greater activation in the left VLPFC and right PreC at the neural level.

On the one hand, the VLPFC has been linked to enhancing goal-directed behavior and improving long-term outcomes during trust decisions (Krueger and Meyer-Lindenberg, 2019). Although a broad array of functions are associated with the VLPFC – language processing (Wagner et al., 2014), mental imagery (Kleider-Offutt et al., 2019), planning (Fincham et al., 2002), selective bias of behaviorally relevant information (Blackwood et al., 2000), and selection among competing information to guide a response (Thompson-Schill et al., 1998) – former findings emphasize the role of the VLPFC in cognitive control and inhibition (Bereczkei et al., 2015). For example, a previous study revealed that control of distrust and individual differences in change of distrust are linked with left VLPFC activity, reflecting an increased engagement of cognitive control in individuals who tend to change their distrust evaluations (Filkowski et al., 2016).

In our study, increased activation of the left VLPFC was found in men compared to women, suggesting that men engaged more cognitive control while trusting for the purpose of maximizing the reward. The VLPFC may play a crucial role in the evaluation of signals associated with the others' social behavior and show elevated activity when the trustor faces a moderate or high risk of betrayal. Increased recruitment of the left VLPFC in men may indicate that they are more willing to pursue larger rewards by trust despite the increased social risk and uncertainty-related costs of doing so (as reflected by the insensitivity to payoff levels). To maximize their profit, men may selectively inhibit information about social risk and maintain a positively biased expectation of their partner's reciprocity. This interpretation is supported by the questionnaire data from the post-scanning phase of the experiment: men reported using fewer strategies in the trust game to fulfill their economic goals. In addition, men showed longer response times than women across payoff levels, indicating probably that additional time was spent on the inhibition process.

On the other hand, the right PreC (stronger engaged by men than women) is a key region of the default-mode network (Cavanna and Trimble, 2006), which has been implicated in social cognition processes such as mentalizing (Zaki and Ochsner, 2012; Suzuki et al., 2019), tracking social distance (Tavares et al., 2015), and processing social identity (Volz et al., 2009). Previous trust studies have identified the involvement of the PreC in trust decisions (Emonds et al., 2014), and its activation is positively associated with an individuals' level of trust as the result of perspective-taking (Prochazkova et al., 2018). More generally, the PreC may be involved in the evaluation of positive social interactions such as other's benevolence and trustworthiness (Fett et al., 2014; McAdams et al., 2015). The greater PreC response observed in men in our study suggest that the PreC may also play a role in strategic trust decisions. In a multi-round trust game with the same partner and role-switching, men may have been more motivated to always trust as trustors to signal their trustworthiness to maximize profits. Greater PreC responses are consistent with males' enhanced tendency to use self-referencing to infer the strategies of partners in the repeated trust game (Lambert et al., 2017). This confirms a prior study suggesting that PreC activations are linked to attempts to understand the responsiveness of others (Sakaiya et al., 2013).

Finally, the SgACC was differently activated for genders depending on the payoff level. A recent review on moral motivation examined the role of the SgACC in moral choices, with stronger activation associated with higher empathic concern, more donations, enhanced aversion to third-party harm, feelings of guilt, and group affiliation (Zahn et al., 2020). This review posited that the SgACC may represent attachment-related values of social outcomes. This suggestion is in line with the finding that the SgACC is activated both in response to personal and vicarious reward as identified by a meta-analysis (Morelli et al., 2015) as well as with the finding of another meta-analysis that the SgACC is preferentially recruited during altruistic giving (Cutler and Campbell-Meiklejohn, 2019).

The SgACC activation pattern in our study – decreased activation with increased social risk in women and no activation differences across all payoff levels in men – suggests that women adapt their strategy toward partners when the social risk increases. Both women and men could quickly form identification-based trust in the multi-round role-switching trust game and built up partnership in their pair. However, women were more sensitive to the social risk of betrayal. When the social risk became salient, women chose to detach themselves from the partnership and decrease their concern for their partner. In contrast, men were more focused on implementing a profit-maximizing strategy and gaining rewards, pursuing this target irrespective of payoff level. Overall, the described pattern was captured by SgACC activation, probably representing decreased pair bonding among women in moderate and high-risk contexts.

Unlike the behavioral results, we failed to find a main effect of payoff at the neural level. This may be due to a modest degree of repetition for each payoff level (12 trials for each level), which may have resulted in a lack of power to detect a subtle effect of risk on brain activity. Another possibility is that the social risk effect was masked by gender. As shown before, men were more focused on rewards irrespective of payoff level as evidenced by both behavioral and neural results. Only women distinguished between low and moderate/high risk levels; therefore, the risk effect might have been mitigated by this diversity.

Some limitations of our study must be considered. First, participants only played the game with a partner of the same gender. Whether playing with the opposite gender would show the same gender effect needs further investigation. As our study was the first to test the interaction effect between social risk and gender, it was more appropriate to use same-gender and same-race dyads to establish a valid start for this line of research. Future studies should investigate whether the observed interaction between social risk and gender can also be observed in mixed-gender dyads as well as in multiple, one-shot trust games with different partners (e.g., in-group vs. out-group members, same vs. different sexual orientation).

Second, our current neural findings should be considered with caution due to the lowered significance threshold for the whole-brain analysis (uncorrected $p < 0.005$ before correcting for multiple comparisons at the cluster level). However, the activation pattern replicated previous trust studies, which provides evidence the reliability of the current results (Emonds et al., 2014; Prochazkova et al., 2018). Future studies may shed further light on the neural mechanisms underlying gender differences by including large samples to effectively detect the subtle gender differences in associated brain regions.

CONCLUSION

Despite those limitations, our study is the first to manipulate social risk in the trust game to investigate its moderating role in the observed gender differences in trust behavior using functional neuroimaging and show neural-psychological evidence that there is a tradeoff between social risk and partnership in women. Women trust less than men, and this effect is stronger when the social risk increased for the trust decision. We also demonstrated that men are more determined to adopt a constant strategy to maximize resources and to stick to this goal, as reflected by the greater activation in the VLPFC and PreC. In contrast, women were more sensitive to the social risk of betrayal, as indicated by the decreased SgACC activation when social risk increased. These results provide support for the prediction made by the parental investment theory. Our findings that women are more sensitive to social risk while trusting have implications for various aspect of social life such as financial interactions, interpersonal relationships, and social structures.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the NINDS Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JG and FK designed the study. FK collected the data. AH, SS, and FK analyzed the data. YW, AH, and FK prepared the first draft of the manuscript. All authors contributed to the final version.

FUNDING

This work was supported in part by the National Institute of Neurological Disorders and Stroke/National Institutes of Health (NINDS/NIH) postdoctoral competitive fellowship award (to FK) and the Intramural Research Program of the Cognitive Neuroscience Section, NINDS/NIH (to JG).

ACKNOWLEDGMENTS

We thank E. Wassermann for performing the neurological exams, N. Armstrong, and M. Ulrich for help in various stages of this project, and A. Butz for help in programming the fMRI experiment. We also thank J. Bodurka and S. Marrett for technical assistance for the hyper-fMRI setup and E. Condon, J. Moll, R. Zahn, and M. Strenziok for scanning assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2020.00225/full#supplementary-material

REFERENCES

Alacreu-Crespo, A., Costa, R., Abad-Tortosa, D., Hidalgo, V., Salvador, A., and Serrano, M. Á (2019). Hormonal changes after competition predict sex-differentiated decision−making. *J. Behav. Decis. Making* 32, 550–563. doi: 10.1002/bdm.2128

Apicella, C. L., Crittenden, A. N., and Tobolsky, V. A. (2017). Hunter-gatherer males are more risk-seeking than females, even in late childhood. *Evol. Human Behav.* 38, 592–603. doi: 10.1016/j.evolhumbehav.2017.01.003

Bell, V., Robinson, B., Katona, C., Fett, A.-K., and Shergill, S. (2019). When trust is lost: the impact of interpersonal trauma on social interactions. *Psychol. Med.* 49, 1041–1046. doi: 10.1017/S0033291718001800

Ben-Ner, A., and Halldorsson, F. (2010). Trusting and trustworthiness: what are they, how to measure them, and what affects them. *J. Econ. Psychol.* 31, 64–79. doi: 10.1016/j.joep.2009.10.001

Bereczkei, T., Papp, P., Kincses, P., Bodrogi, B., Perlaki, G., Orsi, G., et al. (2015). The neural basis of the Machiavellians' decision making in fair and unfair situations. *Brain Cogn.* 98, 53–64. doi: 10.1016/j.bandc.2015.05.006

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

Blackwood, N. J., Howard, R. J., ffytche, D. H., Simmons, A., Bentall, R. P., and Murray, R. M. (2000). Imaging attentional and attributional bias: an fMRI approach to the paranoid delusion. *Psychol. Med.* 30, 873–883. doi: 10.1017/s0033291799002421

Büchel, C., Holmes, A. P., Rees, G., and Friston, K. J. (1998). Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. *Neuroimage* 8, 140–148. doi: 10.1006/nimg.1998.0351

Camerer, C. F. (2011). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press. doi: 10.1007/s11615-004-0067-y

Carr, P. B., and Steele, C. M. (2010). Stereotype threat affects financial decision making. *Psychol. Sci.* 21, 1411–1416. doi: 10.1177/0956797610384146

Cavanna, A., and Trimble, M. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain J. Neurol.* 129, 564–583. doi: 10.1093/brain/awl004

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., and Ferrari, P. F. (2014). Empathy: gender effects in brain and behavior. *Neurosci. Biobehav. Rev.* 46, 604–627. doi: 10.1016/j.neubiorev.2014.09.001

Cohen, J. (1988). *The Effect Size Index: d. Statistical Power Analysis for the Behavioral Sciences. The Analysis of Variance. Statistical Power Analysis for the Behavioral Sciences. 20-26*. Lawrence: Erlbaum Associates, Publishers. doi: 10.4324/9780203771587

Cutler, J., and Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *Neuroimage* 184, 227–241. doi: 10.1016/j.neuroimage.2018.09.009

Darwin, C. (1871). *The Descent of Man and Selection In Relation To Sex. Facsimile of the First Edition. 2 vols*. London: Murray. doi: 10.1037/12294-002

Dreber, A., and Hoffman, M. (2010). *Biological Basis of Sex Differences in Risk Aversion and Competitiveness*. Stockholm: Stockhom School of Economics.

Eagly, A. H., and Wood, W. (2012). "Social role theory," in *Handbook of Theories of Social Psychology*, Vol. 2, eds P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins (Thousand Oaks, CA: Sage Publications Ltd.), 458–476. doi: 10.4135/9781446249222.n49

Emonds, G., Declerck, C. H., Boone, C., Seurinck, R., and Achten, R. (2014). Establishing cooperation in a mixed-motive social dilemma. An fMRI study investigating the role of social value orientation and dispositional trust. *Soc. Neurosci.* 9, 10–22. doi: 10.1080/17470919.2013.858080

Fett, A.-K. J., Gromann, P. M., Giampietro, V., Shergill, S. S., and Krabbendam, L. (2014). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Soc. Cogn. Affect. Neurosci.* 9, 395–402. doi: 10.1093/scan/nss144

Filkowski, M. M., Anderson, I. W., and Haas, B. W. (2016). Trying to trust: brain activity during interpersonal social attitude change. *Cogn. Affect. Behav. Neurosci.* 16, 325–338. doi: 10.3758/s13415-015-0393-0

Fincham, J. M., Carter, C. S., van Veen, V., Stenger, V. A., and Anderson, J. R. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3346–3351. doi: 10.1073/pnas.052703399

Fischer, D., and Hills, T. T. (2012). The baby effect and young male syndrome: social influences on cooperative risk-taking in women and men. *Evol. Human Behav.* 33, 530–536. doi: 10.1016/j.evolhumbehav.2012.01.006

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster−size threshold. *Magn. Reson. Med.* 33, 636–647. doi: 10.1002/mrm.1910330508

Friston, K. J., Holmes, A. P., and Worsley, K. J. (1999). How many subjects constitute a Study? *Neuroimage* 10, 1–5. doi: 10.1006/nimg.1999.0439

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single−subject to cortically aligned group general linear model analysis and self−organizing group independent component analysis. *Human Brain Mapp.* 27, 392–401. doi: 10.1002/hbm.20249

Haselhuhn, M. P., Kennedy, J. A., Kray, L. J., Van Zant, A. B., and Schweitzer, M. E. (2015). Gender differences in trust dynamics: women trust more than men following a trust violation. *J. Exp. Soc. Psychol.* 56, 104–109. doi: 10.1016/j.jesp.2014.09.007

Johnson, N. D., and Mislin, A. A. (2011). Trust games: a meta-analysis. *J. Econ. Psychol.* 32, 865–889. doi: 10.1016/j.joep.2011.05.007

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *Am. Econ. Rev.* 95, 1688–1699. doi: 10.1257/000282805775014407

Kéri, S., Kiss, I., and Kelemen, O. (2009). Sharing secrets: oxytocin and trust in schizophrenia. *Soc. Neurosci.* 4, 287–293. doi: 10.1080/17470910802319710

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

Kleider-Offutt, H. M., Grant, A., and Turner, J. A. (2019). Common cortical areas involved in both auditory and visual imageries for novel stimuli. *Exp. Brain Res.* 237, 1279–1287. doi: 10.1007/s00221-019-05492-4

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

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

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

Lambert, B., Declerck, C. H., Emonds, G., and Boone, C. (2017). Trust as commodity: social value orientation affects the neural substrates of learning to cooperate. *Soc. Cogn. Affect. Neurosci.* 12, 609–617. doi: 10.1093/scan/nsw170

Lemmers-Jansen, I. L., Fett, A. K. J., Shergill, S. S., Van Kesteren, M. T., and Krabbendam, L. (2019). Girls-boys an investigation of gender differences in the behavioral and neural mechanisms of trust and reciprocity in adolescence. *Front. Human Neurosci.* 13:257. doi: 10.3389/fnhum.2019.00257

Lemmers-Jansen, I. L., Krabbendam, L., Veltman, D. J., and Fett, A. K. J. (2017). Boys vs. girls: gender differences in the neural development of trust and reciprocity depend on social context. *Dev. Cogn. Neurosci.* 25, 235–245. doi: 10.1016/j.dcn.2017.02.001

Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., and Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: task or ask? An empirical comparison. *J. Econ. Behav. Organ.* 119, 254–266. doi: 10.1016/j.jebo.2015.08.003

McAdams, C. J., Lohrenz, T., and Montague, P. R. (2015). Neural responses to kindness and malevolence differ in illness and recovery in women with anorexia nervosa. *Human Brain Mapp.* 36, 5207–5219. doi: 10.1002/hbm.23005

Mehrabian, A., and Epstein, N. (1972). A measure of emotional empathy. *J. Pers.* 40, 525–543. doi: 10.1111/j.1467-6494.1972.tb00078.x

Miles, J., and Shevlin, M. (2001). *Applying Regression and Correlation: A Guide for Students and Researchers*. Thousand Oaks, CA: Sage.

Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., et al. (2002). Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* 16, 1159–1164. doi: 10.1006/nimg.2002.1150

Morelli, S. A., Sacchet, M. D., and Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: a quantitative meta-analysis. *Neuroimage* 112, 244–253. doi: 10.1016/j.neuroimage.2014.12.056

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Poppo, L., Cheng, Z., and Cheng, Z. (2018). *Trust and Contracts: Complements Versus Substitutes in Business-to-business Exchanges*. Abingdon: The Routledge Companion to Trust. doi: 10.4324/9781315745572-16

Prochazkova, E., Prochazkova, L., Giffin, M. R., Scholte, H. S., De Dreu, C. K. W., and Kret, M. E. (2018). Pupil mimicry promotes trust through the theory-of-mind network. *Proc. Natl. Acad. Sci. U.S.A.* 115, E7265–E7274. doi: 10.1073/pnas.1803916115

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manag. Rev.* 23, 393–404. doi: 10.5465/amr.1998.926617

Sakaiya, S., Shiraito, Y., Kato, J., Ide, H., Okada, K., Takano, K., et al. (2013). Neural correlate of human reciprocity in social interactions. *Front. Neurosci.* 7:239. doi: 10.3389/fnins.2013.00239

Slonim, R., and Guillen, P. (2010). Gender selection discrimination: evidence from a trust game. *J. Econ. Behav. Organ.* 76, 385–405. doi: 10.1016/j.jebo.2010.06.016

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

Suzuki, A., Ueno, M., Ishikawa, K., Kobayashi, A., Okubo, M., and Nakai, T. (2019). Age-related differences in the activation of the mentalizing- and reward-related brain regions during the learning of others' true trustworthiness. *Neurobiol. Aging* 73, 1–8. doi: 10.1016/j.neurobiolaging.2018.09.002

Tavares, R., Mendelso, A., Grossman, Y., Williams, C., Shapiro, M., Trope, Y., et al. (2015). A map for social navigation in the human brain. *Neuron* 87, 231–243. doi: 10.1016/j.neuron.2015.06.011

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., and Knight, R. T. (1998). Verb generation in patients with focal frontal lesions: a neuropsychological test of neuroimaging findings. *Proc. Natl. Acad. Sci. U.S.A.* 95, 15855–15860. doi: 10.1073/pnas.95.26.15855

Tov, W., and Diener, E. (2009). "The well-being of nations: linking together trust, cooperation, and democracy," in *The Science of Well-Being: The Collected Works of Ed Diener*, ed. E. Diener (Dordrecht: Springer Netherlands), 155–173. doi: 10.1007/978-90-481-2350-6_7

Trivers, R. L. (1972). "Parental investment and sexual selection," in *Sexual Selection and the Descent of Man 1871 -1971*, ed. B. Campbell (Chicago, IL: Aldine), 136–179. doi: 10.4324/9781315129266-7

Van den Akker, O., van Vugt, M., van Assen, M. A. L. M., and Wicherts, J. M. (2018). Sex differences in trust and trustworthiness—a meta-analysis of the trust game and the gift-exchange game. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/5zbja

van 't Wout, M., and Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition* 108, 796–803. doi: 10.1016/j.cognition.2008.07.002

Volz, K. G., Kessler, T., and von Cramon, D. Y. (2009). In-group as part of the self: in-group favoritism is mediated by medial prefrontal cortex activation. *Soc. Neurosci.* 4, 244–260. doi: 10.1080/17470910802553565

Wagner, S., Sebastian, A., Lieb, K., Tuscher, O., and Tadic, A. (2014). A coordinate-based ALE functional MRI meta-analysis of brain activation during verbal fluency tasks in healthy control subjects. *BMC Neurosci.* 15:19. doi: 10.1186/1471-2202-15-19

Zahn, R., de Oliveira-Souza, R., and Moll, J. (2020). Moral motivation and the basal forebrain. *Neurosci. Biobehav. Rev.* 108, 207–217. doi: 10.1016/j.neubiorev.2019.10.022

Zaki, J., and Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nat. Neurosci.* 15, 675–680. doi: 10.1038/nn.3085

Zhao, N., and Zhang, J. (2016). Gender differences in trusting strangers: role of the target's gender. *PsyCh J.* 5, 83–91. doi: 10.1002/pchj.120

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership