Psychological and Neural Dynamics of Trust

Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaft (Dr. rer. nat.)

am Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

vorgelegt von

Gabriele Bellucci, M.A., M.Sc.

Berlin 2019

Erstgutachter: Prof. Dr. Hauke Heekeren

Zweitgutachter: Prof. Dr. Soyoung Park

Tag der Disputation: 14. November 2019

List of Abbreviations

TG	Trust Game
TAG	Take Advice Game
PE	Prediction error
fMRI	functional magnetic resonance imaging
PFC	prefrontal cortex
VMPFC	ventromedial prefrontal cortex
DLPFC	dorsolateral prefrontal cortex
pTPJ	posterior temporoparietal junction
RSFC	resting-state functional connectivity
MU	monetary units
ANOVA	analysis of variance
AIC	Akaike information criterion
EPI	echo-planar imaging
MP-RAGE	magnetization-prepared rapid acquisition with gradient-echo
SPM	statistical parametric mapping
MNI	Montreal Neurological Institute
FWHM	full width at half maximum
GLM	general linear model
FWE	family-wise error
FDR	false discovery rate
PPI	psychophysiological interaction
BOLD	blood-oxygen-level-dependent signal
MVPA	multivariate voxel pattern analysis
SVM	support vector machine
LOROCV	leave one run out cross validation

LOSOCV	leave one subject out cross validation
SMSE	standardized mean squared error
PCC	posterior cingulate cortex
IPS	intraparietal sulcus
ACC	anterior cingulate cortex
OFC	orbitofrontal cortex
CON	cinguloopercular network
SMN	somatosensory network

Table of contents

Summary

Trust is a key feature of social interactions and central to interpersonal cooperation. Acts of trust are not only pivotal aspects of interpersonal cooperation and group cohesion, they also have important consequences for individual health and life expectancy. However, which social qualities of others foster trust, how individuals learn whom to trust, and how the brain integrates this information for optimal behavioral updating is yet unexplored. Here, I will outline two lines of research. On one hand, I will show the psychological and neural predictors of trust in different social contexts. On the other, pharmacological modulations of the neural brain structures involved in trust will be presented. In the first two behavioral experiments, I show that honesty functions as an antecedent of trustworthiness impressions and that an honest reputation is associated with higher trust during a future social interaction. Next, I delineate the neural signatures of these honesty-based trustworthiness impressions. Notably, similar to the behavioral effects of honesty on future trust decisions, I found that honesty-encoding brain regions predicted those future trust decisions, providing evidence of honesty-related brain regions that entail neural signal predictive of trusting behavior. Furthermore, an honest reputation also modulated neural responses to feedback information. Such neural modulation likely biases information integration during social learning. Consequently, I show in a further behavioral study that an honest reputation seems to indeed impair learning due to an honesty-dependent asymmetry in information weighting. Finally, I demonstrate how the pharmacological modulation of brain dynamics impacts trusting behaviors leaving trustworthiness impressions unchanged. On the one hand, these findings shed light on how honesty not only increases trust in others but also hampers learning processes for optimal behavioral adaptation. On the other, they provide the first pharmacological evidence of how impression-based trust can be changed without impacting those very first trustworthiness

impressions. I finally propose accounts that might explain the observed behavioral and neural patterns and outline potential directions for new studies.

Zusammenfassung

Vertrauen ist ein Hauptmerkmal sozialen Austausches und von wesentlicher Bedeutung für zwischenmenschliche Zusammenarbeit. Vertrauensakte sind aber nicht nur zentrale Aspekte zwischenmenschlicher Zusammenarbeit und des Gruppenzusammenhalts, sondern sie haben auch noch wichtige Folgen für die individuelle Gesundheit und Lebenserwartung. Es ist jedoch noch nicht erforscht, welche sozialen Eigenschaften anderer das Vertrauen fördern, wie Individuen lernen wem sie Vertrauen schenken sollen und wie das Gehirn solche Informationen für eine optimale Verhaltensanpassung integriert. Hier werde ich zwei Forschungslinien auslegen. Einerseits zeige ich die psychologischen und neuronalen Grundlagen von Vertrauen in unterschiedlichen sozialen Kontexten. Andererseits wird eine pharmakologische Modulation von Vertrauensakten zugrundeliegenden Hirnstrukturen dargelegt. In den ersten beiden Verhaltensexperimenten zeige ich, dass Ehrlichkeit einem Vertrauenswürdigkeitseindruck vorausgeht und dass ein Ruf, ehrlich zu sein (d.h. ehrlicher Ruf), mit höherem Vertrauen während einer zukünftigen sozialen Interaktion verknüpft ist. Als Nächstes stelle ich die Hirnmarker dieses Ehrlichkeit-basierten Vertrauenswürdigkeitseindrucks dar. Insbesondere fand ich heraus, dass Ehrlichkeit enkodierende Hirnregionen zukünftige Vertrauensentscheidungen vorhersagen, ähnlich wie die Verhaltenseffekte von Ehrlichkeit auf zukünftige Vertrauensentscheidungen. Dies liefert Evidenz für Ehrlichkeit zugrundeliegende Hirnareale, die die Vertrauensverhalten vorhersagenden Hirnsignal beinhalten. Außerdem wirkt sich der ehrliche Ruf eines anderen auf Feedback verarbeitende Hirnaktivierungen aus, was Informationsintegration während sozialen Lernens verzerren kann. In einer weiteren Verhaltensstudie zeige ich, dass ein ehrlicher Ruf Lernprozesse mittels einer durch Ehrlichkeit verursachten Asymmetrie in Informationsgewichtung zu beeinträchtigen scheint. Schließlich demonstriere ich, wie pharmakologische Modulation von Hirnprozessen Vertrauensverhalten

aber nicht Vertrauenswürdigkeitseindrücke beeinflusst. Zum einen werfen diese Erkenntnisse ein Licht darauf, wie Ehrlichkeit das Vertrauen in andere verstärkt, aber auch wie dies Lernprozesse für optimale Verhaltensanpassung erschweren kann. Zum anderen liefern sie erste pharmakologische Evidenz dafür, wie auf Eindrücken basierendes Vertrauen verändert werden kann, ohne Vertrauenswürdikeitseindrücke zu beeinflussen. Zum Schluss schlage ich Ansätze vor, die dem beobachteten Verhalten und den Hirnaktivierungsmustern eine Erklärung bieten, und entwerfe zukünftige Richtungen für neue Untersuchen.

List of Studies

Bellucci G, Münte T F, Park S Q, *Resting-state dynamics as a neuromarker of dopamine administration in healthy female adults*, J Psychopharmacol, doi: 10.1177/0269881119855983, 2019

Bellucci G, Molter F, Park S Q, *Neural representations of honesty predict trust*, Nature Communications, (in revision, 2nd round, decision letter on minor revision received on August 9, 2019)

Bellucci G, Park S Q, *Honesty biases trustworthiness impressions*, Journal of Experimental Psychology: General, (under review, submitted on March 31, 2019)

Bellucci G, Münte T F, Park S Q, *Effects of a dopamine agonist on trusting behaviors in females*, Psychopharmacology, (under review, submitted on July 18, 2019)

Chapter 1: Introduction

1.1 To trust or not to trust

Across disciplines, trust is defined as the willingness to accept vulnerability based on positive expectations of the intentions and behavior of another (Rousseau, Sitkin, Burt, & Camerer, 1998). Accepting some degree of vulnerability to the other is essential to trust, because the other's behavior is not fully under our control and the other may thus take advantage of us (Mayer, Davis, & Schoorman, 1995). If we want to avoid being exploited by our fellow humans, we might want to engage only in social interactions where we can foresee the outcomes. However, not only is such a degree of control and predictability utterly impossible in real-life interactions, it also hampers cooperation and the ability to learn from others. Cooperation is crucial to navigating a highly complex world. Humans cooperate for help, protection and support (Bicchieri, 1990; Cubitt, Gächter, & Quercia, 2017; E. Fehr & Schmidt, 1999). Further, humans have developed a variety of social learning strategies that enable them to maximize their survival chances by using the knowledge and behavioral patterns of others (Kendal et al., 2018).

Relying on others for help, information, support and the like, may turn out to be highly advantageous. In particular, we might be able to exploit others' knowledge and experience to improve accuracy and speed up of our decision-making. Exploiting others' knowledge represents a central feature for better learning and decision strategies, as it prevents the implementation of more costly exploratory approaches to acquire the required information before making a decision or action (which may be, for instance, the way culture operates) (Heyes, 2016; Kendal et al., 2018). This has led to a "social bias" in information gathering (e.g., by taking advice from others) where individuals prefer to sample information from others over gathering information by themselves (Mesoudi, Whiten, & Dunbar, 2006). Although such a bias may improve the accuracy of our decision and boost our survival chances especially when gathering information from groups of individuals (Bang & Frith, 2017; Galton, 1907), relying on information from others may imply less optimal choices for oneself and more vulnerability to the other's exploitation when this information is itself somehow biased or comes from not well-intended others.

As trust outweighs distrust, humans may have evolved to adopt trust as a default strategy in social interactions, accepting vulnerability to others for the sake of the advantages of social learning and cooperation (Cesarini et al., 2008; Oskarsson, Dawes, Johannesson, & Magnusson, 2012; Reimann, Schilke, & Cook, 2017). At the same time, however, they may have refined tools to "learn" to distrust, namely, to identify untrustworthy others who need to be avoided or ostracized (Reimann et al., 2017; Twenge, Baumeister, DeWall, Ciarocco, & Bartels, 2007; Williams & Sommer, 1997). This, in turn, requires an accurate estimation of the other's character to make adequate predictions about the quality of the other's information or about the reliability of the other's future, cooperative behavior.

Trust might be required both when we do not know the trustee and when we do (**Fig. 1**). When the social interaction is completely anonymous (like many interactions on the Internet), trust relies on social norms recognized by the group and applicable to a particular social circumstance. In such situations, individuals trust on the assumption that the other would comply with a norm of fairness and reciprocity (Bellucci, Feng, Camilleri, Eickhoff, & Krueger, 2018; Berg, Dickhaut, & McCabe, 1995; Bicchieri, 2005, 2014). If, however, the interaction is not completely anonymous, individuals have access to partial information about the other and can form trustworthiness beliefs based on first impressions that emerge effortlessly and rapidly (Engell, Haxby, & Todorov, 2007; A. Todorov, 2008; A. Todorov, Baron, & Oosterhof, 2008; A. Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Alexander Todorov, Pakrashi, & Oosterhof, 2009).

On the contrary, individuals can also have some knowledge about the other prior to trust. On the one hand, individuals can receive information indirectly from others (e.g., indirect reputation). This is the case when people have information about the trustee's previous social behavior or have heard about the trustee's reputation as social partner (Delgado, Frank, & Phelps, 2005; Fouragnan et al., 2013; Hillebrandt, Sebastian, & Blakemore, 2011; Semmann, Krambeck, & Milinski, 2004). On the other, people might have the opportunity to repeatedly interact with the trustee over time. In this case, individuals slowly gather information via direct experience to form subjective, trustworthiness beliefs about the other (Bellucci, Chernyak, Goodyear, Eickhoff, & Krueger, 2017; I. Bohnet & Huck, 2004; Heyes, 2016; Kendal et al., 2018; King-Casas et al., 2005; Krueger, Grafman, & McCabe, 2008; Krueger et al., 2007). Moreover, when individuals have the opportunity to interact over the course of multiple encounters, these different sources of social information can also influence each other. For instance, through direct experience with the other, the initial, trustworthiness beliefs based on first impressions or indirect reputation might be revised to form more precise beliefs (e.g., direct reputation).

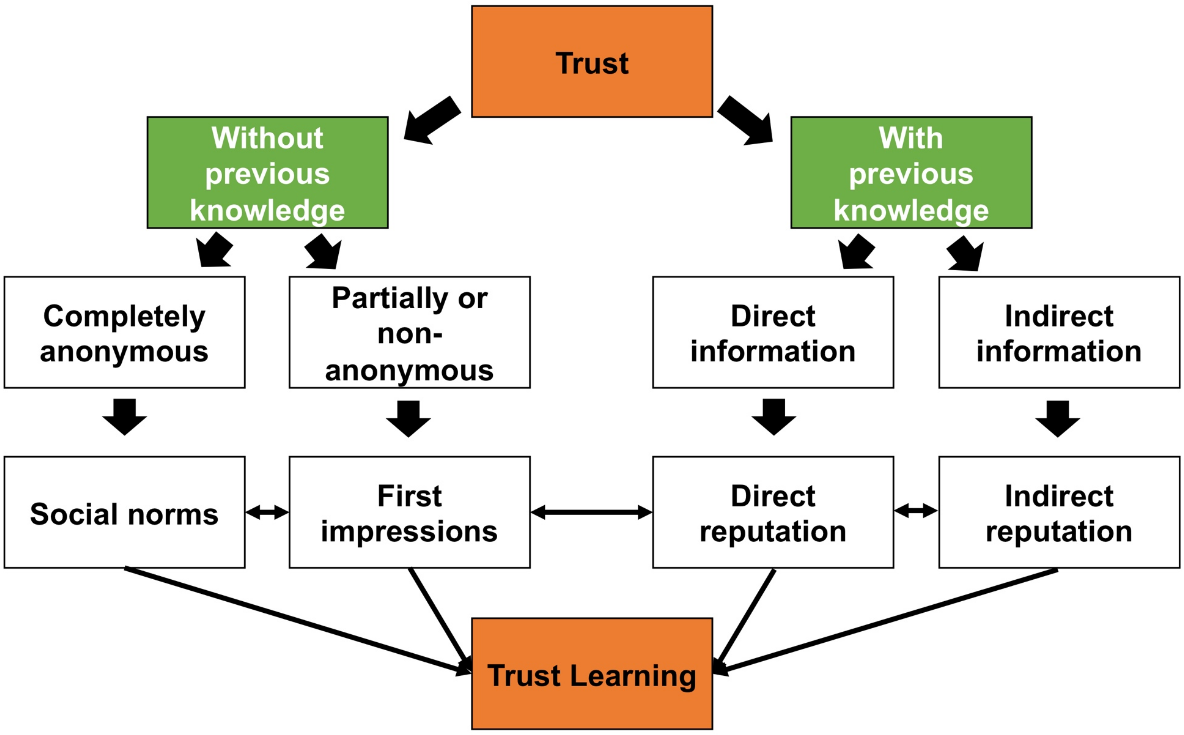


Fig.1. Trusting interactions. Individuals face very different situations in which they would need to trust others. Individuals might be required to put their trust in either unknown or known others, namely, individuals whom they have no previous knowledge of or individuals whom they know. Trusting unknown others might be required in anonymous situations where individuals do not have the opportunity to either meet or see the trustee. In these cases, in order to decide whether to trust, individuals rely on the social norms recognized by the social group they are a member of and applicable to the particular situation they are in. When, however, the social interaction is not completely anonymous, individuals might have access to some (often partial) information about the other. In these cases, individuals can form first impressions about the trustee based on the available information. On the contrary, when individuals trust known others, they might face a trustee they know directly, for instance, through previous experience. In these cases, individuals have beliefs about the other's trustworthiness that ground in the other's direct reputation. In other cases, however, individuals might not have interacted with the other previously but have knowledge about the other's trustworthiness through indirect reputation, for instance, because they have heard what others think of the trustee. Importantly, these different sources of social information might dynamically interact to form, change, revise or update trustworthiness beliefs about others, underlying mechanisms of trust learning.

Here, I will focus in particular on impression-based trust derived from facial information and experience-based trust derived from direct reputation. First impressions can draw on different types of social information from faces (Olivola & Todorov, 2010; Sofer, Dotsch, Wigboldus, & Todorov, 2015; A. Todorov et al., 2015). Previous work has shown that different social dimensions can be inferred from faces, such as, attractiveness, competence, trustworthiness and dominance (A. Todorov et al., 2008; A. Todorov, Mandisodza, Goren, & Hall, 2005; A. Todorov et al., 2015). Such information is processed as quickly as 33 ms and can influence a variety of complex social behaviors from voting to trusting (A. Todorov, 2008; A. Todorov et al., 2005; Alexander Todorov et al., 2009).

In contrast, repetitive social interactions allow abstracting behavioral patterns that constitute reputational priors about the other's character and social preferences. The ABI model, an influential model of trust based on a cross-discipline meta-analysis of the literature on trust, suggests three main factors of trustworthiness: 1) ability; 2) benevolence, and 3) integrity (Mayer et al., 1995). These factors induce perceptions of trustworthiness in others that guide social and prosocial behaviors such as advice-taking, altruistic behavior, unconditional kindness and reciprocity (Ashraf, Bohnet, & Piankov, 2006; Baumert, Schlösser, & Schmitt, 2014; Hilbig, Thielmann, Hepp, Klein, & Zettler, 2015; Thielmann & Hilbig, 2015; Yaniv, 2004; Yaniv & Kleinberger, 2000).

1.2 Taking advice

Seeking advice from others is a highly efficient strategy through which individuals exploit others' knowledge to improve their decisional outcomes (Yaniv, 2016). Sound decision-making is of pivotal importance to the individual because suboptimal choices are perilous to one's survival chances. Individuals have been shown to take advice from different sources and make their decision after integrating this information.

On the one hand, individuals have been seen to rely on their own knowledge more frequently and more knowledgeable individuals take less advice than less knowledgeable individuals—a phenomenon referred to as egocentric advice discounting (Yaniv, 2004, 2016; Yaniv & Kleinberger, 2000). However, in highly unpredictable environments, individuals prefer gathering more information from others, deliberately seeking to acquire knowledge about their past decisions and behaviors (McElreath et al., 2005). Moreover, individuals are also sensitive to the quality of the advice and tend to discount poor advice, suggesting that an accuracy maximization strategy plays a role in advice taking (Budescu & Rantilla, 2000; Yaniv & Kleinberger, 2000). Advice-taking behaviors are also sensitive to the social qualities of the adviser, so that less advice is generally taken from those who do not reciprocate in advice taking (Mahmoodi, Bahrami, & Mehring, 2018; McElreath et al., 2005). In sum, previous work suggests that individuals reach out to others less when uncertainty (about one's decisions or the environment) is low and rely more strongly on what they know when making decisions.

However, when they do seek advice, individuals weigh up the social qualities of the adviser to decide whom to take advice from (Festinger, 1954).

So far, previous work has focused on how and when advice from others affects an individual's learning. However, little is known about how character traits of the adviser are learnt and impact the advisee's decisions and advice-taking behavior. Of particular relevance are perceptions about the adviser's trustworthiness, which might ultimately lead someone to decide whether to take advice and from whom. The trustworthiness of advisers is indeed central to many decisions in everyday life. Recently, it has been suggested that the patients' trustworthiness impressions about their doctor might play a substantial role in the patient's health and life expectancy (Baker et al., 2016; Pereira Gray, Sidaway-Lee, White, Thorne, & Evans, 2018). Higher trust in a doctor might increase patient compliance even when positive outcomes are not readily and clearly predictable, resulting in the long-term benefit of the patient. However, to date, no study has experimentally tested whether and how direct reputation affects trustworthiness impressions and trusting behaviors that guide advice-taking strategies in interactions with others.

1.3 Trusting once and again: The Trust Game

Behavioral trust has been widely investigated across contexts (from economics to psychology to neuroscience) using the well-established trust game (TG) (Berg et al., 1995). The TG is a two-player game. One of the players receives the role of investor. The other player is the trustee. The investor receives an initial, monetary endowment. The investor must decide whether to share part of this initial endowment with the trustee. In some versions of the TG, investors make a binary decision, namely, they can decide either to share nothing or to share half of the initial endowment. Other versions allow for more variability in the investor's decisions, enabling investors to share any amount from nothing to the full amount. Unanimously, the investor's decision is taken as a measurement of the investor's trust. If the investor decides to share any portion of the initial endowment, the shared amount of money is multiplied by the experimenter (generally tripled) and passed on to the trustee. The trustee can then decide whether any portion of the received amount should be sent back to the trustor. The trustee's decision represents the trustee's reciprocity to the investor. Either the absolute amount shared back by the trustee or its proportion relative to the total amount at the trustee's disposal can be used as a measure.

Further, two versions of this game are generally employed. In one version, investors and trustees interact only once (one-shot TG). In many studies employing this version, especially in the functional magnetic resonance imaging (fMRI) literature, participants in the role of investor in general play several rounds of one-shot TGs. That is, investors make repeated trust decisions, yet each time with a different trustee. Thus, the social interaction with each partner lasts per se only one round (hence, one-shot decision). In contrast, in the multi-round TG, investors play multiple rounds (i.e., make multiple trust decisions interacting) with the same trustee. Studies have also examined trusting behavior in investors and trustees playing several multi-round TG, that is, each partner plays multiple rounds with multiple partners. The main advantage of this second TG version is that it allows studying the temporal dynamics of trust in the course of the social interaction. This makes it possible to investigate how trust is established and maintained, evolves, and ultimately breaks down.

The game has reached such popularity because it establishes a highly controlled environment in which a social interaction between people unfolds that has good ecological validity. Furthermore, the TG provides very similar results under varying conditions and across cultures (I. Bohnet, Greig, Herrmann, & Zeckhauser, 2008; Johnson & Mislin, 2011). Based on these studies, investors have been seen to consistently share around 50% of their initial endowment and trustees reciprocate by sending back as much as entrusted to them (C. F. Camerer, 2003a, 2003b, 2003c; Johnson & Mislin, 2011). Studies using the TG have shown that trust in others hinges on both the character and intentions of the other partner. For instance, individuals trust those who have a morally good character, show good intentions or are good cooperators (Delgado et al., 2005; Falk, Fehr, & Fischbacher, 2008; McCabe, Rigdon, & Smith, 2003; Nelson, 2002; van 't Wout & Sanfey, 2008).

The flexibility of the TG lends itself to study different trusting dynamics. Subjective impressions of others' trustworthiness (impression-based trust) has been studied using the one-shot TG. The investor is provided with information to elicit varying trustworthiness impressions prior to a trust decision. Impressions can be elicited by providing vignettes or story lines describing the character of the trustee. Alternatively, participants could interact with their partner in a previous game in which the partner's behavior was modulated to induce impressions of high or low trustworthiness. Finally, impressions of trustworthiness can be triggered by presenting pictures of putative partners whose facial trustworthiness is manipulated ad-hoc.

In contrast, trusting behaviors that rely on the learnt social character of the other (i.e., experience-based trust) can be studied using multi-round TG. Participants interact with each other repeatedly, thereby learning dynamically from each other's decisions and which action may be the best one in the next encounter. Trustees' reciprocal behaviors may be manipulated to examine how individuals adapt their trusting behavior to the other's trustworthiness. In some cases, reciprocal behaviors differed across trustees but remained constant across time. In some other cases, the reciprocal behavior of a trustee changes over time. Others have examined how trust breaks down when healthy individuals interact with patients with social impairments who do not exhibit the same depth of mentalizing to form an adequate model of the partner (Anderl et al., 2018; Maurer, Chambon, Bourgeois-Gironde, Leboyer, & Zalla, 2018; Xiang, Ray, Lohrenz, Dayan, & Montague, 2012).

1.4 Reinforcement learning

Thus, beliefs about the other's trustworthiness may be formed based on subjective impressions from facial information and/or on dynamic integration of new incoming information about the other's behavior that updates trustworthiness beliefs about them. Previous studies have suggested that cues about others' behavior during repeated interactions are integrated via reinforcement learning processes. Reinforcement learning relies on prediction errors that signal the discrepancy between actual and expected rewards (Rudebeck, Saunders, Prescott, Chau, & Murray, 2013; Tsuchida, Doll, & Fellows, 2010). The Rescorla Wagner model posits that the prediction error reflects how much learning occurs based on the unexpected reward (i.e., surprise) (Rescorla & Wagner, 1972). The model formalizes learning as update of the value ascribed to a particular stimulus *V* at a particular time *t* based on the received reward *R*:

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

where PE_t is the prediction error at time *t*. The PE approximates zero when the received reward is close to what expected. The PE increases the more the received and expected rewards diverge. The PE is negative when the received reward is smaller than expected, whereas the PE is positive when the received reward is bigger than expected.

The PE can be thought of as the quantity that determines how much update is needed. The more we learn about the associative strength between a particular stimulus and its reward outcomes, the less learning occurs, as the expected reward approximates the actual reward. This implies that our expectation (prediction) of a reward R given a stimulus S will increase in accuracy with a concomitant reduction of discrepancy (error). However, this value update is not linear but hinges on learning parameters, which affect the magnitude of the changes involved. In their original formulation, Rescorla and Wagner describes at least three sets of parameters that modulates value update (Rescorla & Wagner, 1972). These were: 1) the stimulus salience (α), which indicates the assumption that the associative strength between stimulus and reward may be acquired at different rates despite equal reinforcement; 2) the learning rate associated with a particular stimulus (β), which indicates the assumption that the rate of learning depends on the type of stimulus employed; and 3) the asymptotic level of the associative strength (λ), which describes the associative strength of the stimulus-reward pair. Based on these assumptions, the amount of learning that occurs for each stimulus (i.e., the change of stimulus value, ΔV_s) can be formalized as:

$$\Delta V_s = \alpha_s \beta_s (\lambda_s - V_s). \tag{2}$$

Neuronally, PEs evoke teaching signals for changing synaptic weights in dopaminergic neuronal networks (Sutton & Barto, 1981). Thus, similar to the mathematical formulation of PEs, an unexpected outcome leads to a positive neural signal, a predicted outcome to a zero neural signal and the absence of an expected outcome to a negative neural signal. PEs are encoded, among others, in the midbrain, striatum and orbitofrontal cortex (OFC) (W. Schultz, 2000; W. Schultz & Dickinson, 2000). Notably, these dopaminergic brain regions show neural responses to rewards that depend on their predictability, suggesting that they track reward PE for learning. For instance, Hollerman and Schultz (1998) found that the magnitude of dopamine responses to a reward reflected the degree of reward predictability during individual learning. Responses to unexpected rewards were stronger and decreased with improved performance (i.e., better reward predictions). Moreover, dopamine neurons signal not only the occurrence of a reward but also its timing relative to expectations (Hollerman & Schultz, 1998; W. Schultz, 2000; W. Schultz, Dayan, & Montague, 1997).

Neuroimaging studies in humans have found similar results. A pioneering positron-emission tomography study on stimulus-outcome associations found activations in bilateral OFC during the occurrence of unexpected outcomes (Nobre, Coull, Frith, & Mesulam, 1999). More recently, a neuroimaging study using fMRI has indicated a certain degree of neural differentiation in the representation of different types of PEs. In particular, while value PEs (related to the magnitude of an expected reward) elicit brain activations specifically in the midbrain, identity PEs (related to the sensory features of an expected reward) evoke brain activity not only in the midbrain but also in the OFC (Howard & Kahnt, 2018).

Reinforcement learning models were initially applied to describe how an agent learns the associative strength of two stimuli during instrumental learning (i.e., based on Pavlovian conditioning). More recently, these models have been applied to learning of other forms of stimulus values, such as a person's character traits in social learning. In particular, reinforcement learning has been proposed to allow the formation of beliefs about another that ultimately inform trusting behaviors (Delgado et al., 2005; Fouragnan et al., 2013; King-Casas et al., 2005). Trustworthiness beliefs are dynamically updated based on feedback about the other's reciprocity over multiple interactions with the partner in a TG (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010). Similarly, in an advice-taking paradigm, participants integrate advice by weighting the different outcomes of the recommended and not recommended options (Biele, Rieskamp, & Gonzalez, 2009).

Activity in the OFC is associated with valuation of expert advice before an advice-based decision is made, suggesting that on a neural level, a reinforcement learning mechanism is likely involved in advice utilization (Meshi, Biele, Korn, & Heekeren, 2012). These results provide behavioral and neural evidence that reinforcement-learning models might be best suited to capture the dynamics of trust-based learning in social contexts. In particular, they might provide a mechanistic account to formally describe how individuals learn about another's character traits and reputation before deciding whether to trust.

1.5 The role of dopamine in trust

Dopaminergic neurons are a heterogeneous group of cells situated in the diencephalon, mesencephalon, and olfactory bulb with a large majority localized in the ventral midbrain (Arias-Carrion, Stamelou, Murillo-Rodriguez, Menendez-Gonzalez, & Poppel, 2010; Bjorklund & Dunnett, 2007; Ikemoto, 2010). Although they make up only roughly 1% of all neurons in a human brain, they play a central role in a wide range of human behaviors (Arias-Carrion et al., 2010). Animal and human studies together with clinical investigations have suggested roles of the dopaminergic system in movements, goal-directed behavior, cognition, attention, reward and, as seen in the previous section, reinforcement learning (Boureau & Dayan, 2011; Cools, 2006; Cools, Nakamura, & Daw, 2011; J. P. O'Doherty, 2004; Wolfram Schultz, 2002).

In social contexts, many studies have observed activations in the dopaminergic system for a broad variety of behaviors. For instance, activity in dopaminergic regions has been observed for altruism (Karns, Moore, & Mayr, 2017), charity donations (Moll et al., 2006) and generosity (Hackel, Doll, & Amodio, 2015). A recent fMRI study has further observed that activity in the striatum and OFC during generosity is associated with increased happiness feelings (Park et al., 2017), suggesting a central role for the dopaminergic system not only in prosocial behaviors but also in subjective well-being. Trust has been observed to engage dopaminergic regions such as the striatum and OFC as well. For instance, trust decisions with a social partner as opposed to a computer evoke brain activations in the putamen and ventral striatum. Striatal and OFC responses were further observed to reciprocated trust as opposed to defection of trust (Phan, Sripada, Angstadt, & McCabe, 2010; Sripada et al., 2009).

However, the role of the dopaminergic system in such complex behaviors is yet to be clarified. Especially with respect to the literature on trust, the engagement of the dopaminergic system might be related to other cognitive mechanisms that play an essential, but lateral role in trusting behavior, or might be due to methodological choices of researchers, such as how trust is studied and operationalized.

A pioneering hyperscanning (simultaneous dual-brain) fMRI study (King-Casas et al., 2005) showed that even though neural signals from the caudate initially occurred after each repayment amount was revealed (likely reflecting reward processing), activations shifted over the course of the TG and began to peak before the repayment amount was revealed. Consistent with prediction error signals central to reinforcement learning, these results suggest that neural activity in the caudate underwent a dynamic change in functional role from its early involvement in the response to the received reward outcome to its later involvement in the prediction of the reward outcome based on the investor's actions (King-Casas et al., 2005).

Furthermore, another fMRI study has shown that during trust decisions, individuals recruit prefrontal brain regions involved in mentalizing at the beginning of a social interaction but engage well-known dopaminergic structures in later phases (Krueger et al., 2007). Thus, when interacting with unknown partners, individuals rely more strongly on cognitive processes that aid the formation of beliefs about the partner's character. These beliefs represent reputational priors about the other that are retrieved to support decisions and updated on the basis of prediction errors over time. Accordingly, the engagement of dopaminergic regions in repeated interactions with known others might reflect reinforcement learning mechanisms.

Notably, activity in the caudate during trust decisions is dampened by both positive and negative moral priors, whereas the absence of a prior does not reduce brain activity in this region (Delgado et al., 2005; Fareri, Chang, & Delgado, 2012). These findings suggest that neural responses in the caudate during a decision to trust reflects a learning signal to update one's behavior from feedback on the other's actions. This learning signal is diminished when information about the other's moral character is provided. A recent fMRI study has directly tested the hypothesis that neural signal in the striatum represents a learning mechanism related to updating one's behavioral strategy in a social interaction with a trusting partner (Vanyukov, Hallquist, Delgado, Szanto, & Dombrovski, 2019). The authors compared the results of multiple computational models and showed that activity in the striatum could be best captured by a model that describes one's action policy and is sensitive to counterfactual outcomes of one's untaken actions. It was concluded that activity in the striatum closely tracks the success of one's behavioral strategy (i.e., whether and when to trust) for learning and optimal behavioral adaptation to the other's actions (Vanyukov et al., 2019).

Thus, when individuals focus on the consequences of a trust decision, brain regions signaling actual or hypothetical decision outcomes (likely related to reinforcement learning signals) are recruited in trusting interactions (Bellucci et al., 2017; Bellucci et al., 2018; Chang, Smith, Dufwenberg, & Sanfey, 2011; Delgado et al., 2005). However, creating a context in which participants have to evaluate the other's character before making a trust decision might engage other cognitive processes that recruit a different set of brain regions associated with higher-order cognition. It follows that if trust draws on the social character of the other (e.g., whether the other is trustworthy or honest), brain regions associated with social evaluations (such as the ventromedial prefrontal cortex, VMPFC, and dorsolateral prefrontal cortex, DLPFC) and inferences on the other's intentions (e.g., the posterior temporoparietal junction, pTPJ) should be engaged during trusting behaviors (Buckholtz et al., 2008; Cooper, Kreps, Wiebe, Pirkl, & Knutson, 2010; FeldmanHall et al., 2018; Tusche, Bockler, Kanske, Trautwein, & Singer, 2016). However, to date, evidence on the brain regions representing the other's character traits (such as trustworthiness or honesty) and predictive of trust decisions is still missing.

1.6 Limitations of the extant literature

Both advice-taking paradigms and the different versions of the TG present some issues. On one hand, the quality of the advice has mainly been operationalized through the reward magnitudes associated with the advice (Behrens, Hunt, Woolrich, & Rushworth, 2008; Biele et al., 2009; Biele, Rieskamp, Krugel, & Heekeren, 2011; A. O. Diaconescu et al., 2014; Andreea O. Diaconescu et al., 2017; Meshi et al., 2012; Rodriguez Buritica, Heekeren, & van den Bos, 2019; Yaniv & Kleinberger, 2000). In other words, good, informative advice was generally associated with higher reward outcomes and poor, uninformative advice with smaller reward outcomes.

However, advice can be informative without it being the best or near-best option (e.g., advice not to do something or how to make a decision) (Dalal & Bonaccio, 2010). This information-reward confound may have reduced social information processing to reward processing in previous studies. That is, in previous studies, learning and cognitive processes related to estimations of reward outcome contingencies are difficult to disentangle from evaluations about the partner's character, such as their competence, honesty and generosity. Further, in advice-taking paradigms, advisers are generally incentivized to give accurate advice (Bonaccio & Dalal, 2006). This might have indeed preserved the face-validity of the experiment, but it might also have encouraged participants to track the partner's motives (Behrens et al., 2008; A. O. Diaconescu et al., 2014), disincentivizing the learning of the other's character.

Similarly, in the TG, trust might be associated with other behaviors that might arise from causes other than the partner's trustworthiness, such as one's own benefits associated with the act of trust (Dirks & Ferrin, 2002; Kramer, 1999). Indeed, previous studies have observed that trust ceases when external incentives are no longer available or when trust leads to monetary losses (Jason A. Aimone & Houser, 2012; Rode, 2010). These findings indicate some sort of strategic, reward-driven thinking intertwined with trust decisions in the TG (C. F. Camerer, 2003a).

Further, sharing behaviors in the one-shot TG might reflect cognitive mechanisms other than trust. For example, in many studies, investors and trustees start the game with a difference in monetary budget with investor being endowed with a certain monetary amount, whereas trustees having no monetary endowment. In this situation, investors might consider sharing out of other-regarding concerns (J. C. Cox, 2002; James C. Cox, 2004) or might feel obliged to comply with a fairness norm and share enough to counterbalance the initial inequality of money distribution (C. Camerer & Fehr, 2004; Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007). If trust decisions in the TG are confounded by these factors unrelated to trust, it is difficult to uniquely assign neural activity observed during those decisions. Thus, neural activations in the anterior insula in the one-shot TG might be associated with uncertainty, betrayal aversion or anticipation of hypothetical norm violation by the partner (Bellucci et al., 2018; Chang et al., 2011; Engelmann, Meyer, Ruff, & Fehr, 2019). Similarly, activations in the striatum in the multi-round TG are as likely evoked by reward anticipation or learning mechanisms as by trust (Bellucci et al., 2017).

1.7 Overcoming limitations

Thus, the outstanding question is how can we disentangle trusting behaviors from factors not related to trust? One possible solution might be to specifically modulate participants' impressions about others' trustworthiness and investigate how such a manipulation impacts trust. For instance, impression-based trust can be manipulated by presenting faces that vary on the trustworthiness dimension. A pharmacological intervention might be employed to investigate how dopamine affects individual trust. This approach might be quite effective because previous work has already shown that trustworthiness impressions are not impacted by dopaminergic manipulation (Zebrowitz et al., 2018). Thus, differences in how trustworthy faces lead to trust decisions might be uniquely attributed to the dopaminergic modulation.

However, when using faces, one important caveat must be addressed. Different types of social information can be inferred from faces such as facial attractiveness, which plays a role in many social behaviors like approach behaviors and partner choice (Fisher, Aron, & Brown, 2005; Little, Jones, Penton-Voak, Burt, & Perrett, 2002; Olson & Marshuetz, 2005). Facial attractiveness can also explain a significant portion of variance in trusting behaviors in the TG (Stirrat & Perrett, 2010; R. K. Wilson & Eckel, 2006). Moreover, facial attractiveness is not only highly correlated with facial trustworthiness (A. Todorov, 2008; A. Todorov et al., 2008) but also evokes reliable neural activations in the dopaminergic system (Aharon et al., 2001). Yet previous work investigating impression-based trust with faces failed to control for variations in facial attractiveness. Thus, although facial trustworthiness plays a role in both building trust in strangers and in experience-based knowledge about the other's trustworthiness (Chang et al., 2010), the specific effects of facial trustworthiness on trust are yet to be explored. Minimizing facial attractiveness information might help investigate the peculiar effects of trustworthiness impressions on trust in unknown others and might allow for the investigation of the specific dopaminergic effects on trusting behavior.

An alternative approach might be to induce trustworthiness impressions through a previous interaction with the other and use the TG later on as a read-out for the transfer effect of established trust (Redcay & Schilbach, 2019), thereby eliciting trustworthiness impressions that steer subsequent trust decisions. For example, benevolent and competent partners are trusted more, and recent research has shown that guilt-proneness makes people more likely to be trustworthy (Burnham, McCabe, & Smith, 2000; Delgado et al., 2005; Falk et al., 2008; Levine, Bitterly, Cohen, & Schweitzer, 2018; Toelch, Bach, & Dolan, 2014; van 't Wout & Sanfey, 2008). Recently, honesty has been seen to play a central role in different social behaviors and suggested as possible antecedent of trustworthiness perceptions (Ashton & Lee, 2007; Ashton, Lee, & de Vries, 2014; Lee & Ashton, 2004). Hence, honesty in advice-giving might be used as a proxy for the other's trustworthiness that likely guides participants' trust decisions later on.

Finally, it is of crucial to control the well-structured economic incentives inherent to most investigations of trust in iterative social interactions. Previous studies have shown that a lack of monetary incentives results in drastically dropping trust rates (J. A. Aimone & Houser, 2011; Jason A. Aimone & Houser, 2012; J. A. Aimone & Houser, 2013; Johnson & Mislin, 2011; Rode, 2010). At the same time, neural signals in striatal regions are elicited by the repayment amount sent back by the partner (Bellucci et al., 2017; Phan et al., 2010). These findings suggest that monetary incentives might drive participants' behavioral choices and evoke the observed activations in dopaminergic regions. Thus, eliminating external incentives in experimental paradigms might both help to disentangle reward-driven, strategic choices from social behaviors (e.g., reputational concerns and reciprocal motives) and might also provide an ecological setting able to capture neural correlates of real-life trust decisions.

Chapter 2

2.1 Research objectives and hypotheses

In the current dissertation, I show how subjective impressions can be harnessed to investigate 1) how beliefs about others' trustworthiness are updated and 2) how a pharmacologically-induced modulation of the dopaminergic system affects trust decisions. In the first part of this dissertation, I describe a novel paradigm (the take advice game, TAG) designed to disentangle social information from reward information and to induce trusting behaviors based on others' honesty. Using this paradigm, I present two behavioral studies (Chapter 4-5) followed by a related fMRI study (Chapter 6) and a final behavioral study in combination with computational modeling (Chapter 7).

The objective of these studies is to investigate the relationships between an antecedent of trustworthiness perception (i.e., honesty) and trust. Given preliminary evidence that honesty elicits a wide array of prosocial behaviors and given the similarity of honesty to the concept of integrity, which has been hypothesized to lead to trust (Ashton & Lee, 2007; Ashton et al., 2014; Ashton et al., 2004; Lee & Ashton, 2004; Mayer et al., 1995), I hypothesized that honest behavior might evoke trustworthiness impressions that guide later trust decisions. Furthermore, in the fMRI study, I addressed the issue as to whether neural patterns of trustworthiness impressions elicited in the TAG are also able to predict future trusting behaviors and can be disentangled from reward-related signal. In particular, I hypothesized that a stronger integration of honesty-related information about the other's behavior should make participants more willing to trust. Finally, in a last behavioral study, I will show how reputational priors of honesty-based trustworthiness are formed and updated using a computational formalization of participants' choices. As reinforcement models have previously been shown to closely capture social learning dynamics (Biele et al., 2009; Biele et al., 2011; Chang et al., 2010), I hypothesized that these models might well describe how participants learn to trust from information on the other's honesty.

In the second part of this dissertation, I address the question of whether the formation of trustworthiness impressions rely on dopaminergic functioning with the help of a pharmacological intervention in combination with an fMRI study. Using different measures of resting-state functional connectivity, I demonstrate how pramipexole, a D2/D3 dopamine agonist targeting well-known dopaminergic brain structures (Ishibashi, Ishii, Oda, Mizusawa, & Ishiwata, 2011; Riba, Kramer, Heldmann, Richter, & Munte, 2008), impacts neural activity of specific brain structures at rest (Chapter 8). With the help of this pharmacological manipulation, the question as to whether pramipexole modulates impression-based trust in unknown others is addressed (Chapter 9).

The objective of these studies is to examine the engagement of the dopaminergic system in trusting behaviors. Given previous evidence that pramipexole targets a specific subset of brain dopamine regions in subcortical and sensorimotor structures (Ishibashi et al., 2011; Riba et al., 2008; Ye, Hammer, & Munte, 2017), I hypothesized that pramipexole administration impacts resting-state brain dynamics of specific functional connectivity networks such as the cinguloopercular (involving subcortical brain regions) and sensorimotor (involving motor and sensorimotor brain regions) networks. Having the same participants play a subsequent one-shot TG after pramipexole administration, I describe how pramipexole affects trusting behaviors based on subjective trustworthiness impressions from faces. By maximally varying facial trustworthiness and minimizing variations in facial attractiveness, I hypothesized that pramipexole administration impacts trusting behavior independently of the modulation of subjective impressions about others' facial trustworthiness.

Chapter 3

3.1 Methodology

3.1.1 Experimental samples

A total of 150 participants were collected for the studies in this dissertation. Study 1 and Study 2 were run in the laboratory on a computer with procedures that would make them suitable for fMRI investigations. Study 3 employed one of these lab paradigms in an fMRI experiment, which, in combination with multivariate decoding analyses, predictive analytics and functional connectivity analysis, examined the neural underpinnings of honesty-based trustworthiness impressions. Finally, a last version of this paradigm was conducted in a behavioral experiment (Study 4), where the computational dynamics of social character learning were examined using reinforcement learning computational models.

Study 5-6 of this dissertation investigated the effects of a dopamine agonist on neural dynamics and impression-based trust. Combining resting-state functional connectivity with multivariate classification and prediction analyses, Study 5 examines the effects of pramipexole on resting-state neural dynamics and the relationships between this neural modulation and pramipexole's effects on attractiveness evaluations. Using a one-shot TG, Study 6 addressed the question as to whether trust in unknown others based on subjective trustworthiness impressions from faces can be modulated by pramipexole administration.

3.1.2 Study 1-4

Twenty-eight participants (18 females; 21.43±3.47, mean age±SD) were invited to the lab for Study 1. G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) was used to calculate the desired sample size, based on the effects of others' moral character on trustworthiness perceptions in a previous study (Delgado et al., 2005). Twenty-eight participants (18 females; 24.54±4.0, mean age±SD) were invited in Study 2 based on the effect size of Study 1. In Study 3, data from 31 participants were acquired (20 females; 24.29±3.81, mean age±SD). This fMRI study was conducted at the Free University Berlin. In Study 4, a sample size similar to the previous behavioral studies was aimed at, ending up with a final sample of 33 participants (23 females; 22.27±3.13, mean age±SD).

For all three studies, exclusion criteria were: 1) present or past neurological and psychiatric disorders; 2) current physical or mental stress and other severe health complications; and 3) pharmacological medication up to 2 weeks prior to the study. All participants had normal or corrected-to-normal vision. Further, all participants of the fMRI study were right-handed. Studies were approved by the local Ethics Committee of the University of Lübeck and conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent and were monetarily compensated for their participation.

3.1.3 Study 5-6

Study 5 and 6 were conducted within the same pharmacological intervention. From the initial sample of 30 participants, 3 participants in Study 5 and 2 participants in Study 6 had to be excluded due to technical problems in data collection, leaving a final sample of 27 healthy, right-handed, female participants in Study 5 (22±2.26, mean age±SD) and 28 healthy, right-handed, female participants in Study 6 (22.11±2.25, mean age±SD). Sample size was based on a previous study using similar procedures (Riba et al., 2008). The same exclusion criteria of Study 1-4 were used. Given sex differences in receptor availability (Pohjalainen, Rinne, Nagren, Syvalahti, & Hietala, 1998), modulation by pharmacological intervention (Munro et al., 2006; Soutschek et al., 2017), dopamine function (Castner, Xiao, & Becker, 1993) and resting-state functional connectivity (RSFC) organization (Weis, Hodgetts, & Hausmann, 2017), only female participants were recruited for this pharmacological intervention study. Data on the use of hormonal contraception were collected, as the estrous cycle has been shown to affect the dopaminergic system (Becker, Perry, & Westenbroek, 2012; Jacobs & D'Esposito, 2011) and hormonal contraceptive use can affect social behaviors as well (Alvergne & Lummaa, 2010; Birnbaum, Zholtack, Mizrahi, & Ein-Dor, 2019).

3.1.4 Tasks: The Take Advice Game

In Study 1-4, different versions of the same task were employed with very similar procedures. Participants were invited to the lab and were made to believe that they were going to play two games with other participants who were in different rooms. The games were the take advice game (TAG; **Fig. 2A-B**) and the TG. In Study 1, 2 and 4, participants performed the tasks on the lab computer. In Study 3, participants played the TAG in the MRI scanner and the TG afterwards.

In the TAG, participants were required to choose the higher of two cards to win money. Importantly, they had no information about the numbers on the cards and needed to completely rely on the advisers who could see one of the cards and pass this information to the participants. Each trial consisted of four phases. Participants were made to believe that they randomly received the role of the advisee and were first matched with an adviser (adviser phase). The adviser gave an advice (advice phase). The advice could be any number between 1 and 9 except for 5. Finally, participants chose a card (decision phase) and received feedback (feedback phase), i.e., the actual numbers on the cards and a green/red circle to signal winnings/losses.

To disentangle honesty from reward information, accurate advice was unpredictive of the winning card. Moreover, as advisers could see only one of the two cards and thus did not know which of the them was the winning card, their advice was not directly related to the participants' winnings/losses. Finally, the accuracy of the advisers' advice was manipulated to induce different honesty impressions in the participants. In Study 1, there were three types of advisers: 1) consistently honest advisers who always gave accurate advice; 2) consistently dishonest advisers who always gave inaccurate advice; and 3) inconsistently honest advisers who were equally probable to give accurate and inaccurate advice. In Study 2, inconsistently honest advisers were replaced by no-reputation advisers. Participants received no feedback information about these advisers and hence could not form beliefs about their honest reputation. In Study 3, there were only two types of advisers, namely, consistently honest and consistently dishonest advisers. Finally, in Study 4, participants played with two advisers (an honest and a dishonest one) whose honesty was probabilistically determined (75-25%) and changed over the course of the game.

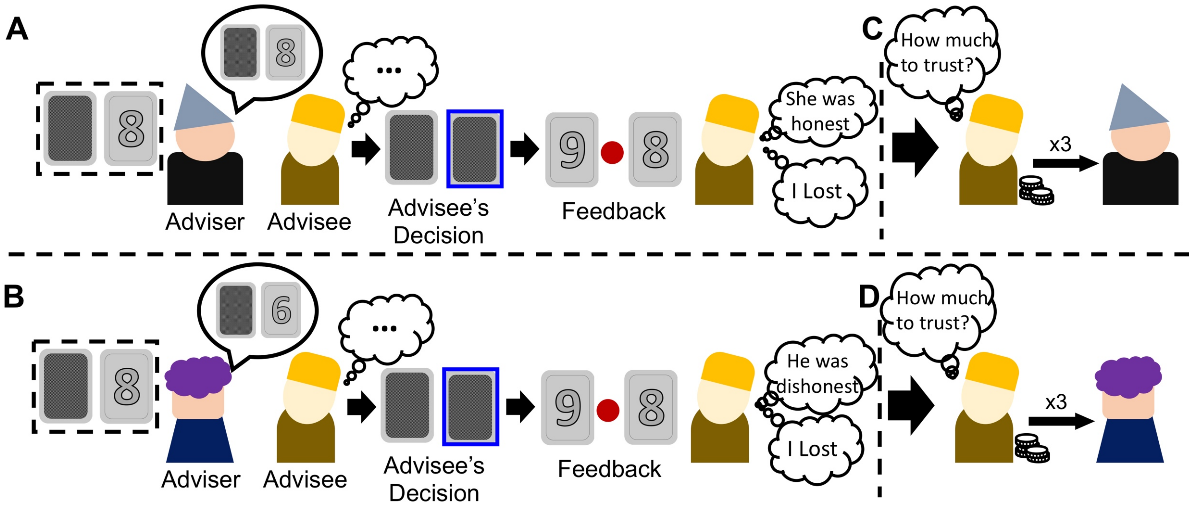


Fig. 2. Paradigms. Schematic representation of the Take Advice Game (TAG). Advisers were given information about one of the two cards and could communicate this information to the advisee. Participants, in the role of advisee, made a decision based on the information received (decision phase). In the feedback phase, advisees received two types of information: 1) social information, i.e., the actual numbers on the cards, which informed them about whether the adviser had been honest (**A.**) or dishonest (**B.**); and 2) non-social information, i.e., a green or red circle, which informed them whether they won or lost, respectively. After the TAG, participants in the role of investor played a one-shot trust game (TG) with both honest (**C.**) and dishonest (**D.**) advisers now in the role of trustee. Investors received a monetary endowment and decided whether they wanted to entrust some of this amount with the trustees. Investors were told that the shared amount was tripled by the experimenter and passed on to the trustee, who could decide to share back any portion of the tripled amount.

3.1.5 Tasks: The Trust Game

In Study 1-4, after the TAG and in Study 6 after the MRI session, participants played a one-shot version of the TG as investor (**Fig. 2C-D**). In the TG, participants shared an initial endowment (trust), namely, 10 monetary units (MUs) with their trustees (in Study 1-4, their previous advisers now in the role of trustee). Any shared amount was tripled and sent to the trustee who could decide to share back (reciprocity) any portion of the received amount of money. Payoffs in MUs were converted in Euros at the end of the experiment. In Study 6, participants were presented with pictures of trustees whose facial trustworthiness was manipulated.

3.1.6 Ratings and open questions

In Study 1-4 and 6, participants rated the trustworthiness of the trustees on a 7-point Likert-scale. Moreover, attractiveness ratings were also provided in Study 2, 4, 5 and 6. Attractiveness and trustworthiness ratings were randomized across participants. At the end of the experiments of Study 1-4, participants were asked to report (with a binary response option: Yes/No) whether they had used a strategy in the TAG, whether they thought that their strategy was successful and in Study 3, which criteria they used for their decisions in the TAG.

3.1.7 Tasks: Facial Evaluations

In Study 5, participants were invited to the lab at two different time points and received either 0.5 mg pramipexole or a placebo. Two hours after drug administration, participants underwent an 8-minute resting-state scan. After the MRI session, they performed a facial evaluation task. Stimulus material was based on Aharon et al. (2001) and consisted of two sets of 40 facial stimuli, i.e., “attractive” (20 pictures, 10 female) and “unattractive” (20 pictures, 10 female) faces. Participants made either a trustworthiness or an attractiveness evaluation of the face on a 7-point Likert-scale.

3.2 Data analyses

3.2.1 Behavioral data analyses

In Study 1-4, differences in advice-taking behaviors in the TAG, investment behavior in the TG and trustworthiness ratings were assessed by computing a one-way analysis of variance (ANOVA) and post-hoc paired t-tests. Mixed-effects logistic regression analyses were employed to investigate variables explaining trial-by-trial advice-taking behavior in the TAG. To compare trust in advisers with different degrees of honesty between Study 1 and Study 2, a two-sample t-test was used. Correlation analyses were computed for relationships between ratings and behaviors in the TAG and TG. In Study 3, mixed-effects regressions were implemented to test time effects on trust in the advisers.

In Study 5, effects of drug administration were tested with a paired sample t-test. In Study 6, mixed-effects regression models were employed to test the effects of drug administration on trustworthiness impressions and trusting behaviors. The best model was selected through a model comparison procedure using the Akaike Information Criterion (AIC). In each mixed-effects regression model, random-effects structure was kept maximal (Barr, Levy, Scheepers, & Tily, 2013). In Study 6, linear regression analyses were performed to investigate the contribution of trustworthiness and attractiveness impressions to trusting behavior in the TG.

3.2.2 Computational analyses in Study 4

To mathematically formalize individual learning of others' honest reputation in Study 4, computational models were fitted to participant's behavior. The winning model was the following:

$$V_{(t)} = V_{t-1} + \tau(I_t - V_{t-1})I_t + \delta(I_t - V_{t-1})(1 - I_t) \tag{3}$$

$$\tau = \begin{cases} \tau_{honest} & \text{if advice from honest adviser} \\ \tau_{dishonest} & \text{if advice from dishonest adviser} \end{cases}$$

$$\delta = \begin{cases} \delta_{honest} & \text{if advice from honest adviser} \\ \delta_{dishonest} & \text{if advice from dishonest adviser} \end{cases}$$

$$I_t = \begin{cases} 1 & \text{if accurate information} \\ 0 & \text{if inaccurate information} \end{cases},$$

where V_t is the subjective value of trusting the adviser on trial t, I_t is the type of social information (accurate or inaccurate advice) received on trial t, τ is the honesty learning parameter and δ is the dishonesty learning parameter. Trial-by-trial subjective values were transformed into trust probabilities with a stochastic decision rule (i.e., softmax function):

$$p_{trust} = \frac{1}{1+e^{-\beta(V_{trust}-V_{distrust})}}, \tag{4}$$

where p_{trust} is the probability of choosing to trust, β is the participant-specific inverse temperature (a free parameter indicative of the stochasticity of participants' choices), and V_{trust} and $V_{distrust}$ represent the value of choosing to trust (i.e., take the advice) and to distrust (i.e., discount the advice), respectively.

3.2.3 Neuroimaging data analyses in Study 3 and 5

Data collection and preprocessing. In Study 3, data were collected with a Siemens MAGNETOM TRIO 3 Tesla scanner at the Freie Universität Berlin. For each participant, an average of 360 contiguous volumes per run were collected with a T2*-weighted echo-planar imaging (EPI) sequence. A total of 5 runs of functional data were collected. High-resolution structural images were acquired through a 3D sagittal T1-weighted magnetization-prepared rapid acquisition with gradient-echo (MP-RAGE) sequence. In Study 5, imaging data were acquired with a 3-Tesla Siemens MAGNETOM Skyra whole-body MRI-scanner at the Center of Brain, Behavior and Metabolism in Lübeck. Each resting-state scan was approximately 8-minute long and consisted of 240 contiguous volumes. High-resolution structural images were acquired with a 3D sagittal T1-weighted MP-RAGE sequence.

Neuroimaging data analyses were performed on SPM12 v6685 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Preprocessing steps for functional images were as follows: 1) slice-timing correction; 2) unwarp for voxel displacement correction based on field maps; 3) realignment for head movement correction to the mean image; 4) co-registration to the structural image using the unified segmentation procedure (Ashburner & Friston, 2005) and normalization into MNI space using deformation fields from the segmentation procedure and a resampling voxel size of 2×2×2 mm^3. Multivariate analyses were based on these normalized functional images. For univariate analyses, functional images were also spatially smoothed using a Gaussian filter (8×8×8 mm^3 full width at half maximum, FWHM) to decrease spatial noise. Movement outliers were identified and excluded if head movements/translations were above 3 mm/rad.

Neuroimaging Analyses. Univariate and multivariate analyses were employed to analyzed fMRI data of Study 3 and 5. General linear models (GLMs) were defined for both univariate and multivariate analyses of fMRI data to estimate voxel-wise beta parameters that capture neural signals related to each effect of interest. Motion parameters were further included as regressors of no-interest in all GLMs. A temporal high-pass filter with a cutoff of 128 seconds was applied for all GLMs. In Study 5, further regressors were introduced to control for white matter and cerebrospinal fluid signal, and a band-pass filter (0.01~0.1 Hz) to remove high-frequency noise and linear drift artifacts. Results were whole-brain corrected for multiple comparisons using a voxel-level threshold of $p < .001$ and a cluster-level, family-wise error (FWE$_c$) corrected threshold of $p < .05$ (Eklund, Nichols, & Knutsson, 2016). Functional connectivity results in Study 5 were corrected on the ROI-level using a false discovery rate (FDR) of FDR < .05. In particular, RSFC was estimated for every participant and each session (dopamine/placebo) running Pearson's bivariate correlations between the average blood-oxygen-level-dependent (BOLD) signals of 142 ROIs as defined by Dosenbach et al. (2010) on the Functional Connectivity toolbox v15 (https://www.nitrc.org/projects/conn). Resting-state functional networks were defined based on the functional atlas of Dosenbach and colleagues (Dosenbach et al., 2007; Dosenbach et al., 2010; Dosenbach et al., 2006).

Task-dependent functional connectivity was implemented in Study 3 using a whole-brain psychophysiological interaction analysis (PPI, Friston et al., 1997). The PPI-GLM consisted of a task regressor, a physiological regressor entailing deconvolved BOLD signal from the seed region and a regressor for the interaction term with movement parameters as regressors of no interest.

In Study 3, decoding analyses were performed using linear support vector machine (SVM) and a whole-brain searchlight (radius = 10 mm). Applying a leave-one-run-out cross-validation (LOROCV), the SVM was trained on all but one run and tested on the left-out run. For searchlight decoding, only voxels within the whole-brain gray matter probability mask provided by SPM were used (white matter probability threshold = 0.1).

Across-subject classification analyses in Study 3 were performed using a leave-one-subject-out cross-validation (LOSOCV) approach in which the SVM was trained on average beta images of all but one participant and tested on the left-out participant. Performance of the across-subject classification accuracy was computed running a permutation test with 10,000 permutations (*n_perm*) and the sum of the models trained on permuted labels that performed better than the true model was computed (*p_models*). The nonparametric *p* value was assessed based on the following formula (Phipson & Smyth, 2010):

$$\frac{1 + p\_models}{1 + n\_perm}\,. \tag{5}$$

A similar LOSOCV procedure was employed for prediction analyses with multivariate regression models in Study 3 and Study 5. However, performance of the multivariate regression models was determined by computing the standardized mean squared error (smse).

Finally, decoding results in Study 3 were functionally characterized by running a meta-analytic image decoding analysis with the help of the Neurosynth Image Decoder (neurosynth.org; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011).

Chapter 4

4.1 Honesty as antecedent of trust

In many circumstances in life, individuals seek advice before making a decision. As bad decisions might jeopardize an individual's survival chances, gathering sufficient information from others helps make more informative decisions. It is thus pivotal to seek advice when making a decision, but it is also central to know whom advice should be sought from. However, little is known about how an adviser's reputation impacts an individual's willingness to trust advice. In particular, the honesty of the adviser might be central to one's trustworthiness perceptions and thus function as antecedent of trust.

As outlined in the introduction, a potential information-reward confound in the current literature employing advice-taking paradigms, that is, the fact that informative advice has in general been operationalized as the best option leading to higher or the highest reward outcomes, might have evoked cognitive processes closely associated with reward processing but unrelated to trust. Further, advisers have in general gained benefits for their advice and had incentives to send accurate advice in previous paradigms (Bonaccio & Dalal, 2006). This made participants in the role of the advisee focus more on the accuracy and congruency of the advisers' advice to track the adviser's motives rather than on learning the advisers' character traits (Behrens et al., 2008; A. O. Diaconescu et al., 2014). However, in real-life situations, informative advice is not always advice about the best action or decision to make (Dalal & Bonaccio, 2010). For instance, in many circumstances, the advice of *not* doing something might be more informative than an advice of carrying out a specific array of actions. Similarly, advisers not always receive proximal benefits for their advice. The TAG was developed to address these issues.

Study 1 shows that the TAG was successful in inducing honesty-based trustworthiness perceptions (**Fig. 3A**). Honest advice increased trustworthiness perceptions and trusting behaviors across contexts. Participants not only took more advice from honest advisers but entrusted them also with more money in a subsequent interaction (i.e., in the TG). Notably, inconsistency in honest behavior reduced willingness to take advice, but inconsistent honest others were still trusted more and perceived as more trustworthy than dishonest advisers (**Fig. 3C**). These results suggest that even small signs of honest behavior induce others to reciprocate. Finally, participants preferred honest advice even though it was not more informative to make better decisions. This finding suggests that honest advice is associated with an information bonus that is integrated into the decision-making process by uninformed decision-makers. It is still unclear, however, whether individuals would take the advice of an honest other even if information about the current honesty of the other is not available, or whether they would be more likely to take advice from those with an established reputation than from those whose reputation is unknown.

Chapter 5

5.1 Honest reputation biases trustworthiness perceptions

We often reach out to others for advice for the most disparate reasons. However, we rarely (if ever) have any control over the quality of the other's advice. Employing a new version of the TAG, Study 2 investigated how individuals integrate information from others when feedback about the accuracy of the other's advice is missing. Moreover, it was also inquired whether in such contexts, individuals would prefer to take advice from advisers with an established good reputation as opposed to advisers without any reputation.

Findings from Study 2 suggest that when it is impossible to check the accuracy of the other's advice, individuals decide whether to take the advice exclusively on the basis of the adviser's reputation (**Fig. 3B**). These results extend previous work by demonstrating that uninformed decision-makers base their advice-taking and advice-discounting strategies on the other's character or reputation (Yaniv & Kleinberger, 2000). Individuals were less likely to take advice from dishonest advisers than honest advisers or advisers without reputation but as likely to take advice from honest advisers as from advisers without reputation (**Fig. 3D**). Participants also preferred to take advice from advisers without an honest reputation than from advisers who showed to be inconsistent in their honesty (**Fig. 3E**). Moreover, participants also trusted advisers without reputation significantly more than dishonest advisers in a subsequent trusting interaction. These findings suggest that individuals had positive initial expectations of others and chose to take advice from them as a default trusting strategy. A decision to disregard an adviser's advice was made only after participants learnt the other's dishonest behavior, namely, when they realized that their trust had been misplaced. These behavioral patterns raise the question as to how individuals dynamically update and revise their beliefs about the other's honest character. I will address this question later on in Chapter 7.

Thus, in these first two behavioral studies, I showed how individuals form beliefs about others' reputation and how these beliefs inform behaviors across contexts. Results reveal that an honest reputation predicts trusting behavior across contexts and even in the absence of feedback about the other's behavior. In particular, consistently and inconsistently honest others were trusted more and perceived as more trustworthy than dishonest ones. However, those whose honest reputation was unknown were trusted more than inconsistently honest others, suggesting that signs of dishonesty negatively impact one's initial positive expectations of others' trustworthiness.

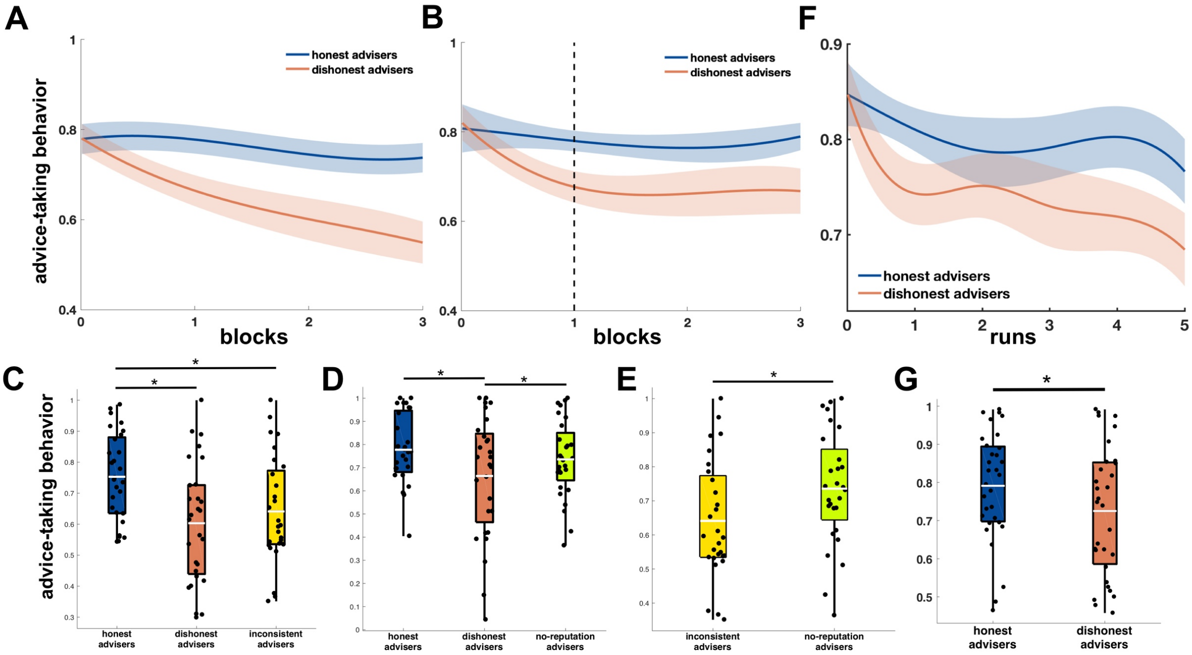


Fig. 3. Behavioral results of Study 1-3. Advice-taking behavior in the Take Advice Game toward honest and dishonest advisers over time (i.e., blocks) in Study 1 (**A.**) and Study 2 (**B.**). In Study 2, participants did not receive any feedback after the first block (dashed line). Average advice-taking behavior toward honest, dishonest, inconsistent and no-reputation advisers in Study 1 (**C.**) and 2 (**D.**) for each participant. Between-subject difference in advice-taking behavior toward inconsistently honest advisers in Study 1 and advisers without reputation in Study 2 (**E.**). Replication of the behavioral results from Study 1 in the MRI scanner (Study 3): advice-taking behavior over functional MRI runs (**F.**) and average advice-taking behavior for each participant (**G.**). Dots in the boxplots represent participants. * *p* < .05.

Honest and dishonest others are thus quickly identified, and a reputational tag is attached to them that might ease the decision-making processes in future encounters. In difficult situations, in which we are required to make a decision, relying on others might reduce the burden of the decision-making process. Thus, trust might buffer the stress related to a difficult decision in our everyday life (Kikusui, Winslow, & Mori, 2006; Rapoza et al., 2016; Thorsteinsson, James, & Gregg, 1998; Yanagisawa et al., 2011). Conversely, interacting with dishonest others might be more stressful, as dishonest others do not appear to have an intrinsic motivation to commit to the other's well-being. As such, they might exploit us or simply let us down at any time by providing false or poor information. A decision-maker who has observed signs of dishonesty might prefer to refrain from interacting with those unreliable others. However, when this cannot be avoided, a decision-maker might be in a state of high alertness and might constantly track the other's actions to anticipate disadvantageous outcomes. This leads to specific hypotheses on the neural correlates of beliefs about honest character traits and on how these beliefs are dynamically updated and revised—research questions that will be addressed in the next two chapters.

Chapter 6

6.1 Neural representations of honesty predict future trust

Neuroimaging studies investigating trust have largely employed the TG. However, as mentioned in the introduction, the structure of the game induces individuals to trust as long as they will be better off with trusting than distrusting (E. Fehr, 2009), thereby focusing on maximizing their personal payoffs (Ernst Fehr & Fischbacher, 2002). Thus, when there are no external incentives or when trust is associated with monetary losses (Johnson & Mislin, 2011; Rode, 2010), individuals cease to trust (Chang et al., 2010; Hula, Vilares, Dayan, & Montague, 2017). This raises the question as to whether neural activity in striatal and orbitofrontal regions observed during a trust decision in the TG underlies the act of trust as such or rather represents reinforcement learning mechanisms signaling reward outcomes (Bellucci et al., 2017; Bellucci et al., 2018; Chang et al., 2011; Delgado et al., 2005).

Indeed, other neuroimaging studies investigating social behaviors have observed a different set of brain regions when individuals interact with others, understand their intentions and learn their character, such as the DLPFC, the pTPJ, and the VMPFC, respectively (Buckholtz et al., 2008; Cooper et al., 2010; FeldmanHall et al., 2018; Igelstrom & Graziano, 2017; Igelström, Webb, & Graziano, 2015; Koster-Hale et al., 2017; Saxe & Powell, 2006; Tusche et al., 2016; L. Young & Saxe, 2008). Thus, it is plausible to hypothesize that similar brain regions are engaged when individuals evaluate each other's character to decide whether to trust.

In Study 3, using multivariate voxel pattern analysis (MVPA) in combination with fMRI, the relationships between honesty, dishonesty and trust on both behavioral and neural level were analyzed. On the behavioral level, the findings observed in the previous behavioral studies (Chapter 4 and 5) were replicated (**Fig. 3F-G**). Honest behavior increases trust irrespective of proximal benefits associated with the act of trust. Further, the honest character of an advisor makes others more likely to accept the adviser's advice and more willing to trust the adviser in a later interaction (i.e., the TG).

On the neural level, the other's honest character was decoded in brain regions associated with higher-order cognition, such as the DLPFC, posterior cingulate cortex (PCC) and intraparietal sulcus (IPS), whereas striatum and anterior cingulate cortex (ACC) significantly decoded individual feedback information about one's winnings and losses (**Fig. 4A**). Interestingly, honesty-decoding neural patterns in the DLPFC, PCC and IPS (but not reward-decoding neural patterns in the striatum and ACC) predicted individual trust in the TG. In particular, honesty more strongly recruited the VMPFC than dishonesty (**Fig. 4B**). The VMPFC was in addition functionally coupled with the pTPJ during honesty as opposed to dishonesty (**Fig. 4C**). Further, stronger VMPFC-pTPJ coupling correlated with higher trust in the TG (**Fig. 4D**), suggesting that a stronger integration of the honesty signal increases an individual's willingness to trust the other later on.

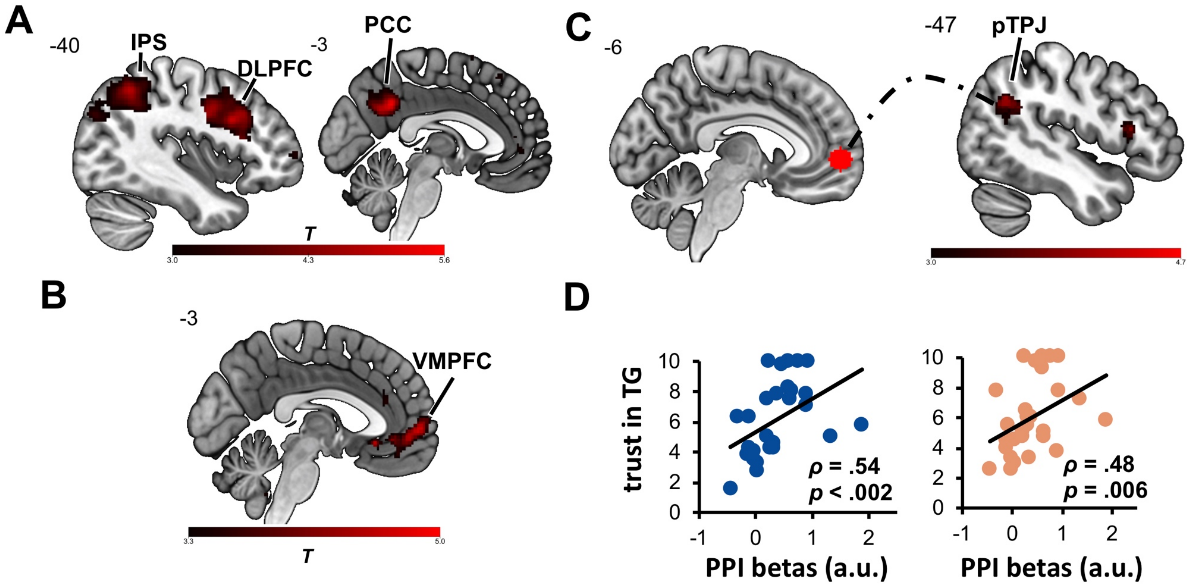


Fig. 4. Neuroimaging results of Study 3. Multivariate analyses revealed that honesty-based trustworthiness was decoded in the PCC, bilateral DLPFC and left IPS (**A.**). Honesty, as opposed to dishonesty, engaged the VMPFC (**B.**). A PPI analysis revealed that the VMPFC was more strongly functionally coupled to the pTPJ for

honest than dishonest advisers (**C.**). VMPFC-pTPJ functional connectivity correlated with subsequent trust decisions for honest and dishonest advisers (**D.**). IPS, intraparietal sulcus; DLPFC, dorsolateral prefrontal cortex; PCC, posterior cingulate cortex; VMPFC, ventromedial prefrontal cortex; pTPJ, posterior temporoparietal junction; TG, trust game; PPI, psychophysiological interaction; a.u. arbitrary units; *T*, *t*-values; results of multivariate analyses are cluster-level family-wise error corrected (cFWE) for multiple comparisons at cFWE < .05 with an uncorrected, cluster-forming threshold of $p < .001$; results of functional connectivity analyses are small-volume, cluster-level family-wise error corrected within the pTPJ (FWE_{svc}) at $\text{FWE}_{\text{svc}} < .05$ with an uncorrected, cluster-forming threshold of $p < .001$.

Finally, an asymmetry in the OFC activity in response to positive feedback due to the honest reputation of the adviser was also observed. Such asymmetry in feedback encoding likely jeopardizes an individual’s ability to optimally update one’s beliefs about the other, promoting judgmental biases. It still remains unclear, however, whether the other’s honest character impairs learning processes underlying character trait learning. For instance, how do individuals learn the honest character of the other to be able to optimally revise their behavior and avoid being exploited by the other? Better insights into these learning processes might clarify whether and how honest reputation impairs social learning—an empirical question that will be addressed in the next chapter.

Chapter 7

7.1 Honest reputation impairs learning

In the previous experiments, I have shown how honesty induces trustworthiness perceptions that inform trust decisions across contexts. Individuals initially trusted unknown others, and this level of trust was similar to their level of trust in advisers with a good reputation. On the contrary, individuals adapted their trusting behavior as they slowly gathered evidence that their trust was misplaced. A set of frontoparietal and mentalizing brain regions was engaged, which likely allowed for this behavioral adaptation. These results indicate that in social interactions with unknown others, people tend to trust first if they do not have evidence to behave otherwise or as long as they do not learn that trust might make them vulnerable to others' exploitation. However, evidence on how information about the other's trustworthiness character is integrated to inform and, eventually, revise these trusting behaviors is still missing.

In this chapter, a new version of the TAG is described that allowed to apply reinforcement-learning models to mathematically formalize how social information is processed and integrated to form and update beliefs about the other's honest reputation. Results from this experiment replicated and extended the behavioral patterns observed in previous chapters. First, participants were seen to initially trust both advisers. However, after a couple of trials, when they realized that their trust in the dishonest adviser was misplaced, they immediately adapted their behavior, discounting advice from the dishonest partner. However, and most interestingly, the results further showed that the same did not happen for those who could establish a reputation as an honest partner over the course of the first period of the interaction (**Fig. 5A**). In other words, participants kept trusting the advice of those advisers who initially showed to be honest, disregarding information inconsistent with their honest reputation. This suggests that individuals integrate information from those with an honest reputation differently from those with a dishonest reputation over the course of the social interaction. That is, once the initial positive expectations that the adviser is trustworthy were confirmed, individuals placed more weight on new incoming information consistent with the honest reputation of the adviser. On the contrary, when the initial positive expectations that the adviser behaves in a trustworthy manner were not confirmed, individuals tended to value new consistent and inconsistent information equally.

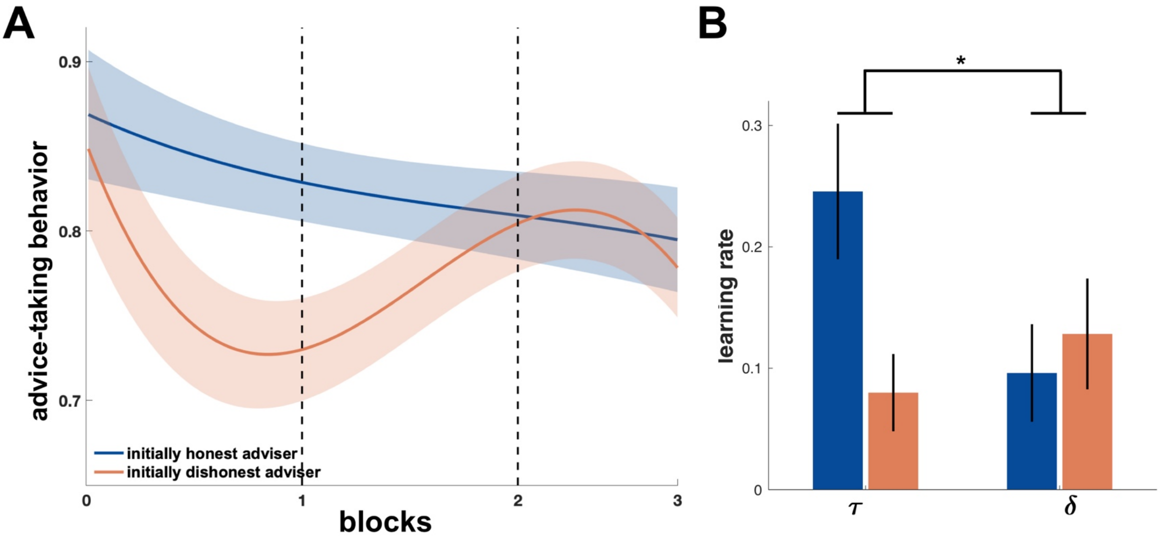


Fig. 5. Behavioral and learning results of Study 4. Advice-taking behavior toward initially honest and dishonest advisers over time (i.e., blocks) in Study 4 (**A.**). Dashed lines separate blocks. Advisers changed their honesty in advice giving across blocks. Participants closely tracked changes in honesty of the initially dishonest advisers, but not of the initially honest advisers. Such difference in behavioral adaptation points to a learning impairment related to a failure in successful belief update. Results from computational modeling reveal that participants weighted accurate information from the initially honest adviser significantly more than inaccurate information or accurate information from the initially dishonest adviser (**B.**). τ, learning rate for accurate information; δ, learning rate for inaccurate information. * $p < .05$.

Notably, the asymmetry in information weighting for honest advisers impaired beliefs updating and behavior change once honest advisers turned dishonest. On the contrary, individuals were still able to change their behavior toward initially dishonest advisers after learning that they become increasingly honest. This likely hinges on a significant difference in the weights given to the information coming from the advisers (**Fig. 5B**). That is, accurate information from initially honest advisers was valued significantly more than the same information provided by an initially dishonest adviser. Importantly, initial beliefs formed from direct reputation (being honest or dishonest) were so enduring that participants trusted the advisers significantly differently in a subsequent interaction (i.e., in the TG), although they behaved on average equally honestly.

Chapter 8

8.1 Neural modulation of resting-state brain dynamics

Results from the previous chapters, provide novel evidence that representations of character traits informative of social behaviors are encoded in a distributed neural network associated with higher-order cognition and are separable from a reward encoding network. However, these studies still leave unanswered the question as to what role the dopaminergic system plays in trusting behaviors.

In Study 5 and 6, a pharmacological intervention was conducted to investigate the relationships between dopamine and trust. In the current chapter, the effects of the pharmacological manipulation on neural dynamics is described. A D2/D3 dopamine agonist (i.e., pramipexole) was administrated to participants in a double-blind, placebo-controlled, within-subject design. The drug was administrated before participants underwent a resting-state MRI scan. Resting-state brain dynamics reflected the neural responsiveness to acute drug administration (**Fig. 6A**). In particular, subcortical brain regions within the cinguloopercular network were mostly affected by the dopaminergic manipulation. The most affected resting-state brain dynamics were in well-known dopaminergic brain areas such as the striatum and medial PFC previously shown to be modulated by pramipexole (Gurevich & Joyce, 1999; Hall et al., 1996; Ishibashi et al., 2011; A. M. Murray, Ryoo, Gurevich, & Joyce, 1994; Riba et al., 2008). Notably, these neural signatures of pramipexole's administration significantly predicted the drug's effects on subsequent facial attractiveness evaluations (**Fig. 6B-C**). In particular, stronger functional connectivity within the cinguloopercular network predicted increased facial attractiveness evaluations following pramipexole intake.

Chapter 9

9.1 Dopaminergic effects on behavioral trust

After having shown that administration of a dopamine agonist successfully modulates neural dynamics, Study 6 addressed the question as to whether this dopaminergic modulation impacted trusting behavior. However, as pointed out in the introduction, eliciting trusting behaviors by inducing trustworthiness perceptions over the course of repeated interactions is problematic because learning mechanisms hinge on dopaminergic functioning, which might as well be impacted by a pharmacological intervention of the dopaminergic system. This, in turn, might introduce a serious confound that makes disentangling the effects of dopamine on trust from its effects on learning processes preceding trusting behavior impossible.

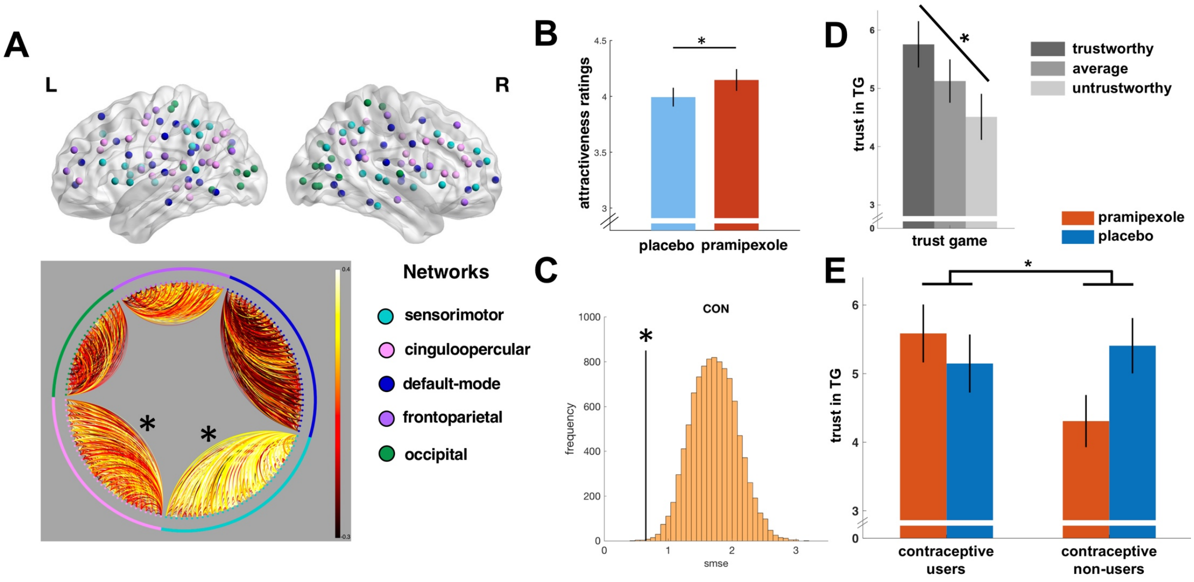


Fig. 6. Neuroimaging and behavioral results of Study 5 and 6. In Study 5, resting-state network-level analyses revealed that only the SMN and the CON were significantly modulated by pramipexole intake (**A.**). Pramipexole intake significantly impacted participants' perceptions of facial attractiveness (**B.**). The behavioral effects of pramipexole on attractiveness perceptions could be predicted by pramipexole's modulation of resting-state CON dynamics (**C.**). In Study 6, participants' trusting behaviors were modulated by the facial trustworthiness but not the facial attractiveness of the trustee (**D.**). Pramipexole affected participants' trust by

interacting with hormonal contraceptive use (**E.**). SMN, somatosensory motor network; CON, cinguloopercular network; smse, standardized mean squared error; TG, trust game. * $p < .05$.

One possible solution is to induce trustworthiness perceptions of others without having participants learn of the others' trustworthiness. Hence, in this Study 6, participants' trust during the one-shot TG was manipulated by presentation of faces that varied in their facial trustworthiness. Further, to minimize facial attractiveness confounds, faces maximally differed on the trustworthiness dimension with minimal variations on the attractiveness dimension. The facial trustworthiness manipulation was successful, as participants trusted the trustworthy-looking partners significantly more (**Fig. 6D**). Moreover, trusting behavior could significantly be explained by the trustee's facial trustworthiness independently of facial attractiveness information. Furthermore, administration of pramipexole decreased trust in others. This drug effect was further mediated by hormonal contraceptive use, as women who did not use hormonal contraceptives trusted less after pramipexole intake, whereas pramipexole increased trust among contraceptive users (**Fig. 6E**).

Chapter 10: General Discussion

In this dissertation, I pursued two lines of research that aimed at providing insights into the psychological antecedents and neural determinants of trusting behaviors. In the first part, I showed how individuals form beliefs about others' reputation and how these beliefs inform trusting behaviors across contexts. Results showed that an honest reputation predicts trusting behavior across different social interactions even in the absence of current feedback about the other's behavior. Honesty-based trustworthiness was encoded in an extended network involving the lateral PFC, IPS and PCC that predicted future trust. Stronger integration of an honesty signal from the VMPFC correlated with higher trust in a later interaction with the other. Notably, an honest reputation modulated how feedback information was encoded in the OFC. Further, participants had initial positive expectations of others' trustworthiness, employing trust as a default behavioral strategy in interactions with new partners. Predictably however, signs of dishonesty negatively impacted trust. On the contrary, confirmation of these initial positive expectations produced strong trustworthiness perceptions about others that impaired social learning and hindered adaptive, flexible behavior. This learning impairment was due to an asymmetry in the weighting of information and its integration when interacting with honest others. This difference in information weights might in turn depend on the honesty-induced differences in feedback processing in the OFC observed in the fMRI study.

In the second part, I provided novel pharmacological evidence of the role of dopamine in trust. This included delineating the effects of pramipexole (a D2/D3 dopamine agonist) on neural dynamics at rest. These analyses revealed that pramipexole administration successfully modulated different metrics of resting-state brain dynamics, in particular, in brain regions known to be targeted by pramipexole's dopamine agonist effects. Once the modulation of the dopaminergic system was assured, the effects of pramipexole on behavioral trust in a one-shot TG were tested outside the MRI. Thereby, it was shown that the contribution of facial trustworthiness to subjective impressions that guide subsequent trust decisions could be disentangled from other types of social information from faces (e.g., attractiveness). Although pramipexole did not impact subjective, trustworthiness impressions, trusting behavior was significantly modulated by the dopamine agonist. Notably, the effects of pramipexole on behavioral trust interacted with hormonal contraceptive use in the female sample. Increased trust was observed in women using hormonal contraceptives and decreased trust in naturally cycling women after pramipexole intake.

10.1 Honesty as antecedent of trust

To choose the proper course of actions in a dynamically changing world, decision-makers need to gather information about the structure of the world (R. C. Wilson, Geana, White, Ludvig, & Cohen, 2014). Even though a decision's outcomes remain to a certain degree always uncertain, gathering more information allows one to make more reliable inferences about the future state of the world. More reliable inferences imply a more accurate model of the world, reduced surprise about future events and better survival chances (Badcock, Friston, & Ramstead, 2019). Human beings prefer to make decisions whose outcomes are known or can be known probabilistically (i.e., under risk), whereas they show a strong aversion to ambiguous situations in which no inference on future outcomes can be made (Platt & Huettel, 2008).

In a social interaction, the world whose hidden states need to be inferred is another human being. In this context, a human agent tries to first gather information about the other that can be based on a previous experience with that person, i.e., direct reputation, or on indirect information about the other, i.e., indirect reputation (Izuma, 2012; Li, Meng, & Ma, 2017). With the help of this information, the decision-maker can make inferences on the other's character to deduce the other's behavior in different contexts. For instance, I may trust you if you have previously been trustworthy to me or if I have heard that you are a trustworthy person.

Previous research has suggested that individuals trust others because they have positive expectations about the good intentions and behaviors of others (Rousseau et al., 1998). As outlined in the introduction, models of trust have proposed qualities of the other that promote trustworthiness impressions, which ultimately guide trusting behaviors. However, empirical evidence of honesty as an antecedent of trust was still lacking. Only recently, some studies have pointed to honesty as a fundamental character trait that correlates with a variety of prosocial behaviors, such as altruistic behavior, unconditional kindness and reciprocity (Ashraf et al., 2006; Baumert et al., 2014; Hilbig et al., 2015; Thielmann & Hilbig, 2015).

Study 1 provides evidence of honesty as antecedent of trust. Results showed that individuals prefer taking advice when there are no reasons to believe the other to be untrustworthy or dishonest. These findings concur with evidence showing that individuals prefer options that are either preferred or suggested by others (Biele et al., 2009; Mahmoodi et al., 2018). When, however, participants learnt over the course of the interaction that some partners are honest and others are not, participants revised their behavior by discounting the advice of the dishonest partners. As participants could not know at the time of the decision whether the advice was accurate or not, participants based their decisions on the reputation of the other. Honest others built their reputation on the accurate information they shared and their honest reputation might have worked as a proxy for the quality of the information shared in a future encounter (Gordon & Spears, 2012). Participants' preference for advice from honest others may thus reflect their attempt to improve their decisions and reduce uncertainty by using information that is likely accurate. This finding accords with previous results showing that individuals more strongly rely on others' advice in highly uncertain situations like the TAG (McElreath et al., 2005; Sniezek & Buckley, 1995; Sniezek, May, & Sawyer, 1990; Van Swol & Sniezek, 2005).

Even though taking accurate advice from honest others did not yield higher gains, participants showed a consistent preference for truthful advice. Previous paradigms have not properly controlled for the reward-information confound, as advice usually took the form of the best option in the task, generally associated with higher rewards (Behrens et al., 2008; Biele et al., 2009; A. O. Diaconescu et al., 2014; Rodriguez Buritica et al., 2019). Thus, patterns of advice-taking behaviors in previous studies might well be described by more parsimonious explanations, such as by classic mechanisms of reward learning. However, good, informative advice is rarely advice about the best and most rewarding decision to make. Examples are, for instance, disclosing information about certain decisions, sharing one's own experience after certain choices or advice of not to take certain actions (Bonaccio & Dalal, 2006; Dalal & Bonaccio, 2010). Results reported in this dissertation suggest that uninformed decision-makers prefer to take informative advice irrespective of their proximal benefits. Moreover, in a subsequent interaction, participants repaid their advisers for their honesty by entrusting more money to the honest advisers in the TG, whereas no relationships were found between the amount of gains derived by the honest advice in the TAG and money entrusted in the TG. These results confirm that in a social exchange, individuals are motivated by reputational concerns and decide whether to trust based on the social qualities of the other. Further, they validate the strength of the task in disentangling social learning processes from others, related but exogenous, learning processes (such as reward learning). This was of pivotal importance to Study 3 to capture the specific neural signatures of honest character learning (see below).

10.2 Initial expectations are disrupted by dishonesty

Even if a decision-maker lacks information about the interacting partner, individuals still have a way to make good-enough inferences about the other. In particular, individuals use their knowledge of social norms as priors to infer the other's likely behavior (Bellucci et al., 2018; Bicchieri, 2014). In the TAG, at the beginning of a social interaction, participants might have assumed that the other complies with social norms of fairness and equity. Despite the betrayal aversion associated with acts of trust (J. A. Aimone & Houser, 2011, 2013; I. Bohnet et al., 2008; Iris Bohnet & Zeckhauser, 2004), these expectations of a compliant behavior help individuals overcome concerns of betrayal and exploitation by the partner (Thomas Baumgartner, Fischbacher, Feierabend, Lutz, & Fehr, 2009; Masuda & Nakamura, 2012; van 't Wout & Sanfey, 2008). In Study 2, this was supported by the fact that participants trusted advisers with an unknown reputation as much as advisers with an honest reputation. The same is proven by participants' initial behavior toward dishonest advisers at the beginning of the social interactions in Study 1-4. In all experiments, participants revised their advice-taking behavior toward dishonest advisers only after they realized that their trust in those advisers was misplaced.

Thus, when a history of interactions with the partner is possible, individuals integrate information that informs their beliefs about the other's character. Thereby, they attach "reputation tags" to the other that can be thought of as priors allowing good-enough estimations of the other's (future) behavior (E. Fehr & Fischbacher, 2003). Based on the other's reputation, individuals decide on the best trusting strategy (Milinski, Semmann, & Krambeck, 2002; Semmann et al., 2004; Wedekind & Braithwaite, 2002). Therefore, people grant others a trustworthy character recognizing distrust as the optimal strategy in case of an untrustworthy reputation. Such reputation likely informs one's future behavior during interactions with the recognized untrustworthy other. Consistently with this notion, Study 1-4 showed that an honest reputation in advice giving has an impact not only on advice-taking behaviors in the current situation, but also generalizes to trust in a different context and situation.

Interestingly, signs of dishonesty negatively impacted one's trustworthiness expectations of others. This was evidenced by the fact that participants trusted inconsistently honest advisers less than advisers without any reputation. Nonetheless, even small signs of honesty increased participants' trust, which remained at higher levels for inconsistently honest advisers as opposed to dishonest advisers. These findings suggest that honesty plays a more central role in building and maintaining trust than simple behavioral predictability (e.g., being predictably dishonest) (G. R. Jones & George, 1998). However, signs of dishonesty have deleterious consequences on initial expectations of others' trustworthiness, as, once it is lost, trust might be difficult to regain and might never be entirely regained (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000).

10.3 Neural signatures of honesty predict future trust

Study 3 showed that the trustworthiness inferred from the other's honest or dishonest behavior was decoded in four brain regions (i.e., the PCC, IPS and bilateral DLPFC). These neural signatures of honesty-based trustworthiness was able to successfully classify neural responses to honesty and dishonesty in out-of-sample individuals. Importantly, brain signals from these regions was informative of future trust decisions in a subsequent interaction with the partners. Thus, these brain regions might play a central role in understanding others, learning their character and revise one's behavior to tailor it to the other's behavior.

This is consistent with previous work showing that these brain regions are associated with judgments about others' traits (PCC) (Schurz, Radua, Aichhorn, Richlan, & Perner, 2014), attribution of temporary beliefs to others (IPS) (Igelstrom & Graziano, 2017; Schurz et al., 2014) and a variety of prosocial behaviors such as generous decisions (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006) and group-based cooperation (DLPFC) (Lemmers-Jansen, Krabbendam, Veltman, & Fett, 2017; Wills, FeldmanHall, Collaboration, Meager, & Van Bavel, 2018). Importantly, the PCC, IPS and bilateral DLPFC have been observed to form an interconnected brain network during interpersonal interactions (Hackel et al., 2015; Igelström, Webb, Kelly, & Graziano, 2016; Mende-Siedlecki, Cai, & Todorov, 2013). Given that Study 3 showed that these regions are not only engaged during an online interaction but further contain neural signal informative of individual future trust, these regions likely build an intertwined brain network engaged in representations of behaviorally-relevant qualities of others, such as social character traits. These representations likely entail information retrieved to choose the optimal behavioral strategy during present and future social interactions.

Critically, these brain regions were also more strongly recruited by dishonesty than honesty. As participants revised their behavior toward dishonest partners over the course of the interaction but not toward honest others, these findings confirm the role of these regions in representing social character traits of others for optimal social strategy selection. This concords with previous evidence that these regions are engaged by others' non-cooperative behavior (Yang, Zheng, Yang, Li, & L., 2018) and violations of expectations in social contexts (e.g., decisions to lie) (Chang, Yarkoni, Khaw, & Sanfey, 2013; Cloutier, Gabrieli, O'Young, & Ambady, 2011; Greene & Paxton, 2009). Hence, recruitment of these brain regions by dishonesty might reflect the online tracking of the other's norm-deviant behavior and the updating of one's beliefs for flexible behavioral adjustments.

Honesty, on the other hand, recruited the VMPFC and brain signal in the VMPFC was functionally coupled with the pTPJ during honesty encoding. Notably, the strength of this functional connectivity correlated with higher trust in the partner during the future interaction in the TG. Given the role of the VMPFC in representations of positive traits of others (Hackel et al., 2015; R. J. Murray, Schaer, & Debbane, 2012; Welborn & Lieberman, 2015) and of the pTPJ in inferences on others' intentions (Saxe & Powell, 2006; L. Young & Saxe, 2008), these findings suggest that a stronger integration of the honesty signal from the VMPFC supports inferences on the other's good intentions undertaken by the pTPJ, resulting in more positive beliefs about the other. These positive beliefs, in turn, lead to an increased willingness to trust the other. Thus, the interplay between these two brain regions likely represents a neural mechanism underlying integration of character information for belief formation about the other's behavior.

Finally, honesty modulated neural responses to value information in the OFC during outcome evaluations. In particular, positive outcomes received when interacting with honest partners elicited significantly higher responses in the OFC. In line with its role in processing subjective values (Sescousse, Redoute, & Dreher, 2010; Valentin, Dickinson, & O'Doherty, 2007), higher neural activity in the OFC might reflect an enhanced subjective value of rewards induced by the honest character of the other. These findings indicate a possible mechanistic explanation to the positivity bias toward individuals with a good reputation that has been observed to influence learning processes (Corriveau & Harris, 2009; Sabbagh & Shafman, 2009). An honesty-dependent asymmetry in valuation of outcomes in the OFC might promote stronger susceptibility to reputational priors and less flexibility in revising one's beliefs about the other. This complies with previous work showing that decreased OFC activity is associated with stronger resistance to belief change during information encoding (Kaplan, Gimbel, & Harris, 2016). Such honesty-based asymmetry in information encoding might jeopardize an individual's ability to optimally form and update one's beliefs and so foster a broad array of judgmental biases. Study 1-3, however, still do not provide compelling evidence as to whether the other's honest character impairs learning processes underlying character trait learning. Study 4 was conducted to answer this open question.

10.4 Impact of honest reputation on social learning

Results from Study 4 replicated and extended the findings in previous experiments. First, participants were seen to initially trust all advisers. However, when they realized that their trust in the dishonest adviser was misplaced, they adapted their behavior, discounting advice from the dishonest partner. Notably, this behavioral adjustment was not observed after the adviser could establish a reputation as an honest partner over the course of the first block. So, participants were seen to keep trusting the honest partner even after the partner stopped being honest. This suggests that participants were integrating information from the two advisers differently over the course of the social interaction. At the beginning of the social interaction, when no reputational knowledge about the others was yet available, all information was integrated in a similar fashion. However, once the initial positive expectation that the partner would behave in a trustworthy fashion was confirmed, participants might have reduced the integration of new incoming information about the current reputation of the initially honest adviser. On the contrary, when the initial positive expectation that the partner would be trustworthy was violated, participants readily changed their behavior and more closely tracked the other's trial-by-trial decisions, allowing for optimal behavioral revision over the course of the social exchange.

These behavioral patterns, which appear to reveal a learning impairment for honest partners, are likely due to the honesty-dependent difference in information weighting. In particular, accurate and inaccurate information from the initially dishonest adviser were weighted in a similar fashion, which likely explains why participants could readily update their beliefs about the initially dishonest adviser in the second block of the task when the adviser became trustworthy. On the contrary, participants placed more weight on accurate than inaccurate information from the honest partner, which impaired the flexible revision of one's advice-taking behavior. This might be due to the fact that information consistent with one's initial positive expectations led to the formation of strong priors that are less likely to be subject to revision. Concomitantly, such strong beliefs might have made individuals more likely to discount evidence of behaviors inconsistent with the other's reputation, like when individuals rationalize inconsistent policy contents on the basis of party membership (Cohen, 2003).

These results might explain a wide array of perceptual and judgmental biases in different domains. For instance, recent work has indicated a perceptual bias that contributes to prejudicial judgments of young Black men, whereby young Black men are judged as bigger and more threatening than young White men (J. P. Wilson, Hugenberg, & Rule, 2017). As perceptual evidence is integrated following a reinforcement learning mechanism (Badcock et al., 2019), such perceptual bias might rely on asymmetric weights of perceptual information similar to the one observed in Study 4. Social phenomena like the "do-gooder derogation" or "self-licensing" might be tracked back to a similar information-weighting asymmetry as well (Merritt, Effron, & Monin, 2010; Minson & Monin, 2011; Monin, Sawyer, & Marquez, 2008). For example, morally dubious actions might be licensed by past “good” deeds through reputation-based reinforcement mechanisms based on which we learn that others are likely to discount evidence inconsistent with our good reputation (Merritt et al., 2010).

Overall, findings from these studies might reflect a learning strategy optimization in social interactions that explains the need of reputational priors. Identifying and being able to keep track of free-riders is of pivotal importance to the individual survival chances, as free-riders may jeopardize one’s existence through perilous exploitation. However, tracking the intentions and motives of every single action of our social fellows implies enormous and unsustainable energy costs that call for better strategies to track and control the behaviors of others. Reputation might offer the tool that solves this conundrum allowing for efficient resource distribution in social behavior control. In particular, using reputational tags to quickly identify who deserves our trust enables one to track only a limited number of interactions in which the risk of exploitation is more likely to occur, resulting in an efficient energy saving. Although efficient in most situations, this strategy also lurks the danger of biased estimations that negatively influence the integration of new inconsistent information. This, in turn, might result in suboptimal decisions, for instance, as shown in Study 4, when individuals trust no-longer trustworthy others. Future studies are needed to shed light on the neural mechanisms underlying this social learning impairment.

10.5 Effects of a dopamine agonist on brain and behavior

Study 5-6 implemented a pharmacological intervention to investigate the second line of this dissertation’s research, namely, the relationships between dopamine and trust.

Study 5 first examined whether the pharmacological intervention was successful in modulating neural dynamics within well-known dopaminergic brain structures. Analyses of resting-state dynamics after drug administration show that brain regions in the striatum and medial PFC were modulated by pramipexole, concurring with previous evidence on D2/D3 dopamine receptor availability in the human brain and pramipexole's modulation of neural dynamics (Gurevich & Joyce, 1999; Hall et al., 1996; Ishibashi et al., 2011; A. M. Murray et al., 1994; Riba et al., 2008). In particular, pramipexole administration significantly increased functional connectivity strength within two resting-state networks (i.e., the cinguloopercular network and the somatosensory network), and regional BOLD signal variability in subcortical and prefrontal regions. Pramipexole particularly increased BOLD signal variability in the striatum, OFC and ACC. Although the functional role of BOLD signal variability and its relationship to behavior is yet to be clarified (Aguirre, Zarahn, & D'Esposito, 1998; T. B. Jones, Bandettini, & Birn, 2008; Miller et al., 2002; Neumann, Lohmann, Zysset, & von Cramon, 2003; S. M. Smith et al., 2005), an increasing number of recent investigations points to a link between BOLD signal variability and cognitive abilities (Alavash et al., 2018; Garrett, Kovacevic, McIntosh, & Grady, 2013).

Moreover, the selective effect of pramipexole on facial evaluations could be predicted by this modulation of resting-state dynamics. In particular, pramipexole increased impressions of facial attractiveness, in line with previous studies showing that facial attractiveness evokes activity in dopaminergic regions like the striatum and medial PFC (Aharon et al., 2001; J. O'Doherty et al., 2003; Pegors, Kable, Chatterjee, & Epstein, 2015; D. V. Smith, Clithero, Boltuck, & Huettel, 2014; Winston, O'Doherty, Kilner, Perrett, & Dolan, 2007). In addition, this increase in attractiveness impressions was predicted by the enhanced connectivity strength within the cinguloopercular network, suggesting a direct link between pramipexole's modulation of resting-state dynamics and the drug's effects on behavior. In particular, functional connectivity between the striatum and pTPJ was more strongly associated with higher attractiveness evaluations after pramipexole administration, suggesting that pramipexole might enhance the socially-relevant, reward-based information flow between a pivotal region in reward processing (i.e., the striatum) and another pivotal region in mental states attribution (i.e., the pTPJ) (J. O'Doherty et al., 2004; Saxe & Powell, 2006; W. Schultz et al., 1997; L. Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; L. Young & Saxe, 2008).

An interesting research question for future studies relates to understanding whether fMRI neural dynamics can be used as biomarkers of the pharmacologically-induced neurochemical changes on the neural level. In particular, although widely used as dopamine agonist, it has been suggested that pramipexole might behave as a dopamine antagonist as well, especially when acting on D3 autoreceptors. In fact, D3 autoreceptor activity has been suggested to inhibit the reward-related phasic firing of dopaminergic neurons (Sokoloff et al., 2006). Further, a previous fMRI study has observed reduced fMRI activations in brain structures rich in D3 autoreceptors after pramipexole intake (Riba et al., 2008), which might be linked to an inhibition of dopamine activity. It remains, thus, an open question whether changing dynamics in BOLD signal, and, in particular, functional connectivity might be informative of how drugs operate on receptors. BOLD signal reflects post-synaptic activity and both neurotransmitter agonists and antagonists modulate blood flow (Attwell et al., 2010; Norup Nielsen & Lauritzen, 2001; Zonta et al., 2003). As recent evidence shows that excitatory and inhibitory activity can be modeled from BOLD data (Havlicek, Ivanov, Roebroeck, & Uludag, 2017; Sotero & Trujillo-Barreto, 2007; Sten et al., 2017), variations in BOLD dynamics and functional connectivity might also reflect the impact of a pharmacological administration on neural activity.

10.6 Reducing trust by dopamine-agonist administration

After having shown that administration of a dopamine agonist successfully modulates neural dynamics, Study 6 was conducted to investigate whether pramipexole administration impacts behavioral trust. As mentioned, individuals may ground their trust decisions in subjective impressions about the partner’s trustworthiness that are formed rapidly and effortlessly

(Alexander Todorov et al., 2009) or may dynamically update their beliefs about the partner's trustworthiness based on previous experience with the partner in repeated social interactions (Hula, Vilares, Lohrenz, Dayan, & Montague, 2018). However, on the one hand, the engagement of dopaminergic brain structures in repeated trusting interactions might be related to reward anticipation or reinforcement learning processes (Chang et al., 2010; van 't Wout & Sanfey, 2008). On the other, the effects of dopamine on impression-based trust might be confounded by different types of social information from faces other than facial trustworthiness, such as facial attractiveness (Stirrat & Perrett, 2010; R. K. Wilson & Eckel, 2006).

Thus, Study 6 employed a one-shot TG, where participants were presented with trustees' faces that maximally varied along the trustworthiness dimension with minimal variations in facial attractiveness. Our results first replicated previous evidence that pharmacologically-modulation of dopaminergic functioning does not alter trustworthiness perceptions of others despite successful neural modulation of the brain's reward system (Zebrowitz et al., 2018). Collectively, findings from previous and current studies suggest that neural dynamics and brain regions other than dopaminergic areas likely underlie first subjective impressions of others' social character.

Moreover, we disentangled for the first time the contribution of trustworthiness information to trusting behavior from attractiveness evaluations. Indeed, by reducing the attractiveness information in faces, it was possible to single out the specific effects of facial trustworthiness. Future studies might consider conducting similar experiments by constructing the stimulus material as we did to investigate whether different types of social information from faces induce different trust motives in individuals. Trustworthy-looking individuals, for instance, are likely to be trusted because they signal to be good cooperators (Dunbar, 2004), while trust in attractive others might be driven by reward-based processes, for instance, because of a "beauty premium" associated with attractive individuals (Mobius & Rosenblat, 2006).

Finally, pramipexole affected behavioral trust in the one-shot TG. However, effects of pramipexole on impression-based trust were modulated by hormonal contraceptive use in the female sample. In particular, women using hormonal contraceptives trusted more after pramipexole intake, whereas trust was reduced in non-users. Consistently with the absence of any dopaminergic effect on trustworthiness impressions, such effects on trust were observed across facial trustworthiness dimensions.

The effects of pramipexole on impression-based trust might be due to pramipexole modulation of dopaminergic brain structures as a dopamine agonist or as dopamine antagonist. As dopamine agonist, pramipexole might reduce participants' sensitivity to social contact and feedback by saturating the human need to belong, which would then result in reduced willingness to relate to and connect with others (Baumeister & Leary, 1995). On the contrary, inhibition of dopamine activity might silence the ability to form and maintain satisfying social relationships, equally reducing an individual's willingness to trust.

Unexpectedly, the effects of pramipexole on trust in women using hormonal contraceptives were reversed as compared to non-users. Previous work has shown that hormonal contraceptive use impacts both neural dynamics and behavior in women. On the neural level, functional connectivity in higher-order brain areas associated with social cognition and brain structures related to reward-processing are altered in women using hormonal contraception as compared to naturally cycling women (Bonenberger et al., 2013; Petersen, Kilpatrick, Goharzad, & Cahill, 2014). On the behavioral level, partner choice, attraction to other-sex features (Alvergne & Lummaa, 2010; Feinberg, DeBruine, Jones, & Little, 2008; Roberts, Gosling, Carter, & Petrie, 2008), personal satisfaction and quality of life, especially in relation to romantic relationships (Roberts et al., 2012), have also been shown to change as a function of hormonal contraceptive use.

In particular, contraceptive use shifts women's preferences of partner features to less masculine features (indicative of low testosterone levels) (Gangestad, Garver-Apgar, Simpson,

& Cousins, 2007; Little et al., 2002). Because contraceptive use increases preferences for features such as safety and future security, women using hormonal contraceptive are likely to be more attracted by more trustworthy partners. This preference was indeed observed in Study 6, where women using hormonal contraceptives perceived trustworthy faces as more attractive despite comparable levels of facial attractiveness across trustworthiness dimensions. These findings suggest that pramipexole might intensify such preferences in women using hormonal contraceptive.

10.7 Limitations

These achievements notwithstanding, some limitations have to be addressed that future studies need to overcome for better insights into the psychological and neural dynamics that bring about trusting behaviors.

The TAG allowed to disentangle social information from reward information to test whether these two types of information independently affect trusting behavior. Further, this paradigm allowed me to test whether social and reward information recruit differing brain signatures predictive of trust decisions. However, one of the major limitations of the TAG relates to the unclear motivational structure in the game for the advisers. As pointed out in the introduction, minimizing external incentives for the advisers was intended to prevent participants from tracking the changing probabilistic structure of the incentives of the other and to focus on learning about the other's character instead (Behrens et al., 2008; A. O. Diaconescu et al., 2014; Andreea O. Diaconescu et al., 2017). Moreover, participants knew that they were going to interact with each other on two consecutive games. Importantly, dependency roles were reversed in the two games, as participants depended on the advisers for the outcomes of their decisions in the TAG and the advisers on participants for decision outcomes in the TG. Thus, participants were given the impression that their advisers were motivated to behave honestly in the TAG to form a good reputation that might have paid off in the subsequent interaction in the TG. Despite this, we still lack data about what participants thought the advisers' motivations were to share accurate or inaccurate information or what motivated participants themselves to accept the advice. At the end of the experiments, participants explicitly reported the strategies underlying their decisions. In line with results from their decision patterns, participants' explicit reports suggest they were indeed deciding whether to use advice based on the advisers' honesty. However, as honesty might also elicit impressions of benevolence and good intentions, which are as likely antecedents of trust (Mayer et al., 1995), it is still an open question what inferences participants in the role of advisee were making when they decided whether to trust an adviser's advice.

Secondly, to computationally capture learning dynamics during the trusting interaction, reinforcement learning models were employed that mathematically formalized participants' decisions. This choice was based on previous studies that have provided evidence for their suitability of these models (Biele et al., 2009; Biele et al., 2011; Chang et al., 2010). One of these studies has also shown that reinforcement learning models outperform other models such as Bayesian models (Biele et al., 2009). However, Study 4 did not confirm this assumption with a direct comparison of the employed reinforcement learning models with other models. Trust learning dynamics in the TAG might have been described by other, equally likely models. For example, Bayesian modeling has been shown to optimally describe how individuals integrate information about others' competence (Toelch et al., 2014). Bayesian models might help gain insights into the asymmetry in information weighting observed in Study 4. For instance, the observed learning impairment for honest advisers is likely due to participants' initial positive expectations of the other. These expectations might have functioned as priors for participants' behaviors in the very first stages of the social interaction. As positive expectations are reinforced, they plausibly fostered the formation of strong posteriors that contributed to the discounting of new inconsistent evidence. Bayesian accounts could have captured these dynamics of belief formation and updating. Thus, the absence of a comparison between reinforcement learning models and other types of computational models represents an important limitation to the generalizability of the formalization of the observed learning patterns.

Thirdly, results of the pharmacological studies need to be replicated in a bigger sample size. In particular, the predictive framework based on a leave-one-subject-out cross-validation approach in our small sample of 27 participants might have been affected by unstable and biased estimates that compromise the reliability of our conclusions (Varoquaux, 2017; Varoquaux et al., 2017). The problem is that estimates of variance across cross-validation folds strongly underestimate errors on the prediction accuracy, leading to big error bars. Moreover, future studies might also consider acquiring more subject-level data to use other cross-validation approaches. For instance, the 80-20 cross-validation approach (i.e., training the algorithm on 80% of the data and testing it on the remaining 20%) might provide less biased estimates and thus more reliable results (Varoquaux et al., 2017).

Finally, despite the relevance of providing novel pharmacological evidence on the influence of a dopamine agonist on trusting behaviors, some important issues have to be addressed also for the last study. Due to gender differences in pharmacological interventions using dopaminergic drugs (Munro et al., 2006; Soutschek et al., 2017), we tried to avoid gender variability by limiting our sample to female participants. However, this choice reduces the generalizability of the observed results. Thus, future studies need to replicate these results in a mixed sample. Further, interpretations of the interaction between pramipexole intake and hormonal contraceptive use in affecting behavioral trust are limited by the lack of data on contraceptive type used by the female sample. Different types of hormonal contraceptives may interact in different ways with pharmacological modulations of brain dynamics (Petersen et al., 2014). In addition, a previous study has found weak, but significant evidence on the effects of endogenous sex hormones on interpersonal trust during the preovulatory phase in a sample of 12 naturally cycling women (Ball et al., 2013). As we could not control for menstrual cycle phases, future studies are needed to check whether results hold also after controlling for sex hormones in naturally cycling women. Lastly, the absence of any data on the binding profile of pramipexole limits the insights we can gain into the relationships between dopamine and trust. Hence, future studies using more suitable techniques, such as positron emission tomography, are needed to overcome this issue.

10.8 Future directions

Building on the results provided here, future research might initiate follow-up investigations to extend and complement the knowledge we have acquired from the discussed studies. A possible research line may closely examine the bias in information sampling and processing observed in Chapter 7. For instance, how do decision-makers weigh the same piece of information (e.g., positive and negative outcomes) learnt via social and asocial learning? Do uninformed decision-makers (e.g., who lack previous knowledge about a particular context or about the other) integrate new information in a more biased fashion (for instance, relying more on social learning)? To which extent do decision-makers accept making suboptimal choices and be vulnerable to others' exploitation?

One hypothesis is that information from others, especially from those we trust, is associated with a "social premium" that biases information sampling and processing, particularly in uninformed decision-makers. This bias may be further nourished by reputational concerns, for which individuals may accept taking poor advice to signal trustworthiness and induce the partner to reciprocate in the future. The underlying neural mechanisms might relate to bias-dependent activation patterns in different brain regions: on the one hand, in brain regions associated with learning and prediction error processing (e.g., striatum; Biele et al., 2011); on the other, in regions known to undertake value computations (e.g., VMPFC and OFC; Bartra, McGuire, & Kable, 2013; Sescousse et al., 2010) and inferences on others' mental states (e.g., pTPJ; Igelstrom & Graziano, 2017; Igelström et al., 2015; Igelström et al., 2016; Koster-Hale et al., 2017; Saxe & Powell, 2006; L. Young et al., 2010; L. Young & Saxe, 2008).

Unearthing the psychological and neural mechanisms underlying the integration of information from feedback will shed light on decision-making in particular and human cognition in general. Insights into how information is integrated by a decision-maker before a decision is made will be of pivotal importance to other research fields such as political sciences, economics and nutritional sciences, as similar mechanisms may be involved when individuals decide for whom to vote, which house to buy or what to eat for dinner. Finally, this line of research may provide a testable account to pinpoint the dynamics of clinical symptoms, such as repetitive behaviors in obsessive compulsive disorder or autism in which behavioral rigidity likely hinges on an abnormal integration and evaluation of feedback (Just, Cherkassky, Keller, & Minshew, 2004; Voon et al., 2014). A better understanding of how different sources of information compete and contribute to one's decisions may enrich our knowledge of human cognition, offering a mechanistic framework to improve it.

Moreover, trust and trustworthiness not only foster cooperation and facilitate binding and social integration, they also promote happiness and subjective well-being (Bjørnskov, 2008; McCarthy, Wood, & Holmes, 2017). Recently, a movement in the health and medical domains has advocated for more trust in medicine. The dominant idea, mainly supported by an economic worldview, that a social interaction can be judged solely by its end results (e.g., its profits) is unsatisfying. Not least, because the quality of human relations is defined by their transparency (Bleakley, 2019). Hence, paradoxically, the health of a patient, and thus the success of a therapy, cannot be determined solely by the effectiveness of a drug but also by the type of the relationship between the patient and the doctor (Chen, Tseng, & Cheng, 2013; Elgar, 2010). Trust underlies the formation and maintenance of a long-lasting, supportive and profound relationship between patients and doctors that even predicts reduced mortality (Barker, Steventon, & Deeny, 2017; Pereira Gray et al., 2018). On the contrary, lack of trust leads to over-diagnosing and over-prescribing—the phenomenon of "too much medicine", which has negative health outcomes (Fritz & Holton, 2019).

The psychological and neurobiological mechanisms set in motion by trust and that positively contribute to an individual's subjective well-being are still unexplored. One hypothesis mentioned at the end of Chapter 5 states that trust might act as a social buffer, reducing stress levels and thus improving mental and physical well-being. For example, trust might reduce the perception of decision and outcome uncertainty on the one hand, and the severity of negative outcomes on the other. It might allow for strategies of social support that enable the sharing of responsibility and the burden of difficult decisions with others. In this direction goes preliminary evidence that trust promotes the disclosure of distress (McCarthy et al., 2017). If trust facilitates disclosure of negative emotions, it might set in motion a virtuous loop of emotional disclosure and supportive feedback that improves one's ability to cope with difficult and stressful situations.

The neuropeptide oxytocin, which plays a pivotal role in affiliation, social attachment but also cortisol-level reduction (Panksepp, 2004; L. J. Young & Wang, 2004), might be part of the neurobiological mechanisms that underlie these effects of trust. However, to date, empirical investigations on this topic are still rare. To my knowledge, only one study has pharmacologically investigated the effects of oxytocin on trust in a multi-round TG, finding that oxytocin impairs learning mechanisms that would allow for an adaptive change of trusting strategies. However, it left initial levels of trust intact (T. Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008). Such an impairment was mirrored by reduced activations in the midbrain and striatum, which, as seen in the introduction, are dopaminergic brain regions pivotal to prediction error encoding and learning. Given evidence from animal studies that oxytocin modulates neural activity in the nucleus accumbens and ventral pallidum (Insel & Young, 2001; L. J. Young, Lim, Gingrich, & Insel, 2001), these findings based on a multi-round TG point again to a successful modulation of learning mechanisms during a trusting interaction but leave yet again the question as to what role oxytocin plays in trustworthiness impressions and behavioral trust unanswered.

Conclusions

Taken together, the results outlined in this dissertation demonstrate that honesty is a central determinant of trustworthiness perceptions and trusting behavior. Honest others are quickly identified, and a reputational tag is attached to them that might ease decision-making processes in future encounters. In difficult situations, in which we are required to make a decision, relying on others might reduce the burden of the decision-making process. Preliminary evidence in animal and human research has indicated that trust might serve as a psychological buffer against stress and pain (Burkett et al., 2016; Coan, Schaefer, & Davidson, 2006; Inagaki & Eisenberger, 2012; Kikusui et al., 2006; Yanagisawa et al., 2011). In animals, the mere presence of a peer reduces stress levels in fear-conditioned rats (Davitz & Mason, 1955; Morozov & Ito, 2019). Thus, trust might as well buffer the stress related to a difficult decision in our everyday life.

Conversely, interacting with dishonest others might be more stressful, as dishonest others do not appear to have an intrinsic motivation to commit to the other's well-being. As such, dishonest others might exploit us or simply let us down by providing false or poor information. A decision-maker who has observed signs of dishonesty might prefer refraining from interacting with those unreliable others. However, when this cannot be avoided, a decision-maker might be in a state of high alertness and might constantly track the other's actions to anticipate disadvantageous outcomes. This might explain the flexible behavioral adaption observed in Study 1-4 for dishonest advisers.

Notably, honesty-based trustworthiness was encoded in cortical brain regions associated with higher-order cognition and the neural signal in these regions was informative of future trust decisions. These results outline a specific neural model of honesty-based trust. First, an individual interacting with another and trying to figure out whether she is trustworthy needs to understand the other's intentions and evaluate her character. This elicits cognitive processes that recruit brain regions associated with inferences on others' mental state (i.e., pTPJ and PCC) (Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010; Koster-Hale et al., 2017; Mar, 2011; Saxe & Kanwisher, 2003) and representations of others' character traits (i.e., VMPFC) (Hackel et al., 2015; R. J. Murray et al., 2012; Welborn & Lieberman, 2015). Second, lateral prefrontal (e.g., DLPFC) and parietal regions (e.g., IPL) associated with executive functions and top-down control might be engaged to allow for flexible behavioral revisions based on the updated beliefs about the other's character.

These results raise the question as to how these brain regions are engaged when individuals are presented with contradictory evidence about another's social character. Preliminary evidence from Study 4 suggests that the psychological mechanisms allowing for successful belief updating might be hampered by previous knowledge about the other's honest reputation. This phenomenon could be traced back to a difference in information weighting. Interestingly, a similar asymmetry in the encoding of information from advisers was observed in the OFC (Study 3), suggesting that this brain region might play a pivotal role in successful belief formation and updating in social interactions.

Finally, administration of a dopamine agonist successfully modulated resting-state brain dynamics in subcortical and medial prefrontal regions and was seen to impact trusting behaviors based on subjective impressions of facial trustworthiness. Notably, this effect interacted with the use of common hormonal contraceptives in women. These preliminary results indicate complex neural dynamics between trust and the dopaminergic system. First of all, dopamine does not seem to affect subjective trustworthiness impressions but specifically modulates the behavioral component of trust. Second, this dopaminergic modulation cannot be explained by learning mechanisms, since Study 6 was explicitly designed to control for such a possible confound. Third, the role of dopamine in trust needs to be considered from a broader framework of the interplay between dopamine and other neural dynamics/systems. Four, the possible role of sex hormones in trust might relate not only to sex differences but also to interindividual differences in one's willingness to trust. For instance, hormonal fluctuations in women might be reflected by slight variations of trust over time. Future studies are needed to address these open questions.

By providing, on the one hand, evidence of the psychological and neural dynamics underlying honest reputation and its influence on trusting behaviors, and by highlighting, on the other, the pharmacological impact of a dopamine agonist on impression-based trust, these studies have notable implications for our society and far-reaching consequences not only for research in psychology, neuroscience and pharmacology, but also medicine and politics.

References

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage, 8*(4), 360-369. doi:10.1006/nimg.1998.0369

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

Aimone, J. A., & Houser, D. (2011). Beneficial betrayal aversion. *PLoS One, 6*(3), e17725. doi:10.1371/journal.pone.0017725

Aimone, J. A., & Houser, D. (2012). What you don't know won't hurt you: a laboratory analysis of betrayal aversion. *Experimental Economics, 15*(4), 571-588. doi:10.1007/s10683-012-9314-z

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

Alavash, M., Lim, S. J., Thiel, C., Sehm, B., Deserno, L., & Obleser, J. (2018). Dopaminergic modulation of hemodynamic signal variability and the functional connectome during cognitive performance. *Neuroimage, 172*, 341-356. doi:10.1016/j.neuroimage.2018.01.048

Alvergne, A., & Lummaa, V. (2010). Does the contraceptive pill alter mate choice in humans? *Trends in Ecology & Evolution, 25*(3), 171-179. doi:10.1016/j.tree.2009.08.003

Anderl, C., Steil, R., Hahn, T., Hitzeroth, P., Reif, A., & Windmann, S. (2018). Reduced reciprocal giving in social anxiety - Evidence from the Trust Game. *J Behav Ther Exp Psychiatry, 59*, 12-18. doi:10.1016/j.jbtep.2017.10.005

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron, 65*(4), 550-562. doi:10.1016/j.neuron.2010.02.005

Arias-Carrion, O., Stamelou, M., Murillo-Rodriguez, E., Menendez-Gonzalez, M., & Poppel, E. (2010). Dopaminergic reward system: a short integrative review. *Int Arch Med, 3*, 24. doi:10.1186/1755-7682-3-24

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage, 26*(3), 839-851. doi:10.1016/j.neuroimage.2005.02.018

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

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Pers Soc Psychol Rev, 11*(2), 150-166. doi:10.1177/1088868306294907

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: a review of research and theory. *Pers Soc Psychol Rev, 18*(2), 139-152. doi:10.1177/1088868314523838

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *J Pers Soc Psychol, 86*(2), 356-366. doi:10.1037/0022-3514.86.2.356

Attwell, D., Buchan, A. M., Charpak, S., Lauritzen, M., Macvicar, B. A., & Newman, E. A. (2010). Glial and neuronal control of brain blood flow. *Nature, 468*(7321), 232-243. doi:10.1038/nature09613

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Phys Life Rev*. doi:10.1016/j.plrev.2018.10.002

Baker, R., Honeyford, K., Levene, L. S., Mainous, A. G., 3rd, Jones, D. R., Bankart, M. J., & Stokes, T. (2016). Population characteristics, mechanisms of primary care and premature mortality in England: a cross-sectional study. *BMJ Open, 6*(2), e009981. doi:10.1136/bmjopen-2015-009981

Ball, A., Wolf, C. C., Ocklenburg, S., Herrmann, B. L., Pinnow, M., Brune, M., . . . Gunturkun, O. (2013). Variability in ratings of trustworthiness across the menstrual cycle. *Biol Psychol, 93*(1), 52-57. doi:10.1016/j.biopsycho.2013.01.005

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *R Soc Open Sci, 4*(8), 170193. doi:10.1098/rsos.170193

Barker, I., Steventon, A., & Deeny, S. R. (2017). Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: cross sectional study of routinely collected, person level data. *BMJ, 356*, j84. doi:10.1136/bmj.j84

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang, 68*(3). doi:10.1016/j.jml.2012.11.001

Bartra, O., McGuire, J. T., & 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

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

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic Games. *European Journal of Psychological Assessment, 30*(3), 178-192. doi:10.1027/1015-5759/a000183

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron, 64*(5), 756-770.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*(4), 639-650. doi:10.1016/j.neuron.2008.04.009

Becker, J. B., Perry, A. N., & Westenbroek, C. (2012). Sex differences in the neural mechanisms mediating addiction: a new synthesis and hypothesis. *Biol Sex Differ, 3*(1), 14. doi:10.1186/2042-6410-3-14

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

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

Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S. B., & 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

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

Bicchieri, C. (1990). Norms of Cooperation. *Ethics, 100*(4), 838-861. doi:Doi 10.1086/293237

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bicchieri, C. (2014). Norms, conventions, and the power of expectations. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science: A new introduction* (pp. 208-229). Oxford: Oxford University Press.

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

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

Birnbaum, G. E., Zholtack, K., Mizrahi, M., & Ein-Dor, T. (2019). The Bitter Pill: Cessation of Oral Contraceptives Enhances the Appeal of Alternative Mates. *Evolutionary Psychological Science*. doi:10.1007/s40806-018-00186-6

Bjorklund, A., & Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends Neurosci, 30*(5), 194-202. doi:10.1016/j.tins.2007.03.006

Bjørnskov, C. (2008). Social Capital and Happiness in the United States. *Applied Research in Quality of Life, 3*(1), 43-62. doi:10.1007/s11482-008-9046-6

Bleakley, A. (2019). Invoking the Medical Humanities to Develop a #MedicineWeCanTrust. *Acad Med*. doi:10.1097/ACM.0000000000002870

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

Bohnet, I., & Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review, 94*(2), 362-366. doi:Doi 10.1257/0002828041301506

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127-151. doi:10.1016/j.obhdp.2006.07.001

Bonenberger, M., Groschwitz, R. C., Kumpfmueller, D., Groen, G., Plener, P. L., & Abler, B. (2013). It's all about money: oral contraception alters neural reward processing. *Neuroreport, 24*(17), 951-955. doi:10.1097/WNR.0000000000000024

Boureau, Y. L., & Dayan, P. (2011). Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology, 36*(1), 74-97. doi:10.1038/npp.2010.151

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60*(5), 930-940. doi:10.1016/j.neuron.2008.10.016

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica, 104*(3), 371-398. doi:10.1016/s0001-6918(00)00037-8

Burkett, J. P., Andari, E., Johnson, Z. V., Curry, D. C., de Waal, F. B., & Young, L. J. (2016). Oxytocin-dependent consolation behavior in rodents. *Science, 351*(6271), 375-378. doi:10.1126/science.aac4785

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

Camerer, C., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis (Eds.), *Foundations of Human Sociality - Experimental and Ethnographic Evidence from 15 Small-Scale Societies* (pp. 55-95). Oxford: Oxford University Press.

Camerer, C. F. (2003a). *Behavioral game theory : experiments in strategic interaction*. New York, N.Y., Princeton, N.J.: Russell Sage Foundation; Princeton University Press.

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

Camerer, C. F. (2003c). Psychology and economics. Strategizing in the brain. *Science, 300*(5626), 1673-1675. doi:10.1126/science.1086215

Castner, S. A., Xiao, L., & Becker, J. B. (1993). Sex differences in striatal dopamine: in vivo microdialysis and behavioral studies. *Brain Research, 610*(1), 127-134. doi:10.1016/0006-8993(93)91225-h

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., & Wallace, B. (2008). Heritability of cooperative behavior in the trust game. *Proc Natl Acad Sci U S A, 105*(10), 3721-3726. doi:10.1073/pnas.0710069105

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn Psychol, 61*(2), 87-105. doi:10.1016/j.cogpsych.2010.03.001

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron, 70*(3), 560-572. doi:10.1016/j.neuron.2011.02.056

Chang, L. J., Yarkoni, T., Khaw, M. W., & Sanfey, A. G. (2013). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cereb Cortex, 23*(3), 739-749. doi:10.1093/cercor/bhs065

Chen, C. C., Tseng, C. H., & Cheng, S. H. (2013). Continuity of care, medication adherence, and health care outcomes among patients with newly diagnosed type 2 diabetes: a longitudinal analysis. *Med Care, 51*(3), 231-237. doi:10.1097/MLR.0b013e31827da5b9

Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *Neuroimage, 57*, 583-588. doi:j.neuroimage.2011.04.051

Coan, J. A., Schaefer, H. S., & Davidson, R. J. (2006). Lending a hand: social regulation of the neural response to threat. *Psychol Sci, 17*(12), 1032-1039. doi:10.1111/j.1467-9280.2006.01832.x

Cohen, G. L. (2003). Party over Policy: The Dominating Impact of Group Influence on Politcal Beliefs. *Journal of Personality and Social Psychology, 85*(5), 808-822. doi:10.1037/0022-3514.85.5.808

Cools, R. (2006). Dopaminergic modulation of cognitive function-implications for L-DOPA treatment in Parkinson's disease. *Neurosci Biobehav Rev, 30*(1), 1-23. doi:10.1016/j.neubiorev.2005.03.024

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: unifying affective, activational, and decision functions. *Neuropsychopharmacology, 36*(1), 98-113. doi:10.1038/npp.2010.121

Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T., & Knutson, B. (2010). When giving is good: ventromedial prefrontal cortex activation for others' intentions. *Neuron, 67*(3), 511-521. doi:10.1016/j.neuron.2010.06.030

Corriveau, K., & Harris, P. L. (2009). Choosing your informant: weighing familiarity and recent accuracy. *Dev Sci, 12*(3), 426-437. doi:10.1111/j.1467-7687.2008.00792.x

Cox, J. C. (2002). Trust, Reciprocity, and Other-Regarding Preferences: Group vs. Individuals and Males vs. Females. In R. Zwick & A. Rapoport (Eds.), *Experimental business research* (pp. 331-350). Boston, MA: Kluwer Academic Publisher.

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

Cubitt, R., Gächter, S., & Quercia, S. (2017). Conditional cooperation and betrayal aversion. *Journal of Economic Behavior & Organization, 141*, 110-121. doi:10.1016/j.jebo.2017.06.013

Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes, 112*(1), 11-23. doi:10.1016/j.obhdp.2009.11.007

Davitz, J. R., & Mason, D. J. (1955). Socially facilitated reduction of a fear response in rats. *J Comp Physiol Psychol, 48*(3), 149-151.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature, 446*(7137), 794-796. doi:10.1038/nature05651

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci, 8*(11), 1611-1618. doi:10.1038/nn1575

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., . . . Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput Biol, 10*(9), e1003810. doi:10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Soc Cogn Affect Neurosci*. doi:10.1093/scan/nsw171

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology, 87*(4), 611-628. doi:10.1037/0021-9010.87.4.611

Dosenbach, N. U., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A., . . . Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proc Natl Acad Sci U S A, 104*(26), 11073-11078. doi:10.1073/pnas.0704320104

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., . . . Schlaggar, B. L. (2010). Prediction of individual brain maturity using fMRI. *Science, 329*(5997), 1358-1361. doi:10.1126/science.1194144

Dosenbach, N. U., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron, 50*(5), 799-812. doi:10.1016/j.neuron.2006.04.031

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*(2), 100-110. doi:10.1037/1089-2680.8.2.100

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A, 113*(28), 7900-7905. doi:10.1073/pnas.1602413113

Elgar, F. J. (2010). Income inequality, trust, and population health in 33 countries. *Am J Public Health, 100*(11), 2311-2315. doi:10.2105/AJPH.2009.189134

Engell, A., Haxby, J. V., & Todorov, A. (2007). Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience, 19*(9). doi:10.1162/jocn.2007.19.9.1508

Engelmann, J. B., Meyer, F., Ruff, C. C., & Fehr, E. (2019). The neural circuitry of affect-induced distortions of trust. *Sci Adv, 5*(3), eaau3413. doi:10.1126/sciadv.aau3413

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior, 62*, 287-303. doi:10.1016/j.geb.2007.06.001

Fareri, D. S., Chang, L. J., & 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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of the European Economic Association, 7*(2-3), 235-266.

Fehr, E., & Fischbacher, U. (2002). Why Social Preferences Matter - the Impact of Non-Selfish Motives on Competition, Cooperation and Incentives. *The Economic Journal, 112*(478), C1-C33. doi:10.1111/1468-0297.00027

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

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

Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Little, A. C. (2008). Correlated preferences for men's facial and vocal masculinity. *Evolution and Human Behavior, 29*(4), 233-241. doi:10.1016/j.evolhumbehav.2007.12.008

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1715227115

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations, 7*(2), 117-140. doi:10.1177/001872675400700202

Fisher, H., Aron, A., & Brown, L. L. (2005). Romantic love: an fMRI study of a neural mechanism for mate choice. *J Comp Neurol, 493*(1), 58-62. doi:10.1002/cne.20772

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *Journal of Neuroscience, 33*(8), 3602-3611. doi:Doi 10.1523/Jneurosci.3086-12.2013

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage, 6*(3), 218-229. doi:10.1006/nimg.1997.0291

Fritz, Z., & Holton, R. (2019). Too much medicine: not enough trust? *J Med Ethics, 45*(1), 31-35. doi:10.1136/medethics-2018-104866

Galton, F. (1907). Vox Populi. *Nature, 75*(1949), 450-451. doi:10.1038/075450a0

Gangestad, S. W., Garver-Apgar, C. E., Simpson, J. A., & Cousins, A. J. (2007). Changes in women's mate preferences across the ovulatory cycle. *J Pers Soc Psychol, 92*(1), 151-163. doi:10.1037/0022-3514.92.1.151

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2013). The modulation of BOLD variability between cognitive states varies by age and processing speed. *Cereb Cortex, 23*(3), 684-693. doi:10.1093/cercor/bhs055

Gordon, R., & Spears, K. (2012). You don't act like you trust me: dissociations between behavioural and explicit measures of source credibility judgement. *Q J Exp Psychol (Hove), 65*(1), 121-134. doi:10.1080/17470218.2011.591534

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences, 106*(30), 12506-12511. doi:10.1073/pnas.0900152106

Gurevich, E., & Joyce, J. N. (1999). Distribution of Dopamine D3 Receptor Expressing Neurons in the Human Forebrain Comparison with D2 Receptor Expressing Neurons. *Neuropsychopharmacology, 20*(1), 60-80. doi:10.1016/s0893-133x(98)00066-9

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nat Neurosci, 18*(9), 1233-1235. doi:10.1038/nn.4080

Hall, H., Halldin, C., Dijkstra, D., Wikström, H., Wise, L. D., Pugsley, T. A., . . . Sedvall, G. (1996). Autoradiographic localisation of D 3 -dopamine receptors in the human brain using the selective D 3 -dopamine receptor agonist (+)-[3 H]PD 128907. *Psychopharmacology, 128*(3), 240-247. doi:10.1007/s002130050131

Harris, P. L. (2007). Trust. *Developmental Science, 10*(1), 135-138. doi:10.1111/j.1467-7687.2007.00575.x

Havlicek, M., Ivanov, D., Roebroeck, A., & Uludag, K. (2017). Determining Excitatory and Inhibitory Neuronal Activity from Multimodal fMRI Data Using a Generative Hemodynamic Model. *Front Neurosci, 11*, 616. doi:10.3389/fnins.2017.00616

Heyes, C. (2016). Who Knows? Metacognitive Social Learning Strategies. *Trends Cogn Sci, 20*(3), 204-213. doi:10.1016/j.tics.2015.12.007

Hilbig, B. E., Thielmann, I., Hepp, J., Klein, S. A., & Zettler, I. (2015). From personality to altruistic behavior (and back): Evidence from a double-blind dictator game. *Journal of Research in Personality, 55*, 46-50. doi:10.1016/j.jrp.2014.12.004

Hillebrandt, H., Sebastian, C., & Blakemore, S. J. (2011). Experimentally induced social inclusion influences behavior on trust games. *Cogn Neurosci, 2*(1), 27-33. doi:10.1080/17588928.2010.515020

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci, 1*(4), 304-309. doi:10.1038/1124

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

Hula, A., Vilares, I., Dayan, P., & Montague, P. R. (2017). A Model of Risk and Mental State Shifts during Social Interaction. *preprint arXiv*. doi:1704.03508v2

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

Igelstrom, K. M., & Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia, 105*, 70-83. doi:10.1016/j.neuropsychologia.2017.01.001

Igelström, K. M., Webb, T. W., & Graziano, M. S. (2015). Neural Processes in the Human Temporoparietal Cortex Separated by Localized Independent Component Analysis. *Journal of Neuroscience, 35*(25), 9432-9445. doi:10.1523/JNEUROSCI.0551-15.2015

Igelström, K. M., Webb, T. W., Kelly, Y. T., & Graziano, M. S. (2016). Topographical Organization of Attentional, Social, and Memory Processes in the Human Temporoparietal Cortex. *eNeuro, 3*(2). doi:10.1523/ENEURO.0060-16.2016

Ikemoto, S. (2010). Brain reward circuitry beyond the mesolimbic dopamine system: a neurobiological theory. *Neurosci Biobehav Rev, 35*(2), 129-150. doi:10.1016/j.neubiorev.2010.02.001

Inagaki, T. K., & Eisenberger, N. I. (2012). Neural correlates of giving support to a loved one. *Psychosom Med, 74*(1), 3-7. doi:10.1097/PSY.0b013e3182359335

Insel, T. R., & Young, L. J. (2001). The neurobiology of attachment. *Nat Rev Neurosci, 2*(2), 129-136. doi:10.1038/35053579

Ishibashi, K., Ishii, K., Oda, K., Mizusawa, H., & Ishiwata, K. (2011). Binding of pramipexole to extrastriatal dopamine D2/D3 receptors in the human brain: a positron emission tomography study using 11C-FLB 457. *PLoS One, 6*(3), e17723. doi:10.1371/journal.pone.0017723

Izuma, K. (2012). The social neuroscience of reputation. *Neurosci Res, 72*(4), 283-288. doi:10.1016/j.neures.2012.01.003

Jacobs, E., & D'Esposito, M. (2011). Estrogen shapes dopamine-dependent cognitive processes: implications for women's health. *J Neurosci, 31*(14), 5286-5293. doi:10.1523/JNEUROSCI.6394-10.2011

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

Jones, G. R., & George, J. M. (1998). The Experience and Evolution of Trust: Implications for Cooperation and Teamwork. *Academy of Management Review, 23*(3), 531-546. doi:10.5465/amr.1998.926625

Jones, T. B., Bandettini, P. A., & Birn, R. M. (2008). Integration of motion correction and physiological noise regression in fMRI. *Neuroimage, 42*(2), 582-590. doi:10.1016/j.neuroimage.2008.05.019

Just, M. A., Cherkassky, V. L., Keller, T. A., & Minshew, N. J. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain, 127*, 1811-1821. doi:10.1093/brain/awh199

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Sci Rep, 6*, 39589. doi:10.1038/srep39589

Karns, C. M., Moore, W. E., 3rd, & 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

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social Learning Strategies: Bridge-Building between Fields. *Trends Cogn Sci*. doi:10.1016/j.tics.2018.04.003

Kikusui, T., Winslow, J. T., & Mori, Y. (2006). Social buffering: relief from stress and anxiety. *Philos Trans R Soc Lond B Biol Sci, 361*(1476), 2215-2228. doi:10.1098/rstb.2006.1941

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78-83. doi:10.1126/science.1108062

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science, 314*(5800), 829-832. doi:10.1126/science.1129156

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

Kramer, R. M. (1999). TRUST AND DISTRUST IN ORGANIZATIONS: Emerging Perspectives, Enduring Questions. *Annual Review of Psychology, 50*(1), 569-598. doi:10.1146/annurev.psych.50.1.569

Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philos Trans R Soc Lond B Biol Sci, 363*(1511), 3859-3874. doi:10.1098/rstb.2008.0165

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., . . . Grafman, J. (2007). Neural correlates of trust. *Proc Natl Acad Sci U S A, 104*(50), 20084-20089. doi:10.1073/pnas.0710103104

Lee, K., & Ashton, M. C. (2004). Psychometric Properties of the HEXACO Personality Inventory. *Multivariate Behav Res, 39*(2), 329-358. doi:10.1207/s15327906mbr3902_8

Lemmers-Jansen, I. L. J., Krabbendam, L., Veltman, D. J., & Fett, A. 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

Levine, E. E., Bitterly, T. B., Cohen, T. R., & Schweitzer, M. E. (2018). Who is trustworthy? Predicting trustworthy intentions and behavior. *J Pers Soc Psychol, 115*(3), 468-494. doi:10.1037/pspi0000136

Li, D., Meng, L., & Ma, Q. (2017). Who Deserves My Trust? Cue-Elicited Feedback Negativity Tracks Reputation Learning in Repeated Social Interactions. *Front Hum Neurosci, 11*, 307. doi:10.3389/fnhum.2017.00307

Little, A. C., Jones, B. C., Penton-Voak, I. S., Burt, D. M., & Perrett, D. I. (2002). Partnership status and the temporal context of relationships influence human female preferences for sexual dimorphism in male face shape. *Proc Biol Sci, 269*(1496), 1095-1100. doi:10.1098/rspb.2002.1984

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

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annu Rev Psychol, 62*, 103-134. doi:10.1146/annurev-psych-120709-145406

Masuda, N., & Nakamura, M. (2012). Coevolution of trustful buyers and cooperative sellers in the trust game. *PLoS One, 7*(9), e44169. doi:10.1371/journal.pone.0044169

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

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review, 20*(3), 709-734.

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

McCarthy, M. H., Wood, J. V., & Holmes, J. G. (2017). Dispositional pathways to trust: Self-esteem and agreeableness interact to predict trust and negative emotional disclosure. *Journal of Personality and Social Psychology, 113*(1), 95-116. doi:10.1037/pspi0000093

McElreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., . . . Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior, 26*(6), 483-508. doi:10.1016/j.evolhumbehav.2005.04.003

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Soc Cogn Affect Neurosci, 8*(6), 623-631. doi:10.1093/scan/nss040

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral Self-Licensing: When Being Good Frees Us to Be Bad. *Social and Personality Psychology Compass, 4*(5), 344-357. doi:10.1111/j.1751-9004.2010.00263.x

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

Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *Br J Psychol, 97*(Pt 3), 405-423. doi:10.1348/000712605X85871

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature, 415*(6870), 424-426. doi:10.1038/415424a

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., . . . Gazzaniga, M. S. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *J Cogn Neurosci, 14*(8), 1200-1214. doi:10.1162/089892902760807203

Minson, J. A., & Monin, B. (2011). Do-Gooder Derogation. *Social Psychological and Personality Science, 3*(2), 200-207. doi:10.1177/1948550611415695

Mobius, M. M., & Rosenblat, T. S. (2006). Why Beauty Matters. *American Economic Review, 96*(1), 222-235. doi:10.1257/000282806776157515

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc Natl Acad Sci U S A, 103*(42), 15623-15628. doi:10.1073/pnas.0604475103

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: resenting those who do the right thing. *J Pers Soc Psychol, 95*(1), 76-93. doi:10.1037/0022-3514.95.1.76

Morozov, A., & Ito, W. (2019). Social modulation of fear: Facilitation vs buffering. *Genes Brain Behav, 18*(1), e12491. doi:10.1111/gbb.12491

Munro, C. A., McCaul, M. E., Wong, D. F., Oswald, L. M., Zhou, Y., Brasic, J., . . . Wand, G. S. (2006). Sex differences in striatal dopamine release in healthy adults. *Biol Psychiatry, 59*(10), 966-974. doi:10.1016/j.biopsych.2006.01.008

Murray, A. M., Ryoo, H. L., Gurevich, E., & Joyce, J. N. (1994). Localization of dopamine D3 receptors to mesolimbic and D2 receptors to mesostriatal regions of human forebrain. *Proc Natl Acad Sci U S A, 91*(23), 11271-11275.

Murray, R. J., Schaer, M., & Debbane, M. (2012). Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. *Neurosci Biobehav Rev, 36*(3), 1043-1059. doi:10.1016/j.neubiorev.2011.12.013

Nelson, W. R. (2002). Equity or intention: it is the thought that counts. *Journal of Economic Behavior & Organization, 48*(4), 423-430. doi:10.1016/s0167-2681(01)00245-1

Neumann, J., Lohmann, G., Zysset, S., & von Cramon, D. Y. (2003). Within-subject variability of BOLD response dynamics. *Neuroimage, 19*(3), 784-796. doi:10.1016/s1053-8119(03)00177-0

Nobre, A. C., Coull, J. T., Frith, C. D., & Mesulam, M. M. (1999). Orbitofrontal cortex is activated during breaches of expectation in tasks of visual attention. *Nat Neurosci, 2*(1), 11-12. doi:10.1038/4513

Norup Nielsen, A., & Lauritzen, M. (2001). Coupling and uncoupling of activity-dependent increases of neuronal activity and blood flow in rat somatosensory cortex. *J Physiol, 533*(Pt 3), 773-785. doi:10.1111/j.1469-7793.2001.00773.x

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

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia, 41*(2), 147-155.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr Opin Neurobiol, 14*(6), 769-776. doi:10.1016/j.conb.2004.10.016

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315-324. doi:10.1016/j.jesp.2009.12.002

Olson, I. R., & Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion, 5*(4), 498-502. doi:10.1037/1528-3542.5.4.498

Oskarsson, S., Dawes, C., Johannesson, M., & Magnusson, P. K. (2012). The genetic origins of the relationship between psychological traits and social trust. *Twin Res Hum Genet, 15*(1), 21-33. doi:10.1375/twin.15.1.21

Panksepp, J. (2004). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford: Oxford University Press.

Park, S. Q., Kahnt, T., Dogan, A., Strang, S., Fehr, E., & Tobler, P. N. (2017). A neural link between generosity and happiness. *Nat Commun, 8*, 15964. doi:10.1038/ncomms15964

Pegors, T. K., Kable, J. W., Chatterjee, A., & Epstein, R. A. (2015). Common and unique representations in pFC for face and place attractiveness. *J Cogn Neurosci, 27*(5), 959-973. doi:10.1162/jocn_a_00777

Pereira Gray, D. J., Sidaway-Lee, K., White, E., Thorne, A., & Evans, P. H. (2018). Continuity of care with doctors-a matter of life and death? A systematic review of continuity of care and mortality. *BMJ Open, 8*(6), e021161. doi:10.1136/bmjopen-2017-021161

Petersen, N., Kilpatrick, L. A., Goharzad, A., & Cahill, L. (2014). Oral contraceptive pill use and menstrual cycle phase are associated with altered resting state functional connectivity. *Neuroimage, 90*, 24-32. doi:10.1016/j.neuroimage.2013.12.016

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proc Natl Acad Sci U S A, 107*(29), 13099-13104. doi:10.1073/pnas.1008137107

Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol, 9*, Article39. doi:10.2202/1544-6115.1585

Platt, M. L., & Huettel, S. A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nat Neurosci, 11*(4), 398-403. doi:10.1038/nn2062

Pohjalainen, T., Rinne, J. O., Nagren, K., Syvalahti, E., & Hietala, J. (1998). Sex differences in the striatal dopamine D2 receptor binding characteristics in vivo. *Am J Psychiatry, 155*(6), 768-773. doi:10.1176/ajp.155.6.768

Rapoza, K. A., Vassell, K., Wilson, D. T., Robertson, T. W., Manzella, D. J., Ortiz-Garcia, A. L., & Jimenez-Lazar, L. A. (2016). Attachment as a Moderating Factor Between Social Support, Physical Health, and Psychological Symptoms. *SAGE Open, 6*(4), 215824401668281. doi:10.1177/2158244016682818

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat Rev Neurosci*. doi:10.1038/s41583-019-0179-4

Reimann, M., Schilke, O., & Cook, K. S. (2017). Trust is heritable, whereas distrust is not. *Proc Natl Acad Sci U S A, 114*(27), 7007-7012. doi:10.1073/pnas.1617132114

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: current research and theory*: Appleton-Century-Crofts.

Riba, J., Kramer, U. M., Heldmann, M., Richter, S., & Munte, T. F. (2008). Dopamine agonist increases risk taking but blunts reward-related brain activity. *PLoS One, 3*(6), e2479. doi:10.1371/journal.pone.0002479

Roberts, S. C., Gosling, L. M., Carter, V., & Petrie, M. (2008). MHC-correlated odour preferences in humans and the use of oral contraceptives. *Proc R. Soc. B, 275*, 2715-2722. doi:10.1098/rspb.2008.0825

Roberts, S. C., Klapilova, K., Little, A. C., Burriss, R. P., Jones, B. C., DeBruine, L. M., . . . Havlicek, J. (2012). Relationship satisfaction and outcome in women who meet their partner while using oral contraception. *Proc Biol Sci, 279*(1732), 1430-1436. doi:10.1098/rspb.2011.1647

Rode, J. (2010). Truth and trust in communication: Experiments on the effect of a competitive context. *Games and Economic Behavior, 68*(1), 325-338. doi:10.1016/j.geb.2009.05.008

Rodriguez Buritica, J. M., Heekeren, H. R., & van den Bos, W. (2019). The computational basis of following advice in adolescents. *J Exp Child Psychol, 180*, 39-54. doi:10.1016/j.jecp.2018.11.019

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not So Different after All: A Cross-Discipline View of Trust. *Academy of Management Review, 23*(3), 393-404. doi:10.5465/amr.1998.926617

Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S., & Murray, E. A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nat Neurosci, 16*(8), 1140-1145. doi:10.1038/nn.3440

Sabbagh, M. A., & Shafman, D. (2009). How children block learning from ignorant speakers. *Cognition, 112*(3), 415-422. doi:10.1016/j.cognition.2009.06.005

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in “theory of mind”. *Neuroimage, 19*(4), 1835-1842. doi:10.1016/s1053-8119(03)00230-1

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychol Sci, 17*(8), 692-699. doi:10.1111/j.1467-9280.2006.01768.x

Schultz, W. (2000). Multiple reward signals in the brain. *Nat Rev Neurosci, 1*(3), 199-207. doi:10.1038/35044563

Schultz, W. (2002). Getting Formal with Dopamine and Reward. *Neuron, 36*(2), 241-263. doi:10.1016/s0896-6273(02)00967-4

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275*(5306), 1593-1599. doi:10.1126/science.275.5306.1593

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu Rev Neurosci, 23*, 473-500. doi:10.1146/annurev.neuro.23.1.473

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev, 42*, 9-34. doi:10.1016/j.neubiorev.2014.01.009

Semmann, D., Krambeck, H.-J., & Milinski, M. (2004). Strategic investment in reputation. *Behavioral Ecology and Sociobiology, 56*(3). doi:10.1007/s00265-004-0782-9

Sescousse, G., Redoute, J., & Dreher, J. C. (2010). The architecture of reward value coding in the human orbitofrontal cortex. *Journal of Neuroscience, 30*(39), 13095-13104. doi:10.1523/JNEUROSCI.3501-10.2010

Smith, D. V., Clithero, J. A., Boltuck, S. E., & Huettel, S. A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Soc Cogn Affect Neurosci, 9*(12), 2017-2025. doi:10.1093/scan/nsu005

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., . . . McGonigle, D. J. (2005). Variability in fMRI: a re-examination of inter-session differences. *Hum Brain Mapp, 24*(3), 248-257. doi:10.1002/hbm.20080

Sniezek, J. A., & Buckley, T. (1995). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes, 62*(2), 159-174. doi:10.1006/obhd.1995.1040

Sniezek, J. A., May, D. R., & Sawyer, J. E. (1990). Social uncertainty and interdependence: A study of resource allocation decisions in groups. *Organizational Behavior and Human Decision Processes, 46*(2), 155-180. doi:10.1016/0749-5978(90)90027-7

Sofer, C., Dotsch, R., Wigboldus, D. H., & Todorov, A. (2015). What is typical is good: the influence of face typicality on perceived trustworthiness. *Psychol Sci, 26*(1), 39-47. doi:10.1177/0956797614554955

Sokoloff, P., Diaz, J., Le Foll, B., Guillin, O., Leriche, L., Bezard, E., & Gross, C. (2006). The Dopamine D3 Receptor: A Therapeutic Target for the Treatment of Neuropsychiatric Disorders. *CNS & Neurological Disorders - Drug Targets, 5*(1), 25-43. doi:10.2174/187152706784111551

Sotero, R. C., & Trujillo-Barreto, N. J. (2007). Modelling the role of excitatory and inhibitory neuronal activity in the generation of the BOLD signal. *Neuroimage, 35*(1), 149-165. doi:10.1016/j.neuroimage.2006.10.027

Soutschek, A., Burke, C. J., Raja Beharelle, A., Schreiber, R., Weber, S. C., Karipidis, I. I., . . . Tobler, P. N. (2017). The dopaminergic reward system underpins gender differences in social preferences. *Nature Human Behaviour, 1*(11), 819-827. doi:10.1038/s41562-017-0226-y

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

Sten, S., Lundengard, K., Witt, S. T., Cedersund, G., Elinder, F., & Engstrom, M. (2017). Neural inhibition can explain negative BOLD responses: A mechanistic modelling and fMRI study. *Neuroimage, 158*, 219-231. doi:10.1016/j.neuroimage.2017.07.002

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: male facial width and trustworthiness. *Psychol Sci, 21*(3), 349-354. doi:10.1177/0956797610362647

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*(2), 135-170. doi:10.1037/0033-295x.88.2.135

Thielmann, I., & Hilbig, B. E. (2015). The Traits One Can Trust: Dissecting Reciprocity and Kindness as Determinants of Trustworthy Behavior. *Pers Soc Psychol Bull, 41*(11), 1523-1536. doi:10.1177/0146167215600530

Thorsteinsson, E. B., James, J. E., & Gregg, M. E. (1998). Effects of video-relayed social support on hemodynamic reactivity and salivary cortisol during laboratory-based behavioral challenge. *Health Psychol, 17*(5), 436-444.

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., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Soc Cogn Affect Neurosci, 3*(2), 119-127. doi:10.1093/scan/nsn009

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626. doi:10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & 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

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Social Cognition, 27*(6), 813-833. doi:10.1521/soco.2009.27.6.813

Toelch, U., Bach, D. R., & Dolan, R. J. (2014). The neural underpinnings of an optimal exploitation of social information under uncertainty. *Soc Cogn Affect Neurosci, 9*(11), 1746-1753. doi:10.1093/scan/nst173

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

Tusche, A., Bockler, A., Kanske, P., Trautwein, F. M., & Singer, T. (2016). Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving. *J Neurosci, 36*(17), 4719-4732. doi:10.1523/JNEUROSCI.3392-15.2016

Twenge, J. M., Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Bartels, J. M. (2007). Social exclusion decreases prosocial behavior. *J Pers Soc Psychol, 92*(1), 56-66. doi:10.1037/0022-3514.92.1.56

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci, 27*(15), 4019-4026. doi:10.1523/JNEUROSCI.0564-07.2007 van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796-803. doi:10.1016/j.cognition.2008.07.002

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *Br J Soc Psychol, 44*(Pt 3), 443-461. doi:10.1348/014466604X17092

Vanyukov, P. M., Hallquist, M. N., Delgado, M., Szanto, K., & Dombrovski, A. Y. (2019). Neurocomputational mechanisms of adaptive learning in social exchanges. *Cogn Affect Behav Neurosci*. doi:10.3758/s13415-019-00697-0

Varoquaux, G. (2017). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*. doi:10.1016/j.neuroimage.2017.06.061

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage, 145*(Pt B), 166-179. doi:10.1016/j.neuroimage.2016.10.038

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., . . . Bullmore, E. T. (2014). Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry, 20*(3), 345-352. doi:10.1038/mp.2014.44

Wedekind, C., & Braithwaite, V. A. (2002). The Long-Term Benefits of Human Generosity in Indirect Reciprocity. *Current Biology, 12*(12), 1012-1015. doi:10.1016/s0960-9822(02)00890-4

Weis, S., Hodgetts, S., & Hausmann, M. (2017). Sex differences and menstrual cycle effects in cognitive and sensory resting state networks. *Brain and Cognition*. doi:10.1016/j.bandc.2017.09.003

Welborn, B. L., & Lieberman, M. D. (2015). Person-specific theory of mind in medial pFC. *J Cogn Neurosci, 27*(1), 1-12. doi:10.1162/jocn_a_00700

Williams, K. D., & Sommer, K. L. (1997). Social ostracism by coworkers: Does rejection lead to loafing or compensation? *Personality and Social Psychology Bulletin, 23*(7), 693-706. doi:Doi 10.1177/0146167297237003

Wills, J., FeldmanHall, O., Collaboration, N. P., Meager, M. R., & Van Bavel, J. J. (2018). Dissociable Contributions of the Prefrontal Cortex in Group-Based Cooperation. *Soc Cogn Affect Neurosci*. doi:10.1093/scan/nsy023

Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *J Pers Soc Psychol, 113*(1), 59-80. doi:10.1037/pspi0000092

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *J Exp Psychol Gen, 143*(6), 2074-2081. doi:10.1037/a0038199

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

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

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

Yanagisawa, K., Masui, K., Furutani, K., Nomura, M., Ura, M., & Yoshida, H. (2011). Does higher general trust serve as a psychosocial buffer against social pain? An NIRS study of social exclusion. *Soc Neurosci, 6*(2), 190-197. doi:10.1080/17470919.2010.506139

Yang, Z., Zheng, Y., Yang, G., Li, Q., & L., X. (2018). Neural Signatures of Cooperation Enforcement and Violation: A Coordinate-based Meta-analysis. *Human Brain Mapping*.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*(1), 1-13. doi:10.1016/j.obhdp.2003.08.002

Yaniv, I. (2016). The Benefit of Additional Opinions. *Current Directions in Psychological Science, 13*(2), 75-78. doi:10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260-281. doi:10.1006/obhd.2000.2909

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods, 8*(8), 665-670. doi:10.1038/nmeth.1635

Ye, Z., Hammer, A., & Munte, T. F. (2017). Pramipexole Modulates Interregional Connectivity Within the Sensorimotor Network. *Brain Connect, 7*(4), 258-263. doi:10.1089/brain.2017.0484

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc Natl Acad Sci U S A, 107*(15), 6753-6758. doi:10.1073/pnas.0914826107

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *Neuroimage, 40*(4), 1912-1920. doi:10.1016/j.neuroimage.2008.01.057

Young, L. J., Lim, M. M., Gingrich, B., & Insel, T. R. (2001). Cellular mechanisms of social attachment. *Horm Behav, 40*(2), 133-138. doi:10.1006/hbeh.2001.1691

Young, L. J., & Wang, Z. (2004). The neurobiology of pair bonding. *Nat Neurosci, 7*(10), 1048-1054. doi:10.1038/nn1327

Zebrowitz, L. A., Boshyan, J., Ward, N., Hanlin, L., Wolf, J. M., & Hadjikhani, N. (2018). Dietary dopamine depletion blunts reward network sensitivity to face trustworthiness. *J Psychopharmacol, 32*(9), 965-978. doi:10.1177/0269881118758303

Zonta, M., Angulo, M. C., Gobbo, S., Rosengarten, B., Hossmann, K. A., Pozzan, T., & Carmignoto, G. (2003). Neuron-to-astrocyte signaling is central to the dynamic control of brain microcirculation. *Nat Neurosci, 6*(1), 43-50. doi:10.1038/nn980

Original Studies

Study 1

Bellucci G, Münte T F, Park S Q, *Resting-state dynamics as a neuromarker of dopamine administration in healthy female adults*, J Psychopharmacol, 33 (8).

DOI: https://doi.org/10.1177/0269881119855983

Study 2

The article in this PDF may not exactly replicate the final version published. It is not the copy of record.

Please access the published version via the DOI provided below.

Bellucci, G., Molter, F., & Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. Nature Communications, 10(1).

DOI: https://doi.org/10.1038/s41467-019-13261-8

Title

Neural representations of honesty predict future trust behavior

Abbreviated title

Honesty predicts trust

Manuscript Information

Number of pages: *32*

Number of figures: *6*

Number of words for Abstract/Introduction/Discussion: *154/726/1,509*

Key words: trust, honesty, trust game, ventromedial prefrontal cortex, multivariate voxel pattern analysis, functional connectivity

Abstract

Theoretical accounts propose honesty as a central determinant of trustworthiness perceptions and trusting behavior. However, behavioral and neural evidence on the relationships between honesty and trust is missing. Combining a novel paradigm that successfully induces trustworthiness perceptions with functional MRI and multivariate analyses, we demonstrate that honesty-based trustworthiness is represented in the posterior cingulate cortex, dorsolateral prefrontal cortex and intraparietal sulcus. Crucially, brain signal in these regions predicts individual trust in a subsequent social interaction with the same partners after the scanning session. Importantly, honesty recruited the ventromedial prefrontal cortex (VMPFC), and stronger functional connectivity between the VMPFC and temporoparietal junction during honesty encoding was associated with higher trust in the subsequent interaction. These results suggest that honesty signals in the VMPFC are integrated into trustworthiness beliefs to inform present and future social behaviors. These findings improve our understanding of the neural representations of an individual's social character that guide behaviors during interpersonal interactions.

Trust is the essential component of social life enabling successful cooperation and fostering individuals' well-being. The factors that induce trust in others remain, however, still largely unexplored. To date, at least two accounts have been proposed to explain an individual's trust in others.

One account proposes that interacting agents focus on maximizing their personal payoffs during social exchanges[1]. This account assumes that optimally rational agents trust another as long as they will be better off with trusting than distrusting[2]. Empirical investigations implementing economic games such as the trust game (TG) confirm that people are willing to trust others as long as trusting leads to monetary rewards[3,4]. However, trust levels drop significantly when external incentives lack or when trust leads to monetary losses[5,6].

An alternative account argues that individuals take into account the social character and attitudes of the interacting partner when trusting. In this regard, individuals seek to form beliefs about the other's social character by focusing on whether the other's behavior fosters fairness, equality and cooperation[7,8]. Honesty, that is, the quality of being reliable and the tendency to share truthful information, has been proposed as a central determinant of trustworthiness perceptions promoting prosocial behaviors[9,10]. For instance, altruistic behavior, unconditional kindness and reciprocity have been observed in response to others' honesty[11-14]. However, whether honesty also encourages others to trust is yet unexplored.

These two accounts make different predictions on the neural mechanisms underlying trust. When individuals focus on the trade-off between advantageous and disadvantageous consequences following a trust decision, brain regions signaling actual or hypothetical decision outcomes (such as the ventral striatum and dorsal anterior insula) should be recruited in trusting interactions[15-18]. On the contrary, if trust draws on the social character of the other, brain regions associated with social evaluations (such as the ventromedial prefrontal cortex, VMPFC, and dorsolateral prefrontal cortex, DLPFC) and inferences on the other's intentions (e.g., the posterior temporoparietal junction, pTPJ) should be engaged during trusting behaviors[19-22]. However, to date, evidence on the brain regions representing the honest character of another is still missing.

In this study, we investigated for the first time whether information about the other's honest character evokes trustworthiness perceptions that predict future trust in the other. Importantly, a trustworthy reputation has been suggested to impact information processing during social learning. In particular, although individuals prefer to interact with, and learn from, trustworthy partners[23], beliefs about the other's trustworthiness bias how information from the trustworthy other is processed and learnt[24,25]. An explanatory hypothesis for such bias posits that beliefs about the other's trustworthiness modulate evaluations of information from trustworthy others. For instance, previous work has linked biased beliefs about others' reciprocity to differences in how information is encoded in the orbitofrontal cortex (OFC)[26], a region of pivotal importance in value representation[27]. However, it is still unknown whether honest reputation modulates information encoding and whether the OFC plays a role in such biased information processing.

Here, we developed a trust-inducing paradigm (Take Advice Game, TAG), which enables us to isolate social evaluation signals related to the other person's trustworthiness (learnt through the other's honest and dishonest behavior) from nonsocial value signals related to one's task performance (i.e., neural responses to winnings and losses). Being able to disentangle these two types of information was of pivotal importance to the two main objectives of this study. On one hand, it allowed us to isolate brain signals related to representations of the other's honest character. On the other, it enabled us to investigate any modulatory effects of the other's honest character on information processing. In the TAG, participants, in the role of advisee, had to learn the trustworthiness of advisers from feedback about their honest or dishonest advice. After the TAG, participants, in the role of investor, played a one-shot TG with the advisers who advised them previously.

Using multivariate voxel pattern analysis (MVPA) in combination with functional MRI (fMRI), we examined the relationships between honesty, dishonesty and trust on both behavioral and neural level. On the behavioral level, we hypothesized that honest behavior would increase trust irrespective of proximal benefits associated with the act of trust. On the neural level, representations of the other's trustworthiness in brain regions associated with higher-order cognition would predict individual, behavioral trust across contexts. Finally, we hypothesized that the other's honesty would modulate neural responses to positive and negative outcomes in brain regions associated with value computations.

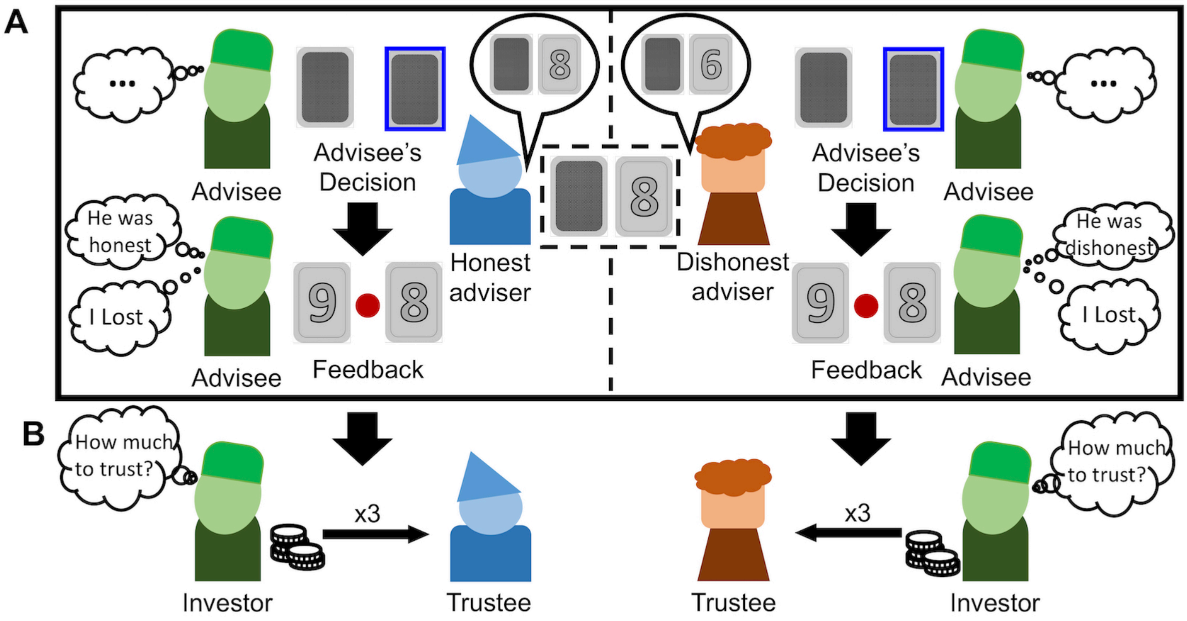


Fig. 1. Paradigms

A. Schematic representation of the Take Advice Game (TAG). Advisers were given information about one of the two cards and could communicate this information to the advisee. Participants, in the role of advisees, made a decision based on the information received (decision phase). In the feedback phase, advisees saw the actual numbers on the cards, which informed them about the adviser's honest behavior (honest vs. dishonest), and a green or red circle, which informed them whether they won or lost, respectively. **B.** After the TAG, participants in the role of investor played a one-shot trust game (TG) with the advisers now in the role of trustee. Investors received a monetary endowment and decided whether they wanted to entrust some of this amount with the trustees. Investors were told that the shared amount was tripled by the experimenter and passed on to the trustee, who could decide to share back any portion of the tripled amount. See also Figure S1.

Results

Paradigms. In the TAG (**Fig. 1A & Fig. S1**), participants in the role of advisee had to rely on the advice of different advisers to choose the highest of two cards. As participants did not have any information about the cards' numbers, they depended on the honesty of the advisers' advice for their decisions. The advisers, on the other hand, could see only one of the two cards, that is, they knew more than the advisees but did not have complete information about the cards. Hence, their advice was not which was the winning card participants should pick but rather additional information about the number of one of the two cards. In each trial, participants were paired with a different adviser (adviser phase). After the adviser sent his advice (advice phase), the advisee decided which card she wanted to pick (decision phase). Finally, the cards were disclosed to the advisee (feedback phase), who could see whether the adviser had been honest and whether she won or lost in that trial. Participants could win/lose €1 in each trial by choosing the card with the higher/lower number. After the TAG, participants in the role of investor played a one-shot TG with the advisers now in the role of trustee (**Fig. 1B**). Investors were paired with each trustee and received an initial endowment of 10 monetary units that they could share with the partner. Investors were told that the shared amount would be tripled by the experimenter and passed on to the trustee who, in turn, could decide to share back any amount of it.

Link between honesty and trusting behavior. First, we tested whether honesty is associated with higher trust levels across contexts and regardless of proximal gains. In the TAG, individuals should be more willing to take the advice of honest advisers and distrust the advice of dishonest advisers. Our results demonstrate that participants took on average more advice from honest than dishonest others ($t_{(30)} = 3.68$; $p < .001$; 95% confidence interval (CI) = [0.03, 0.10]; Cohen's *d* = 0.7; **Fig. 2A**). Importantly, participants grounded their decisions to take an advice in the trustworthiness character of the adviser (i.e., whether the adviser was honest or dishonest; *β* = 0.38; standard error (SE) = 0.12; 95% CI = [0.14, 0.62]; *p* = .007). On the contrary, monetary winnings and losses did not impact participants' decisions to take an adviser's advice (*β* = -0.001; SE = 0.07; 95% CI = [-0.14, 0.14]; *p* = .980; **Tab. 1**). This suggests that our participants trusted an adviser based on the adviser's trustworthy behavior and irrespective of their proximal benefits. Indeed, the majority of our participants (*M* = 88.2%) explicitly reported in an exit questionnaire (see Methods) that their decisions were based on the trustworthiness and advice of the advisers. Importantly, participants applied such trustworthiness-based strategy even though they were aware that it was not successful to gain more benefits (χ^2 = 13.68, *p* = .0002).

Moreover, although trust in the other's advice was comparable for both honest and dishonest advisers in the very first trials of the TAG, participants quickly adjusted their behavior to the other's honesty over the course of the social interaction (**Fig. 2A**). Indeed, participants' advice-taking behaviors toward the two advisers differed increasingly over time (*β* = 0.01; SE = 0.006; 95% CI = [0.0001, 0.024]; *p* = .048), especially due to a significant, linear decrease in trust in the advice of dishonest advisers (*β* = -0.02; SE = 0.007; 95% CI = [-0.028, -0.002]; *p* = .021). On the contrary, advice-taking behavior toward honest advisers did not significantly change over time (*β* = -0.005; SE = 0.006; 95% CI = [-0.016, 0.007]; *p* = .410).

Regressor	***β* (SE)**	**CI**
Intercept	2.01 (0.33)**	1.37, 2.64
Honest adviser	0.38 (0.12)*	0.14, 0.62
Advised Number	-0.17 (0.07)	-0.30, -0.03
Advised Card	0.02 (0.07)	-0.11, 0.15
Feedback previous trial	-0.001 (0.07)	-0.14, 0.14

β coefficients (standard errors) from the generalized mixed-effects logistic regression model with maximal random-effects structure predicting advice-taking behavior (1=advice taken; 0=advice not taken). SE, standard error; CI, confidence interval. **p* < .01; ***p* < .001

Second, we investigated whether these specific effects of the other's trustworthiness on advice-taking behavior in the TAG generalize to a different context and measure of trust (i.e., the TG). Our results confirm this, showing that advice-taking behavior in the TAG correlated with subsequent, economic trust decisions in the TG on average ($\rho_{(29)}$ = .39; *p* = .031), and separately for both honest ($\rho_{(29)}$ = .41; *p* = .021) and dishonest advisers ($\rho_{(29)}$ = -.37; *p* = .040). That is, the more likely participants were to trust the advice of an adviser, the more willing they were to entrust that adviser with money in a subsequent interaction (**Fig. 2B**). As expected, the amount of money shared with the advisers in the TG did not significantly correlate with participants' monetary winnings in the TAG either on average ($\rho_{(29)}$ = .17; *p* = .350) or separately for the two advisers (honest adviser: $\rho_{(29)}$ = .30; *p* = .106; dishonest adviser: $\rho_{(29)}$ = .01; *p* = .978). These results confirm that economic trust decisions in the TG did not represent a form of repayment for the benefits participants obtained from the adviser's advice in the previous interaction but rather reflected participants' willingness to trust the adviser's honesty in advice giving.

Finally, we checked the proportion of positive and negative feedback received by our participants. Participants received on average the same amount of positive and negative feedback (mean difference = 0.0013 ± *SD* = 0.07; $t_{(30)}$ = 0.11; *p* = 0.916), despite more positive feedback for honest than dishonest advisers (honest advisers: *M* = 63.5% ± *SD* = 7.4; dishonest advisers: *M* = 56.7% ± *SD* = 5.0; $t_{(30)}$ = 4.09; *p* < 0.001).

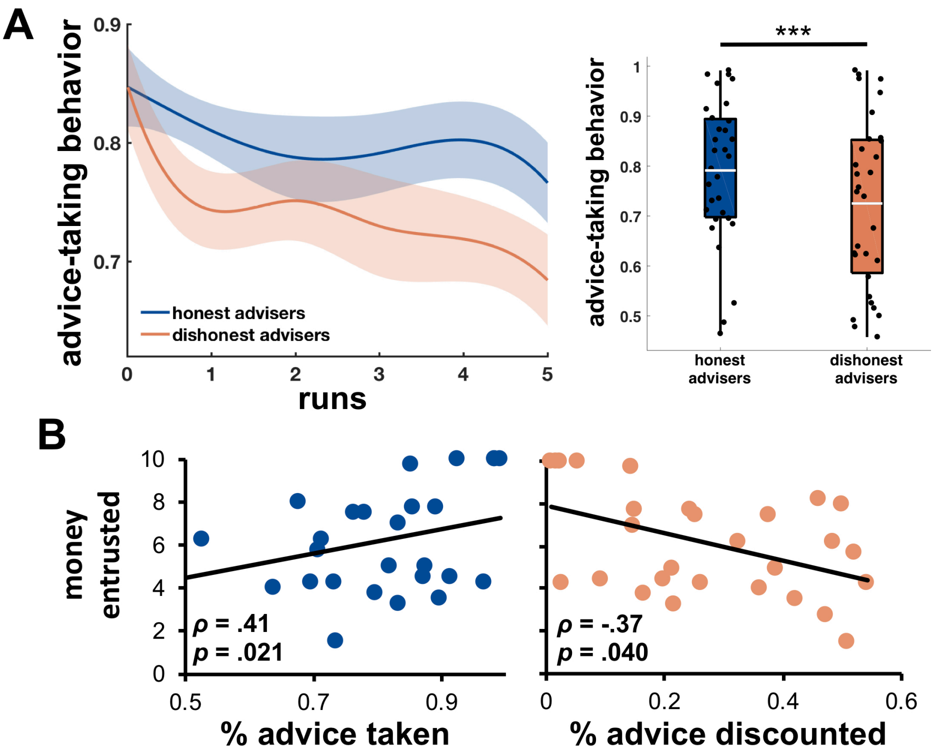


Figure 2. Behavioral results

A. Trusting behavior in the take advice game over runs (left) and on average (right) toward honest and dishonest advisers. Data points on the left were interpolated for visualization purposes and shadowed areas represent standard errors. White lines in the box-plots on the right represent average advice-taking behavior across participants. Each black dot represents one participant. **B.** Amount of money entrusted in the trust game with honest (left) and dishonest (right) others. Each dot represents one participant. *** $p < .001$.

Neural representations of trustworthiness. Next, we examined the common neural patterns of advisers' trustworthiness and value information related to participants' performance. In so doing, we investigated whether these neural patterns capitalize on similar brain regions informative of individual trust. Our task design elegantly allows this, since in the feedback phase, participants received information about the other's trustworthiness (honest/dishonest behavior) and their own task performance (winnings/losses). Hence, applying a whole-brain searchlight MVPA to neural activations during the feedback phase with a leave-one-run-out cross-validation (LOROCV) procedure (**Fig. 3A**), we separately decoded trustworthiness and value information to identify the trustworthiness decoding network and value decoding network, respectively. To this end, a support vector machine (SVM) was trained on beta parameters estimated using two general linear models (GLMs) that coded trustworthiness information (GLM1) and value information (GLM2) in the feedback phase (see Methods).

The trustworthiness decoding network revealed clusters with classification accuracy above chance in the posterior cingulate cortex (PCC), right and left DLPFC, and left IPS (cluster-level, family-wise error corrected, FWEc, < .05; **Fig. 3B & Tab. S1**). Signal in these brain regions was able to classify the neural patterns of honesty and dishonesty of out-of-sample individuals with 68% accuracy (sensitivity: 68%; specificity: 68%; $p < .0001$, based on a nonparametric test of 10,000 permutations; **Fig. 3C**). On the contrary, the value decoding network consisted mainly of regions in the medial PFC extending from the anterior cingulate cortex (ACC) to the striatum (voxel-level FWE < .05; **Fig. 3D & Tab. S1**). Signal in these brain regions was able to classify the neural patterns of positive and negative outcomes of out-of-sample individuals with 82% accuracy (sensitivity: 87%; specificity: 77%; $p < .0002$; **Fig. 3E**). Hence, these analyses indicate a specific neural network representing the other's social character (i.e., trustworthiness) that could be separated from neural signal representing value information. To note, classification accuracy of value information was much better than classification accuracy of social character information. These results concur with previous findings[28] and may hinge on the nature of social concepts, which are distributed neural representations that might be difficult to fully capture using an anatomical-based searchlight approach.

Finally, we set out to characterize the peculiar functional associations of the trustworthiness decoding network. We first ran GLM analyses to control for possible confounds of the observed neural patterns. In particular, we computed another GLM1 adding parametric modulators to the feedback phase for risk (as mean squared deviation from the expected outcome given the adviser's advice) and congruency (as deviance of the adviser's advice from the actual card number on the advised card). These analyses revealed that our results hold also after controlling for these factors (**Fig. S2**). Second, using meta-analytic functional decoding (neurosynth.org)[29], we quantitatively evaluated the representational similarity of the trustworthiness decoding network with neural activation patterns associated with specific psychological components. In particular, we compared the neural signatures of trustworthiness in our study against reverse inference meta-analytic neural patterns of neural images of previous studies stored in the Neurosynth database and associated with particular psychological terms. For this analysis, we chose twelve terms associated with the social and nonsocial domains, such as social cognition, theory of mind, rewards, congruency and risk (**Fig. S3**). Results demonstrate that the trustworthiness decoding network was preferentially associated with psychological terms related to mentalizing and social cognition (**Fig. S3**), validating the ability of our task in singling out neural patterns that likely underlie the formation of trustworthiness beliefs about the advisers. Next, we set up to test this peculiar functional role of the trustworthiness decoding network in representing the trustworthiness character of others.

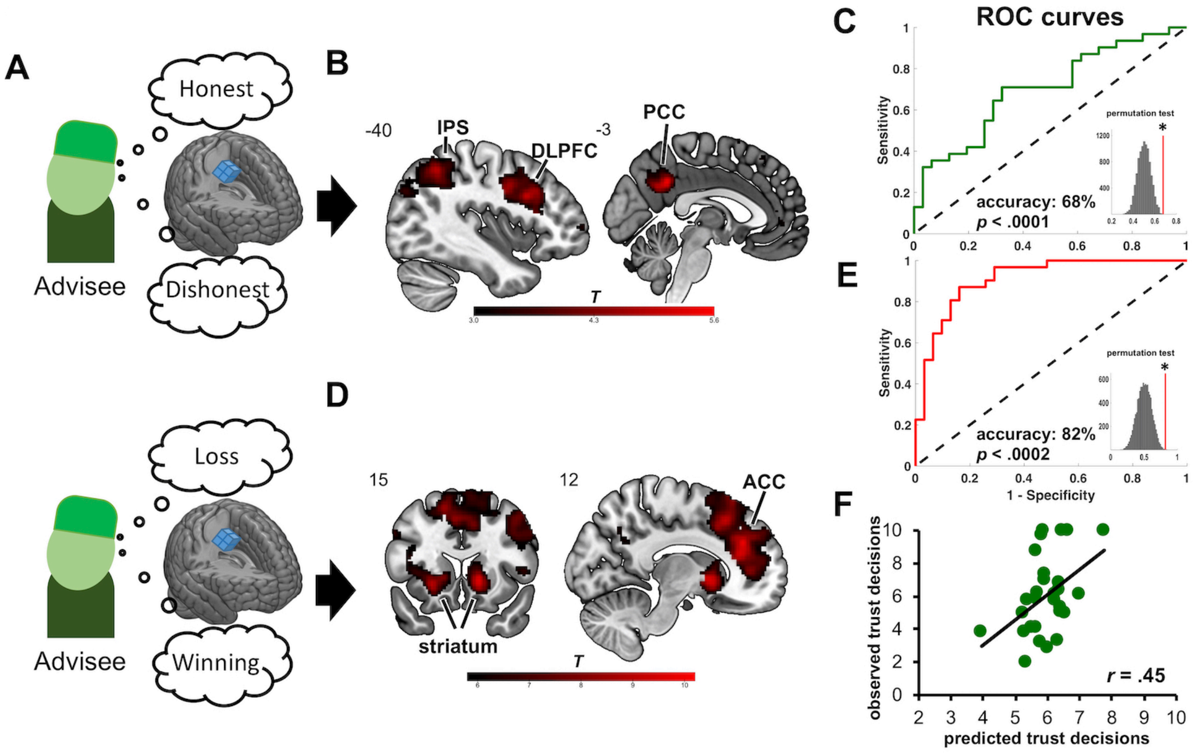


Figure 3. Decoding honesty and predicting trust

In two MVPAs applied to the feedback phase of the TAG (**A**), a support vector machine (SVM) was trained to decode honest and dishonest advice (GLM1) to determine the trustworthiness decoding network (upper), and to decode winnings and losses (GLM2) to determine the value decoding network (lower). The trustworthiness decoding network (**B**) included regions such as the PCC, DLPFC and IPS, and could successfully distinguish neural patterns for honesty and dishonesty in out-of-sample individuals (**C**). The value decoding network (**D**) included the striatum and ACC and could successfully distinguish neural patterns for winnings and losses in out-of-sample individuals (**E**) Finally, a multivariate prediction analysis with support vector regression (SVR) showed that the neural patterns of the trustworthiness decoding network successfully predicted individual economic trust decisions in the TG, thereby showing across-context generalizability (**F**). Both out-of-sample classification and prediction analyses were based on a leave-one-subject-out cross-validation procedure and their significance tested using a permutation test with 10,000 permutations. Each dot represents one participant. See also Figure S2 and Table S1.

MVPA, multivariate voxel pattern analysis; TAG, take advice game; PCC, posterior cingulate cortex; IPS, intraparietal sulcus; DLPFC, dorsolateral prefrontal cortex; ACC, anterior cingulate cortex; TG, trust game. Heatmap represents *t* values.

Neural representations of trustworthiness predict trust. A central feature of the neural representation of a character trait, such as trustworthiness, is its ability to inform decisions across contexts[30]. Thus, neural patterns decoding the other's trustworthiness (i.e., within the trustworthiness decoding network, but not within the value decoding network) should be able to predict individual trust decisions in the TG. To test this, a multivariate prediction analysis with a leave-one-subject-out cross-validation (LOSOCV) procedure was performed. Prediction significance was tested against a random distribution of 10,000 permutations. Results demonstrate that the trustworthiness decoding network significantly predicted the amount of money entrusted in the TG by out-of-sample individuals (standardized mean squared error, smse, = .80; $p < .007$; **Fig. 3F & Fig. S4**). On the contrary, the predictive model based on the value decoding network did not yield a significant prediction (smse = 1.06; $p = .84$; **Fig. S4**). By showing that neural patterns decoding trustworthiness information about others predict an individual's willingness to trust in a different social context, these findings indicate a peculiar functional role of those trustworthiness-decoding brain regions in representing behaviorally-relevant information about others' social characters.

Stronger integration of honesty signals correlates with higher trust. MVPA identified neural patterns of brain signals entailing information about others' trustworthiness that were informative of an individual's trusting behavior and were different from neural patterns related to value information. To further characterize brain regions more strongly recruited by honesty and dishonesty, and to test whether and how honesty modulates neural activations encoding value information, whole-brain univariate analyses were performed on the brain signal during the feedback phase.

Contrast analyses between honesty and dishonesty revealed that dishonesty more strongly activated bilateral DLPFC, left IPS and IPL (FWEc < .05; **Fig. 4A & Tab. S2**), while the VMPFC and ACC were significantly more engaged by honesty (FWEc < .05; **Fig. 4B & Tab. S2**). These results indicate a stronger reliance of dishonesty on brain regions within the trustworthiness decoding network, suggesting that dishonesty likely requires recruitment of brain regions representing the other's character to constantly optimize one's beliefs about the other. On the contrary, honesty more strongly relies on medial prefrontal areas associated with evaluations of positive qualities of others and self.

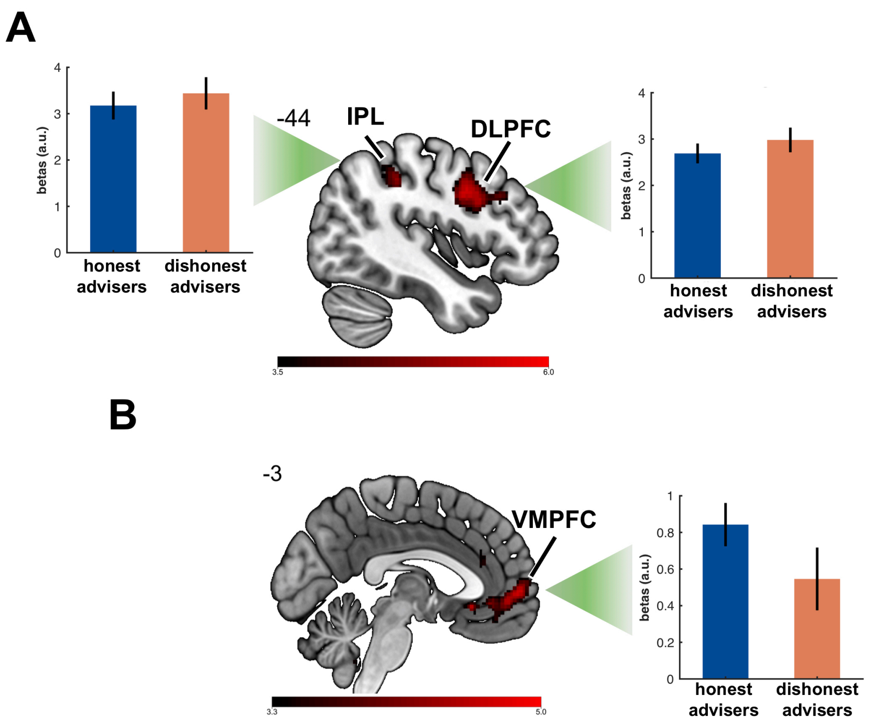


Figure 4. Honesty vs. Dishonesty

Univariate contrasts revealed that brain areas within the trustworthiness decoding network (i.e., IPL and DLPFC) were more engaged by dishonesty than honesty (**A**), whereas honesty more strongly recruited the VMPFC (**B**). Error bars indicate standard errors across participants.

IPL, inferior parietal lobule; DLPFC, dorsolateral prefrontal cortex; VMPFC, ventromedial prefrontal cortex; a.u., arbitary units. Heatmap represents *t* values.

In particular, as the VMPFC has previously been shown to be functionally connected with brain regions associated with social cognition during socially-relevant computations[31], we reasoned that honesty signals in the VMPFC may be integrated into beliefs about the other's social character via functional connectivity with brain regions associated with social cognition. To define these potential pathways, task-dependent functional connectivity analyses were implemented using the VMPFC as seed region. Functional connectivity analyses show that the VMPFC was more strongly functionally coupled to the left pTPJ (-40,-50,30, *x,y,z*; FWEsvc < .05; Fig. 5A) during honesty encoding than dishonesty encoding. Further, if the information flow between the VMPFC and left pTPJ during the feedback phase were specifically associated with the formation of beliefs about the other's trustworthiness, this connectivity signal would be related to subsequent trust decisions but not to individual monetary winnings. Indeed, functional connectivity between the VMPFC and left pTPJ during honesty and dishonesty encoding in the TAG significantly correlated with the amount of money entrusted in the TG to honest ($\rho_{(29)} = .54$; $p < .002$) and dishonest ($\rho_{(29)} = .48$; $p = .006$) advisers (**Fig. 5B**). On the contrary, no significant correlations were found between individual winnings and the VMPFC-pTPJ connectivity for either honest ($\rho_{(29)} = .29$; $p = .111$) or dishonest ($\rho_{(29)} = .11$; $p = .542$) advisers (**Fig. 5C**). These results suggest that functional connectivity between the VMPFC and left pTPJ likely reflects integration of honesty information into knowledge about the other's social character. Specifically, stronger integration of honesty signal from the VMPFC into the pTPJ led to higher trust in others in a subsequent interaction, suggesting that the more participants believed the other to be honest, the more they trusted them.

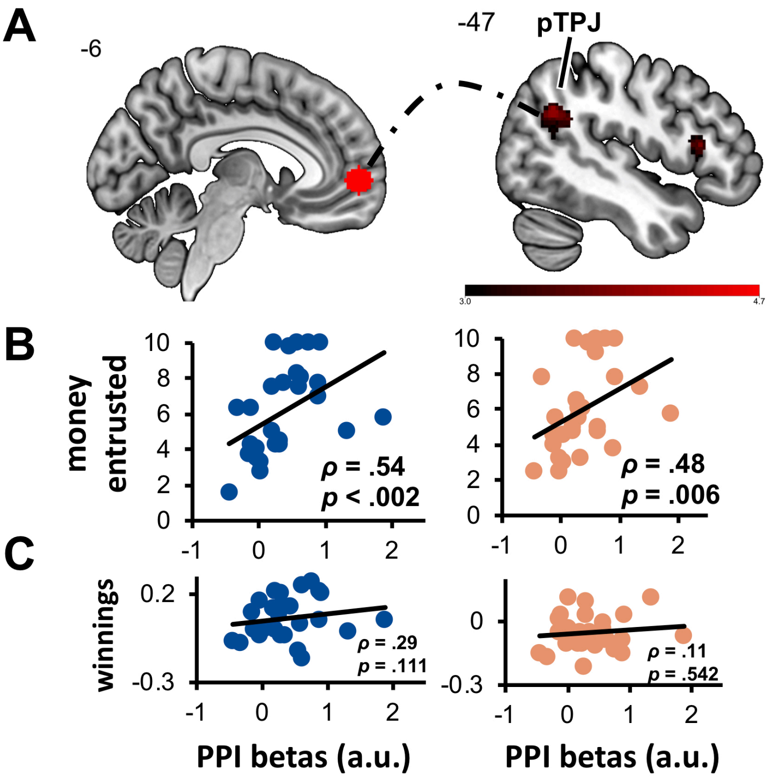


Figure 5. Task-based functional connectivity analysis

Task-based functional connectivity between the VMPFC and left pTPJ was stronger for honesty than dishonesty (**A**). Critically, this functional connectivity correlated with an individual's willingness to trust in the TG (**B**) but not with one's payoffs in the TAG (**C**). Blue dots on correlation plots on the left represent behaviors toward honest advisers, orange dots on correlation plots on the right represent behaviors toward dishonest advisers. Each dot represents one participant.

VMPFC, ventromedial prefrontal cortex; pTPJ, posterior temporo-parietal junction; PPI, psychophysiological interaction; a.u., arbitary units. Heatmap represents *t* values.

Honesty biases value information processing. We then turned to test whether and how these specific activation patterns of honesty and dishonesty modulate brain responses to value information during the feedback phase. Previous behavioral studies have suggested that positive qualities of others bias information processing[24,25]. Such a bias may hinge on trait-dependent differences in neural responses to novel information. We tested this hypothesis by looking at how honesty and dishonesty modulate neural responses to positive and negative outcomes (i.e., GLM3, see Methods).

We first examined the neural responses to positive and negative outcomes during interactions with honest and dishonest advisers separately. Positive outcomes during both interactions with honest and dishonest advisers elicited similar activations in the striatum, and for honest advisers, these activations extended to the OFC (**Tab. S3**). On the other hand, negative outcomes similarly engaged the middle cingulate cortex and inferior frontal gyrus for both honest and dishonest others (**Tab. S4**). Next, we investigated the modulatory effects of honesty and dishonesty on positive and negative outcomes. This analysis revealed that brain regions encoding positive and negative outcomes were differently modulated by honesty and dishonesty. In particular, neural brain signal in the parietal cortex was modulated by dishonesty during both positive (right IPL; FWEc < .05; **Fig. 6A**) and negative (left IPS; FWEc < .05; **Fig. 6B**) outcomes (**Tab. S5**). On the contrary, honesty modulation of neural responses to outcomes was found only in the OFC during positive outcomes (FWEc < .05; Fig. 6C). These results indicate an asymmetry in the neural responses to positive and negative outcomes for honesty.

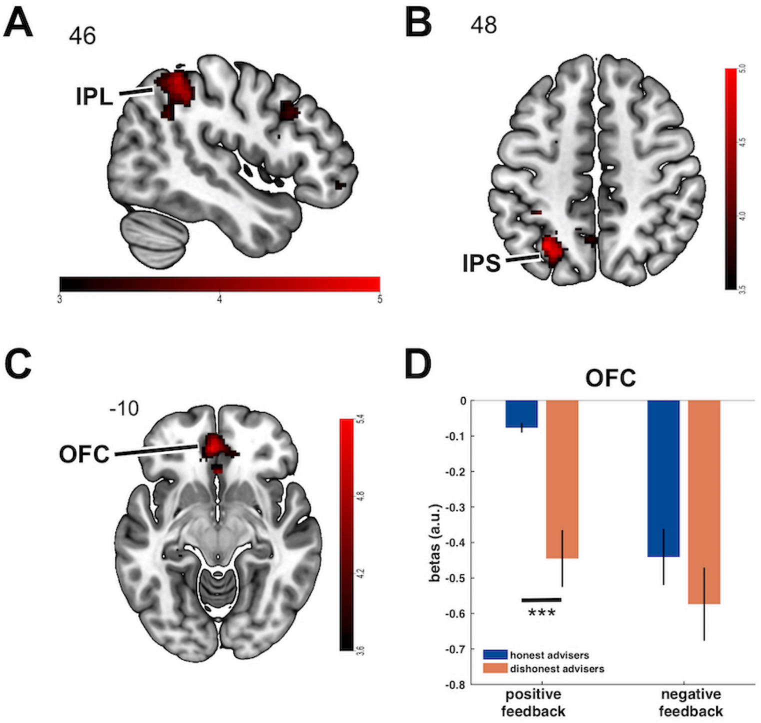


Figure 6. Honesty modulation of feedback-related signal

Whole-brain contrast analyses from GLM3 on the feedback phase yielded significant activations in the parietal cortex for dishonesty during both positive (**A**) and negative feedback (**B**). Honesty, on the contrary, modulated only positive feedback in the OFC (**C**). An ROI analysis (**D**) indicated higher activity in the OFC in response to positive feedback when interacting with honest advisers. Error bars indicate standard errors across participants. See also Table S3, S4 and S5.

ROI, region of interest; IPL, inferior parietal lobule; IPS, intraparietal sulcus; OFC, orbitofrontal cortex; a.u., arbitrary units. Heatmap represents *t* values.

Using an independent ROI in the OFC, we more closely examined in a post-hoc ROI analysis this asymmetric honesty modulation of positive feedback processing (Fig. 6D). Activity in the OFC was significantly higher in response to positive outcomes when interacting with honest advisers as opposed to dishonest advisers (honesty: M = -0.08; SD = 0.46; dishonesty: M = -0.44; SD = 0.35; $t_{(30)}$ = 4.72; p < .0001, CI = [0.21, 0.52]; Cohen's d = 0.85), while OFC activity during negative outcomes was comparable for the two advisers (honesty:

M = -0.45; *SD* = 0.71; dishonesty: *M* = -0.57; *SD* = 0.65; $t_{(30)}$ = 1.16; *p* = .257, CI = [-0.10, 0.36]; Cohen's *d* = 0.21). This asymmetry in the neural responses to positive outcomes in the OFC suggests that feedback information processing may be biased during interactions with honest others.

Discussion

Understanding others is pivotal for successful cooperation. In particular, the other's character may function as a proxy for the other's likely behavior in a future encounter. Thus, trustworthy partners are likely to be trusted in the future, while untrustworthy others are likely to be avoided. In this study, we showed that the honest character of an advisor makes others more likely to accept the adviser's advice and more willing to trust the adviser in a subsequent interaction. Moreover, neural signatures in the DLPFC, IPS and PCC representing the other's trustworthiness predicted individual trust in the partner, and stronger integration of honesty signal from the VMPFC into the pTPJ correlated with higher trust in the other.

If no prior information about how the other will behave is provided, individuals try to gather evidence about the other's social character to inform their decisions. Over the course of multiple interactions, signs of the other's current behavior lay the groundwork for the formation of beliefs about the other's reputation[32]. Consistently with previous models of trust[9], being reliable and telling the truth contributes to an honest reputation that makes others more likely to accept advice. On the contrary, when our participants realized that their initial trust in the other's advice was misplaced, they increasingly discounted the other's advice. Interestingly, even though the adviser's advice was not associated with the best option in the game and did not bring higher benefits to the participants, participants repaid the advisers for their honesty in advice giving in a future trusting interaction.

As there were no incentives for the advisers to help the advisees (except goodwill or a good reputation) and the advisees did not commit to reciprocate, the dynamics in play in our study resemble real-life scenarios in which individuals need to interact with each other without requirements or guarantees of the other's behavior. For instance, trusting someone to give good advice or keep a secret are acts of trust triggered by perceptions of the other's trustworthiness without the requirement of an initial generous act by the trustee[33]. In these contexts, individuals likely assume that the other would comply with the shared social norms, which represent a cluster of expectations an individual can use to make good-enough estimations of the other[34]. In subsequent interactions, individuals would need to quickly learn the trustworthiness of the other based on the other's actual behavior and eventually adopt better behavioral strategies for current and future interactions with the partner[35].

Thus, trusting someone else in a social interaction requires the ability to form a belief about the other's character (i.e., who the other as a person is) and tailor one's behavior to the other's. In our study, we observed that the trustworthiness inferred from the other's honest or dishonest behavior was decoded in four brain regions (i.e., the PCC, IPS and bilateral DLPFC), which were able to successfully classify neural responses to honesty and dishonesty in out-of-sample individuals. In particular, recruitment of the PCC, a central hub of the mentalizing brain system[36,37], is likely related to cognitive processes associated with trait judgments[38], while the IPS, in line with its role in processing expectations related to current goals and stimulus-response selection[39], likely sustains attribution of temporary beliefs to others to tune action selection[38,40]. Finally, the DLPFC might be responsible for translating the knowledge about the partner into action. In particular, in line with its role in generous decisions[41] and group-based cooperation[42,43], the DLPFC might be involved in the decision to engage in prosocial behaviors in response to the other's actions.

Crucially, these brain regions have previously been observed to be interconnected during interpersonal interactions. In particular, the left IPS shows selective connectivity with the DLPFC and PCC while understanding others during social interactions[30,44,45], suggesting that these brain regions build an intertwined brain network engaged in representations of socially-relevant qualities of others. These representations may form an individual's behavioral attitudes based on which adequate behaviors tailored to the other's character are flexibly adopted. Moreover, such representations might be retrieved in future interactions with the partner, as their content is informative of the partner's character and might hence inform individual choices that strongly rely on perceptions of the other's character. In line with this, we observed that neural signal in the IPS, PCC and bilateral DLPFC predicted future individual trust during a social interaction (i.e., in the TG) in which participants had to decide whether to trust the other on the basis of their perceptions of the other's trustworthiness from the previous interaction (i.e., in the TAG).

Critically, the IPS, IPL and DLPFC were also more strongly recruited by dishonesty as opposed to honesty. These findings are in line with previous evidence that the IPS is consistently activated by others' non-cooperative behavior[46], and that the DLPFC, together with the IPL, tracks violations of expectations[47,48] and decisions to lie[49]. The recruitment of these brain regions by dishonesty might reflect the need to constantly track the behaviors of dishonest others for an online update of one's beliefs about them and for flexible revision of one's behavior. In fact, we observed on the behavioral level that advice-taking behavior toward honest advisers did not significantly change over time, while participants continuously adapted their advice-taking behavior for dishonest advisers with a consistent decrease of trust in them over time. These results suggest that the recruitment of the DLPFC, IPS and IPL is more strongly required in cases of norm-deviant behaviors (e.g., being dishonest, unfair or noncooperative) to carefully track the other's actions and optimally adjust one's own behavior.

On the contrary, honesty more strongly recruited the VMPFC, a brain region previously associated with behaviorally-relevant representations of positive traits of others[30,50,51]. In particular, the VMPFC was functionally coupled with the left pTPJ during honesty encoding in the TAG, and the strength of this functional connectivity was further correlated with higher trust in the adviser during a future interaction in the TG. In line with the role of the pTPJ in processing inferences on others' mental states[52,53] and social prediction errors[54,55], these findings suggest that inferences on the other's intentions undertaken by the pTPJ might be supported by integration of novel, incoming information about the other's honesty encoded in the VMPFC. Interestingly, a recent work indicates that connectivity of the left pTPJ with other social cognition regions supports behavioral trust and that an experimentally-induced disruption of trust (via aversive affect) was concomitantly followed by the suppression of pTPJ connectivity during trust decisions. These findings suggest a pivotal role of pTPJ connectivity in integration of behaviorally-relevant signal[56]. In our experiment, stronger integration of an honesty signal likely led to more positive beliefs about the other's intentions, increasing one's willingness to trust. Thus, the interplay between the VMPFC and left pTPJ likely represents a neural mechanism underlying integration of character information for behaviorally-relevant inferences on the other's actions and intentions.

Finally, we observed an honesty modulation of neural responses to value information in the OFC during outcome evaluations. Specifically, higher OFC activity was observed for positive outcomes received when interacting with honest others. These results suggest that in line with its role in processing subjective values of both social and nonsocial rewards[57,58], higher neural activity in the OFC likely reflects an enhanced subjective value of nonsocial rewards induced by the positive character of the other. These neural findings might provide a mechanistic explanation to the positivity bias toward individuals with a good reputation that has been observed to influence learning processes[24,25]. Given the OFC role in learning mechanisms[59,60], an asymmetry in the representation of positive and negative events associated with an individual of good social qualities in the OFC might promote stronger susceptibility to reputational priors and less flexibility in revising one's beliefs about the other. Indeed, previous work has suggested that decreased activity in the OFC is associated with stronger resistance to political belief change during information encoding[61]. Hence, an asymmetric valuation of new incoming information likely contributes to judgmental biases and suboptimal learning.

Taken together, our results improve our understanding of how neural patterns representing honesty-based trustworthiness guide social behaviors in interpersonal interactions. The PCC and frontoparietal brain regions represent behaviorally-relevant knowledge about the other's social character likely taking a role in the flexible revision of one's current behavior for optimal adaptation to the other's actions. Further, social behaviors such as trust are likely enacted based on integration of character information from the VMPFC into the pTPJ for reliable inferences on the good intentions of the partner. Finally, an asymmetric activity in the OFC in response to positive feedback due to the good reputation of the interacting partner likely jeopardizes an individual's ability to optimally form and update one's beliefs about the other, fostering a broad array of judgmental biases. Although we here showed that trustworthiness-related neural signal successfully predicts individual trust decisions in a future social interaction, future studies are still needed to investigate in a brain-to-brain predictive framework whether these neural signatures of trustworthiness are also able to predict the neural patterns recruited by individual trust decisions. Further, future studies, especially in advice-taking paradigms, might also consider controlling for individual susceptibility to social influence, which might, for instance, explain an individual's propensity to take advice from others. Another interesting research question for future studies relates to how other factors of trustworthiness impressions (like competence and benevolence) interact with honesty to elicit trust and/or distrust in others. By shedding light on how social characters are represented in the brain and influence individual decisions, this work makes an important contribution to the extant literature on human cognition in a broad range of scientific fields, such as neuroscience, social psychology, sociology, economics and political sciences.

Materials and Methods

Subjects

Thirty-one participants (20 females) participated in the experiment (age: 24.29±3.81 *M*±*SD*). Participants were recruited from the student community at the University. They were all right-handed and had no history of neurological or psychiatric disorders. Participants gave written informed consent after a complete description of the study was provided. All the procedures involved were in accordance with the Declaration of Helsinki and approved by the Ethical Committee of the University of Lübeck, Lübeck, Germany.

Paradigms

Take Advice Game. In the TAG, participants played as advisee a card game with eight different advisers in a randomized order. Participants were told that these advisers were other participants who were taking part in the same experiment and were preparing themselves in other rooms. Participants were told that roles in the game were randomly assigned by drawing a ball with their role from a lottery box and that all participants were going to do it prior to the experiment. They were told that for transparency reasons, the ball-drawing procedure was going to be performed in front of a camera on top of a screen where each participant could see each of the participants in the other rooms drawing their role. However, to guarantee anonymity, all cameras were mounted on top of the screen so that each participant was recorded only up to the chin. Camera adjustments were performed prior to the ball-drawing procedure to assure this. Moreover, to further guarantee anonymity, each participant needed to choose an avatar that represented themselves in the game (**Fig. S1**). In reality, participants received always the advisee role and the other videos were pre-recorded.

As advisee, participants' task was to draw the card with the higher number. Numbers on the cards ranged from 1 to 9 (except for 5). As participants did not have any information about the card numbers, they needed to rely exclusively on the adviser's advice for their decisions (establishing an adviser-advisee interdependency necessary for trust). Participants were told that the advisers could see only one of the two cards (adviser phase: 2-3s) and could communicate this information to them (advice phase: 1s). This implies that although advisers had more information than our participants, they did not know which card was the winning one, making this setting similar to real-life scenarios in which people generally ask for advice those who may know better, but advisers rarely have complete knowledge of life situations. Participants also knew that advisers could help them but did not have any benefits in doing so. However, both partners knew that after the TAG they were going to play a second game (i.e., the TG, see below), in which participants could repay the advisers for their honesty in advice giving. Thus, in the TAG, advisers were motivated to form a good reputation in the hope that participants would repay them later on. To note, however, participants did not promise or commit to repay the advisers for their advice. The dynamics set into motion by this design resembles real-life interactions in which honest behavior (e.g., giving good advice) has often no proximal benefits to an individual but may help her form a good reputation that might turn out advantageous in the future (a possible, distal benefit).

Moreover, to disentangle trustworthiness information about the advisers from reward information about participants' decisions, the advice of honest advisers was made unpredictive of the winning card (i.e., 50% of the time information about the losing card was given by the honest adviser). Thus, cards were drawn from a uniform distribution with pseudo-random sampling without replacement. The pseudo-random sampling procedure was optimized to have a realized probability of card drawing that approximates chance in both conditions, as would be expected in random drawing. A two-sample Kolmogorov-Smirnov test confirmed that the realized distributions of card numbers did not differ between advisers (*K-S test* = 0.25; $p = .929$). Participants then chose one of the two cards (decision phase: 1s) and saw a final feedback (feedback phase: 1s) in which they received both social information (the card numbers based on which they could infer the adviser's trustworthiness) and nonsocial information (a green or red circle representing winnings and losses, respectively). In each trial, participants could win or lose €1. Intertrial stimulus intervals (ISIs) were 2-8 (*M* = 2.6s) seconds long, whereas jitters between trials were 2-8 (*M* = 4s) seconds long. Participants played a total of 5 runs with 48 trials each (24 with honest and 24 with dishonest advisers) for a total of 240 trials.

Advice-taking behavior in the TAG was operationalized as the probability of choosing a card given the informativeness of the advice received. The optimal strategy in the game would be to choose more frequently a card when the adviser communicated that a number bigger than five is on that card but choose the other card when the adviser communicated that a number smaller than five is on that card. Moreover, as we manipulated the advisers' honesty with the four honest advisers sending accurate information and the four dishonest advisers sending inaccurate information (with 100% contingency), we hypothesized that participants would employ the optimal card-choice strategy differently across honest and dishonest advisers. Analyses of card choice probabilities confirmed our hypotheses (**Fig. S5**). A repeated-measures ANOVA with card numbers as repeated measure yielded a significant main effect of card number ($F_{(7,210)}$ = 83.13; p < .0001; η_p^2 = 0.74) with participants being more likely to choose a card when a number higher than five was said to be on the card and less likely to do so otherwise. Importantly, an interaction effect between card number and advisers was also found ($F_{(7,210)}$ = 4.86; p < .0001; η_p^2 = 0.14). To test the hypothesis that this interaction effect was due to the difference in trust in the advisers and was not simply driven by differences between specific cards, we ran post-hoc t-tests and compared the average choice probability for the honest and dishonest advisers for cards 1-4 and cards 6-9. Results indicate participants were less likely to choose a card when honest advisers told them a low number was on the card (honest vs. dishonest advisers for cards 1-4: t(30) = -2.97; p < 0.006) but more likely to choose a card when honest advisers told them a high number was on the card (honest vs. dishonest advisers for cards 6-9: t(30) = 2.88; p = 0.007). These results suggest that participants were discounting the advice of a dishonest adviser, likely because they did not believe it to be informative. In other words, this decrease in the likelihood of the use of the optimal strategy for dishonest advisers suggests a devaluation of their advice. Overall, these findings indicate that for the same piece of advice, the likelihood someone is going to take the advice hinges on their trust in the adviser or, complementary, on how much they value the adviser's advice (i.e., recognize it as informative).

Trust Game. After the scanning session, participants played as investor a one-shot TG with the same partners who advised them in the TAG. Participants were endowed with 10 monetary units (MUs) for each adviser in the role of trustee and decided whether they wanted to share any of this initial endowment with them (economic trust decision). They were told that any amount they decided to share would be tripled by the experimenter and passed on to the trustee who could in turn decide to share back any portion of this tripled amount (reciprocity decision). The TG was used to probe the transfer effect of the honest reputation established in the TAG on individual trust in a new social interaction.

Exit questionnaire. To acquire an explicit measure of the criteria and motives behind participants' behavior in the TAG, after the experiment, participants were asked to report whether they used any particular strategy and whether they thought this strategy was successful (binary answer option). Although a significant portion of participants reported that they used a strategy in the TAG (χ^2 = 5.89; p = .015), except for 4 participants, no one believed it was successful (χ^2 = 13.68; p = .0002).

Moreover, they were also asked to describe the criteria for their decisions in the TAG (answering the question: "which strategy did you use for your choices in the first game?"). Three researchers blind to the study design and purposes categorized participants' free answers. The first rater identified three main strategies. The second and third raters identified further subcategories for a total of seven and eight categories, respectively. These could be grouped into the three main strategies of the first rater (averaged inter-rater reliability: *r* = .64). For each rater's category, we estimated the percentage of participants using a particular strategy. We then averaged the percentage of participants using each strategy across raters. On average, participants made their decisions 1) intuitively (*M* = 11.8%: rater 1: 9.7%; rater 2: 16%; rater 3: 9.7%), 2) based on the advisers' trustworthiness (*M* = 55.9%: rater 1: 54.8%; rater 2: 51.6%; rater 3: 61.3%), or 3) on the advisers' advice (*M* = 32.3%: rater 1: 35.5%; rater 2: 32.3%; rater 3: 29%). Thus, the majority of our participants (88.2%) explicitly reported to have made their decisions in the TAG based on the adviser's trustworthiness character and advice.

Scanning parameters and preprocessing

Image acquisition. Data were collected with a Siemens MAGNETOM TRIO 3 Tesla scanner at the Freie Universität Berlin. The fMRI scans consisted of an average of 360 contiguous volumes per run (axial slices, 37; slice thickness, 3 mm; interslice gap, 0.6 mm; TR, 2000 ms; TE, 30 ms; flip angle, 70°; voxel size, 3.0×3.0×3.0 mm^3; FOV, 192×192 mm^2). High-resolution structural images were acquired through a 3D sagittal T1-weighted MP-RAGE (sagittal slices, 176; TR, 1900 ms; TE, 2.52 ms; slice thickness, 1.0 mm; voxel size, 1.0×1.0×1.0 mm^3; flip angle, 9°; inversion time, 900 ms; FOV, 256×256 mm^2).

Image preprocessing. Neuroimaging data analyses were performed on SPM12 (v. 6905; http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) in MATLAB 2016b (The Mathworks, Natick, Massachusetts; http://www.mathworks.com/). The functional images were slice-timing corrected, corrected for voxel displacement using field maps and realigned for head movement correction to the mean image. Using the unified segmentation procedure[62], functional images were co-registered to their structural images and subsequently normalized into MNI space using deformation fields (resampling voxel size: 2×2×2 mm^3). Finally, functional images used for univariate analyses were spatially smoothed using a Gaussian filter (8×8×8 mm^3 full width at half maximum, FWHM) to decrease spatial noise. Movement outliers were identified and excluded if head movements/translations were above 3 mm/rad. One run of two participants met these criteria and was therefore excluded from all analyses.

Analyses

Behavioral Analyses. Differences in advice-taking behaviors between honest and dishonest advisers were tested with a one-sample *t*-test. A generalized mixed-effects logistic regression was implemented to investigated whether trial-by-trial advice-taking behavior was predicted by the adviser's honesty irrespective of the benefits associated with the act of trust. A model with the following four regressors was built to predict trust in the adviser's advice (1=trust; 0=distrust): one regressor coding for the adviser's honesty, one for the advised card, one for the advised number, and one for the feedback in the previous trial played with the current adviser. Random-effects structure was based on a 'maximal' approach with by-subject and by-item random intercepts and slopes[63]. *P*-values were computed with a likelihood-ratio test by comparing the full model with the same model without the fixed effect of interest but that it is otherwise identical in random-effects structure[63]. A mixed-effects regression was further fitted to the difference of advice-taking behaviors toward honest and dishonest advisers with run as fixed-effects timing variable and subject as random intercept to test the increase of trust difference over time. Two similar mixed-effects regression models were then separately fitted to each advice-taking behavior toward honest and dishonest advisers in order to examine increases/decreases of trust in the two advisers over time. To test whether trustworthiness relates to subsequent economic trust decisions in a different social context, advice-taking behavior in the TAG was correlated with the amount of money invested in the TG. To further probe that trust decisions in the TG followed from one's perceptions about the other's trustworthiness in the TAG and were not simply reflecting a repaying behavior, correlation analyses (Spearman correlations) were performed between gains in the TAG and money invested in the TG.

Univariate and ROI Analyses. Two general linear models (GLMs) with eight regressors of interest (two for each task phase) on the first level were defined for both univariate and multivariate analyses of fMRI data to be able to estimate beta parameters that uniquely capture neural signals related to trustworthiness and value encoding, respectively. GLM1 consisted of the following regressors: 2 regressors for the advisor phase, 2 regressors for the advice phase, 2 regressors for the decision phase and 2 regressors for the feedback phase coding the adviser's trustworthiness (honesty/dishonesty). GLM2 entailed the same regressors as GLM1 with the exception that the 2 regressors for the feedback phase coded value information (gain/loss). Control analyses were performed to check that the neural signatures of trustworthiness were not confounded by other factors. In particular, we re-ran GLM1 adding further regressors and parametric modulators to account for variance that might be due to risk and congruency effects. To control for risk, two orthogonal parametric modulators were added to the two regressors coding honesty and dishonesty in the feedback phase; namely, a 1st order term for reward probability given the adviser's advice and a 2nd order term for reward variance (i.e., the mean squared deviation from expected outcome), which is quadratic in reward probability *p* and refers to the expected risk given the adviser's advice[64]. Second, to control for contingency effects (i.e., informational deviance between the adviser's advice and the actual card number on the advised card), we added a regressor coding for all feedback phases (i.e., across advisers) with duration 1s and degrees of congruency (continuous variable) as parametric modulator.

Finally, to separately investigate brain activations for responses to positive and negative feedback when interacting with honest and dishonest advisers, and to analyze the honesty modulation of brain regions processing feedback information, GLM3 was defined encompassing a total of 10 regressors of interest. All task phases had the same regressors as

GLM1 and GLM2, except for the feedback phase, for which 4 regressors were defined coding winnings and losses received when advised by honest and dishonest advisers, separately. In all GLMs, conditions were modeled as events using a stick function (i.e., setting the duration of each condition to 0).

Motion parameters were further included as regressors of no-interest in all GLMs. A temporal high-pass filter with a cutoff of 128 seconds was applied for all GLMs. Results were whole-brain corrected for multiple comparison using a voxel-level threshold of $p < .001$ and a family-wise error, cluster-level ($\mathrm{FWE_c}$) corrected threshold of $p < .05$[65]. The ROI analysis for the OFC (area s32) to post-hoc examine the honesty modulation of positive outcomes was based on the probabilistic map provided by the SPM Anatomy toolbox, v. 2.2[66].

Multivariate voxel pattern analyses. Decoding analyses to investigate the neural representations of trustworthiness (honesty/dishonesty) and value (winnings/losses) information were performed using a linear support vector machine (SVM) algorithm for binary classification and a whole-brain searchlight approach with a searchlight's radius size of 10mm. Applying a leave-one-run-out cross-validation (LOROCV), the SVM was trained on all but one run and tested on the left-out run. This procedure was repeated *n* times with *n*=5 (total number of runs) and the algorithm's cross-validated accuracy was computed. To decode character information related to the advisers' trustworthiness, beta images from the feedback phase of GLM1 (fitted to unsmoothed, normalized brain images) were used. To decode feedback information related to winnings and losses, beta images from the feedback phase of GLM2 (fitted to unsmoothed, normalized brain images) were used. Searchlight decoding analyses were applied to all voxels within the whole-brain gray matter probability mask provided by SPM and thresholded at 0.1.

Decoding generalization of the trustworthiness and value decoding networks was tested with a classification analysis using a leave-one-subject-out cross-validation (LOSOCV) approach in which the SVM was trained on *z*-scored average beta images of all but one participant and tested on the left-out participant. Cross-validated accuracy of the group-level classification was tested for significance running a permutation test with 10,000 permutations (*n_perm*). In each permutation, the SVM was trained on randomly permuted labels using the same LOSOCV approach of the true classification model. The sum of models trained on permuted labels that performed better than the true model was then computed (*p_models*). The nonparametric p value was assessed including the observed statistics according to the following formula[67]: ${(1 + p\_models)}/{(1 + n\_perm)}$. Multivariate prediction analyses to predict subsequent, economic trust decisions in the TG from the trustworthiness and value decoding networks were based on the same LOSOCV procedure and permutation test but used support vector regression (SVR) for prediction of continuous variables.

Decoding analyses were run using The Decoding Toolbox TDT, v. 3.99[68] and custom MATLAB scripts.

Meta-analytic functional decoding. To characterize the functional specification of the trustworthiness decoding network, a meta-analytic image decoding analysis was performed using the Neurosynth Image Decoder (neurosynth.org)[29]. The Neurosynth Image Decoder allows to quantitatively estimate the representational similarity between any task-based activation pattern and meta-analytical activation patterns associated with particular terms and generated based on brain images in the Neurosynth database[69]. Similarity was computed as Pearson’s correlations across all voxels between the task-based and the meta-analytical maps. We selected meta-analytic maps based on 12 different terms to test the specific a priori hypothesis that the trustworthiness decoding map more likely related to functional roles in the social domain as opposed to the reward, risk and congruency domains. It has to be noted that the observed correlations are relatively small but in line with previous research[70]. Moreover, while the analysis is quantitative, the conclusions that can be drawn are descriptive in nature, as there is no inference statistics that tested whether any of the observed correlation coefficients is significantly higher than the others.

Task-dependent functional connectivity analyses. To test the information flow between the VMPFC underlying honesty signals and any regions across the whole brain, a task-dependent functional connectivity analysis was implemented using a whole-brain psychophysiological interaction analysis (PPI[71] with seed region (10mm radius) around the VMPFC peak coordinates yielded by the univariate contrast. The PPI-GLM consisted of a task regressor, a physiological regressor entailing deconvolved blood-oxygen-level-dependent (BOLD) signal from the seed region and a regressor for the interaction term with movement parameters as regressors of no interest. Significant connectivity was assessed with a voxel-level threshold of $p < .001$ and an FWE cluster-level threshold of $p < .05$ within the ROI[72].

Labeling and data visualization. The SPM Anatomy toolbox v. 2.2[66] and MRIcron (http://people.cas.sc.edu/rorden/mricron/install.html/) were used for anatomical labeling. MRIcroGL (https://www.mccauslandcenter.sc.edu/mricrogl/home/) was used for brain visualizations.

References

1 Fehr, E. & Fischbacher, U. Why Social Preferences Matter - the Impact of Non-Selfish Motives on Competition, Cooperation and Incentives. *The Economic Journal* **112**, C1-C33, doi:10.1111/1468-0297.00027 (2002).

2 Fehr, E. On the Economics and Biology of Trust. *Journal of the European Economic Association* **7**, 235-266 (2009).

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

4 Hula, A., Vilares, I., Dayan, P. & Montague, P. R. A Model of Risk and Mental State Shifts during Social Interaction. *preprint arXiv*, doi:1704.03508v2 (2017).

5 Rode, J. Truth and trust in communication: Experiments on the effect of a competitive context. *Games and Economic Behavior* **68**, 325-338, doi:10.1016/j.geb.2009.05.008 (2010).

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

7 Barber, B. *The Logic and Limits of Trust*. 190 (Rutgers University Press, 1983).

8 Jones, G. R. & George, J. M. The Experience and Evolution of Trust: Implications for Cooperation and Teamwork. *Academy of Management Review* **23**, 531-546, doi:10.5465/amr.1998.926625 (1998).

9 Mayer, R. C., Davis, J. H. & Schoorman, F. D. An Integrative Model of Organizational Trust. *The Academy of Management Review* **20**, 709-734 (1995).

10 Ashton, M. C., Lee, K. & de Vries, R. E. The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: a review of research and theory. *Pers Soc Psychol Rev* **18**, 139-152, doi:10.1177/1088868314523838 (2014).

11 Thielmann, I. & Hilbig, B. E. The Traits One Can Trust: Dissecting Reciprocity and Kindness as Determinants of Trustworthy Behavior. *Pers Soc Psychol Bull* **41**, 1523-1536, doi:10.1177/0146167215600530 (2015).

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

13 Hilbig, B. E., Thielmann, I., Hepp, J., Klein, S. A. & Zettler, I. From personality to altruistic behavior (and back): Evidence from a double-blind dictator game. *Journal of Research in Personality* **55**, 46-50, doi:10.1016/j.jrp.2014.12.004 (2015).

14 Baumert, A., Schlösser, T. & Schmitt, M. Economic Games. *European Journal of Psychological Assessment* **30**, 178-192, doi:10.1027/1015-5759/a000183 (2014).

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

16 Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S. B. & Krueger, F. 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 (2018).

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

18 Delgado, M. R., Frank, R. H. & Phelps, E. A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature neuroscience* **8**, 1611-1618, doi:10.1038/nn1575 (2005).

19 Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T. & Knutson, B. When giving is good: ventromedial prefrontal cortex activation for others' intentions. *Neuron* **67**, 511-521, doi:10.1016/j.neuron.2010.06.030 (2010).

20 Buckholtz, J. W. *et al.* The neural correlates of third-party punishment. *Neuron* **60**, 930-940,

doi:10.1016/j.neuron.2008.10.016 (2008).

21 Tusche, A., Bockler, A., Kanske, P., Trautwein, F. M. & Singer, T. Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving. *J Neurosci* **36**, 4719-4732, doi:10.1523/JNEUROSCI.3392-15.2016 (2016).

22 FeldmanHall, O. *et al.* Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1715227115 (2018).

23 Harris, P. L. Trust. *Developmental Science* **10**, 135-138, doi:10.1111/j.1467-7687.2007.00575.x (2007).

24 Sabbagh, M. A. & Shafman, D. How children block learning from ignorant speakers. *Cognition* **112**, 415-422, doi:10.1016/j.cognition.2009.06.005 (2009).

25 Corriveau, K. & Harris, P. L. Choosing your informant: weighing familiarity and recent accuracy. *Dev Sci* **12**, 426-437, doi:10.1111/j.1467-7687.2008.00792.x (2009).

26 Ide, J. S. *et al.* 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 (2018).

27 Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S. & Murray, E. A. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nature neuroscience* **16**, 1140-1145, doi:10.1038/nn.3440 (2013).

28 Woo, C. W. *et al.* Separate neural representations for physical pain and social rejection. *Nat Commun* **5**, 5380, doi:10.1038/ncomms6380 (2014).

29 Rubin, T. N. *et al.* Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. *PLoS Comput Biol* **13**, e1005649, doi:10.1371/journal.pcbi.1005649 (2017).

30 Hackel, L. M., Doll, B. B. & Amodio, D. M. Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature neuroscience* **18**, 1233-1235, doi:10.1038/nn.4080 (2015).

31 Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P. & Rangel, A. Value Computations in Ventral Medial Prefrontal Cortex during Charitable Decision Making Incorporate Input from Regions Involved in Social Cognition. *Journal of Neuroscience* **30**, 583-590, doi:10.1523/jneurosci.4089-09.2010 (2010).

32 Krueger, F. *et al.* Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20084-20089, doi:10.1073/pnas.0710103104 (2007).

33 Levine, E. E., Bitterly, T. B., Cohen, T. R. & Schweitzer, M. E. Who is trustworthy? Predicting trustworthy intentions and behavior. *J Pers Soc Psychol* **115**, 468-494, doi:10.1037/pspi0000136 (2018).

34 Bicchieri, C. in *Philosophy of social science: A new introduction* (eds N. Cartwright & E. Montuschi) 208-229 (Oxford University Press, 2014).

35 Falk, A., Fehr, E. & Fischbacher, U. Testing theories of fairness—Intentions matter. *Games and Economic Behavior* **62**, 287-303, doi:10.1016/j.geb.2007.06.001 (2008).

36 Mar, R. A. The neural bases of social cognition and story comprehension. *Annu Rev Psychol* **62**, 103-134, doi:10.1146/annurev-psych-120709-145406 (2011).

37 Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R. & Buckner, R. L. Functional-anatomic fractionation of the brain's default network. *Neuron* **65**, 550-562, doi:10.1016/j.neuron.2010.02.005 (2010).

38 Schurz, M., Radua, J., Aichhorn, M., Richlan, F. & Perner, J. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev* **42**, 9-34, doi:10.1016/j.neubiorev.2014.01.009 (2014).

39 Corbetta, M., Patel, G. & Shulman, G. L. The reorienting system of the human brain: from environment to theory of mind. *Neuron* **58**, 306-324, doi:10.1016/j.neuron.2008.04.017 (2008).

40 Igelstrom, K. M. & Graziano, M. S. A. The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia* **105**, 70-83,

doi:10.1016/j.neuropsychologia.2017.01.001 (2017).

41 Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* **314**, 829-832, doi:10.1126/science.1129156 (2006).

42 Wills, J., FeldmanHall, O., Collaboration, N. P., Meager, M. R. & Van Bavel, J. J. Dissociable Contributions of the Prefrontal Cortex in Group-Based Cooperation. *Social cognitive and affective neuroscience*, doi:10.1093/scan/nsy023 (2018).

43 Lemmers-Jansen, I. L. J., Krabbendam, L., Veltman, D. J. & Fett, A. J. 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 (2017).

44 Igelström, K. M., Webb, T. W., Kelly, Y. T. & Graziano, M. S. Topographical Organization of Attentional, Social, and Memory Processes in the Human Temporoparietal Cortex. *eNeuro* **3**, doi:10.1523/ENEURO.0060-16.2016 (2016).

45 Mende-Siedlecki, P., Cai, Y. & Todorov, A. The neural dynamics of updating person impressions. *Social cognitive and affective neuroscience* **8**, 623-631, doi:10.1093/scan/nss040 (2013).

46 Yang, Z., Zheng, Y., Yang, G., Li, Q. & L., X. Neural Signatures of Cooperation Enforcement and Violation: A Coordinate-based Meta-analysis. *Human Brain Mapping* (2018).

47 Chang, L. J., Yarkoni, T., Khaw, M. W. & Sanfey, A. G. 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 (2013).

48 Cloutier, J., Gabrieli, J. D. E., O'Young, D. & Ambady, N. An fMRI study of violations of social expectations: When people are not who we expect them to be. *Neuroimage* **57**, 583-588, doi:j.neuroimage.2011.04.051 (2011).

49 Greene, J. D. & Paxton, J. M. Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences* **106**, 12506-12511, doi:10.1073/pnas.0900152106 (2009).

50 Murray, R. J., Schaer, M. & Debbane, M. Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. *Neurosci Biobehav Rev* **36**, 1043-1059, doi:10.1016/j.neubiorev.2011.12.013 (2012).

51 Welborn, B. L. & Lieberman, M. D. Person-specific theory of mind in medial pFC. *J Cogn Neurosci* **27**, 1-12, doi:10.1162/jocn_a_00700 (2015).

52 Saxe, R. & Powell, L. J. It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological science* **17**, 692-699, doi:10.1111/j.1467-9280.2006.01768.x (2006).

53 Young, L. & Saxe, R. The neural basis of belief encoding and integration in moral judgment. *Neuroimage* **40**, 1912-1920, doi:10.1016/j.neuroimage.2008.01.057 (2008).

54 Diaconescu, A. O. *et al.* Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput Biol* **10**, e1003810, doi:10.1371/journal.pcbi.1003810 (2014).

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

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

57 Valentin, V. V., Dickinson, A. & O'Doherty, J. P. Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci* **27**, 4019-4026, doi:10.1523/JNEUROSCI.0564-07.2007 (2007).

58 Sescousse, G., Redoute, J. & Dreher, J. C. The architecture of reward value coding in the human orbitofrontal cortex. *J Neurosci* **30**, 13095-13104, doi:10.1523/JNEUROSCI.3501-10.2010 (2010).

59 Tsuchida, A., Doll, B. B. & Fellows, L. K. Beyond reversal: a critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. *J Neurosci* **30**, 16868-16875, doi:10.1523/JNEUROSCI.1958-10.2010 (2010).

60 Gottfried, J. A. & Dolan, R. J. Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. *Nature neuroscience* **7**, 1144-1152, doi:10.1038/nn1314 (2004).

61 Kaplan, J. T., Gimbel, S. I. & Harris, S. Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Sci Rep* **6**, 39589, doi:10.1038/srep39589 (2016).

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

63 Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang* **68**, doi:10.1016/j.jml.2012.11.001 (2013).

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

65 Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 7900-7905, doi:10.1073/pnas.1602413113 (2016).

66 Eickhoff, S. B. *et al.* Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *Neuroimage* **36**, 511-521, doi:10.1016/j.neuroimage.2007.03.060 (2007).

67 Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* **9**, Article39, doi:10.2202/1544-6115.1585 (2010).

68 Hebart, M. N., Gorgen, K. & Haynes, J. D. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front Neuroinform* **8**, 88, doi:10.3389/fninf.2014.00088 (2014).

69 Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* **2**, 4, doi:10.3389/neuro.06.004.2008 (2008).

70 Gao, X. *et al.* Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *PNAS* **115**, E7680-E7689, doi:10.1073/pnas.1802523115 (2018).

71 Friston, K. J. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218-229, doi:10.1006/nimg.1997.0291 (1997).

72 Igelström, K. M., Webb, T. W. & Graziano, M. S. Neural Processes in the Human Temporoparietal Cortex Separated by Localized Independent Component Analysis. *J Neurosci* **35**, 9432-9445, doi:10.1523/JNEUROSCI.0551-15.2015 (2015).

Supplemental Figures

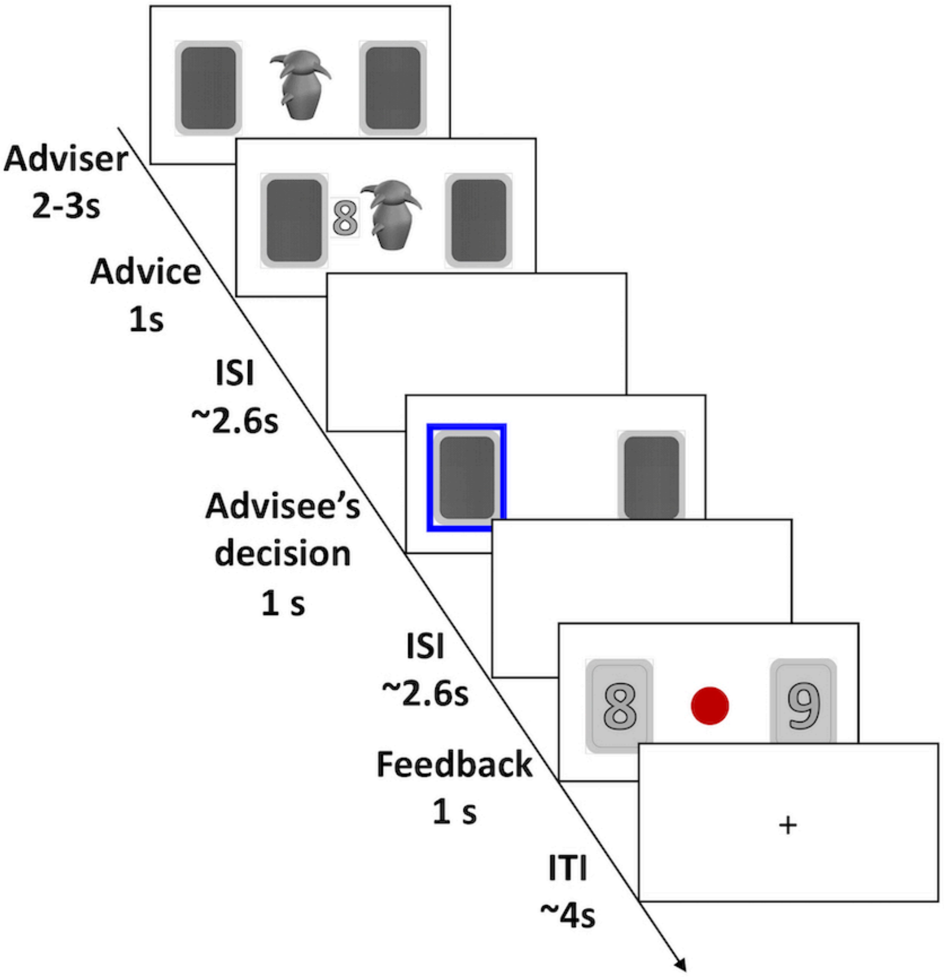


Fig. S1. Timeline of the Take Advice Game.

Timeline of the Take Advice Game (TAG) in the MRI scanner. Before the task, participants had to choose an avatar that represented their identity in the game. They were told that the advisers did the same. Through these avatars, it was assured on one hand that participants knew in each trial who advised them and, on the other, anonymity was guaranteed to each participant. In the adviser presentation phase, participants were told that they see the adviser they were matched with in that trial. In this phase, the adviser was given information about one of the two cards that he or she could communicate to the participants. Thus, in the adviser phase, participants had just to wait that the adviser sends her/his advice. To introduce human-like decisional variability in the communication of the advice, the advice was randomly presented between 2 and 3 seconds after adviser presentation. The advice (presentation time: 1s) was a number between 1 and 9 (expect for 5) either next to the right or the left card. After a variable ISI (range: 2-8s, mean: 2.6s), the two cards were presented one more time and participants were prompted to pick one of the two cards (1s). After another variable ISI, feedback was presented for 1s, revealing the numbers on the cards, based on which participants could judge the honesty of the adviser (social information), and a red or green circle between the cards, representing the participant's performance (nonsocial information about one's payoffs). Finally, an ITI (range: 2-8s, mean: 4s) showing a fixation cross was presented at the end of the trial.
ISI, interstimulus interval; ITI, intertrial interval.

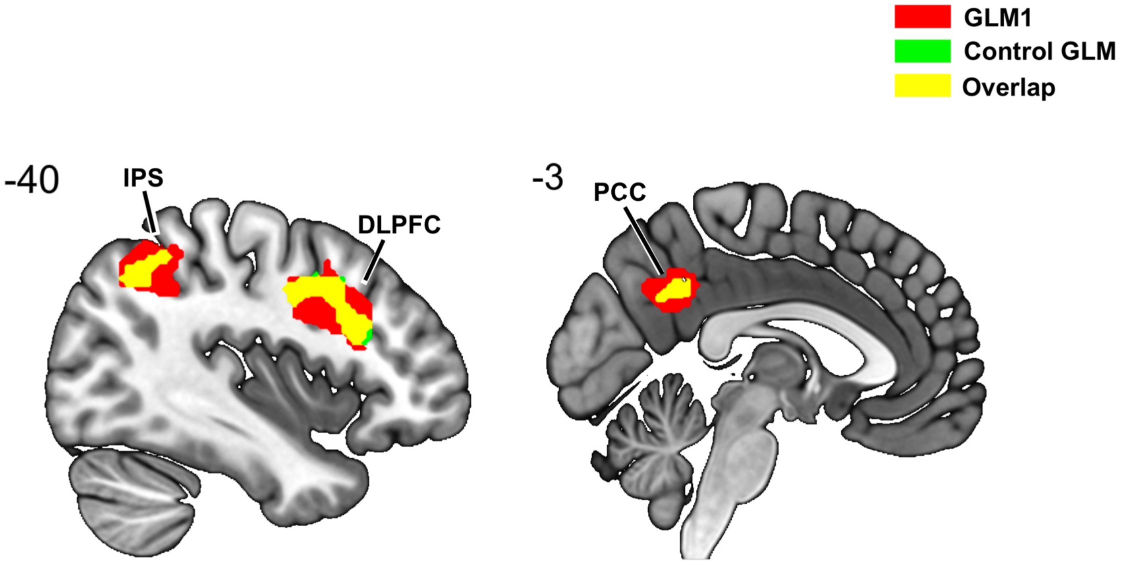


Fig. S2. Control GLM analysis.

Comparison of GLM1 coding for honesty and dishonesty (red), and the control GLM that further controlled for risk and congruency effects (green). The two GLMs yielded similar results. In yellow are the overlaps depicted. GLM, general linear model; IPS, intraparietal sulcus; DLPFC, dorsolateral prefrontal cortex; PCC, posterior cingulate cortex

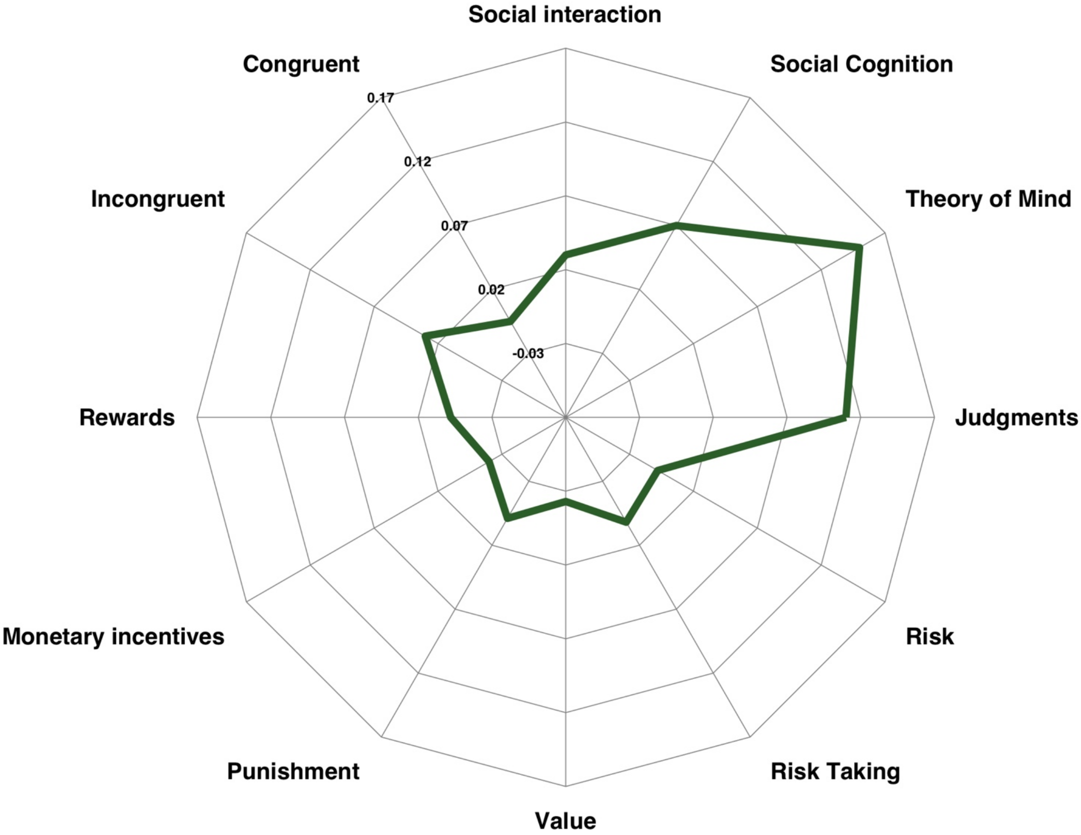


Fig. S3. Meta-analytic functional decoding analysis.

To test the functional specificity of the trustworthiness decoding network, we performed a meta-analytic functional decoding analysis. Using the Neurosynth database, we evaluated the representational similarity between the neural patterns of our trustworthiness map and meta-analytic neural patterns associated with specific terms in the fMRI literature. This way, it was possible to characterize the functional role of our neural patterns by a quantitative comparison with previously observed neural patterns associated with certain cognitive functions. We selected twelve different terms in the social, value, risk and congruency domain. Results show that the neural signatures of the trustworthiness decoding network reveal stronger similarity with neural patterns associated with mentalizing, judgments, social cognition and social interactions than with other cognitive functions. Values on the spider plot (-0.03 — 0.17) represent Pearson's correlation coefficients.

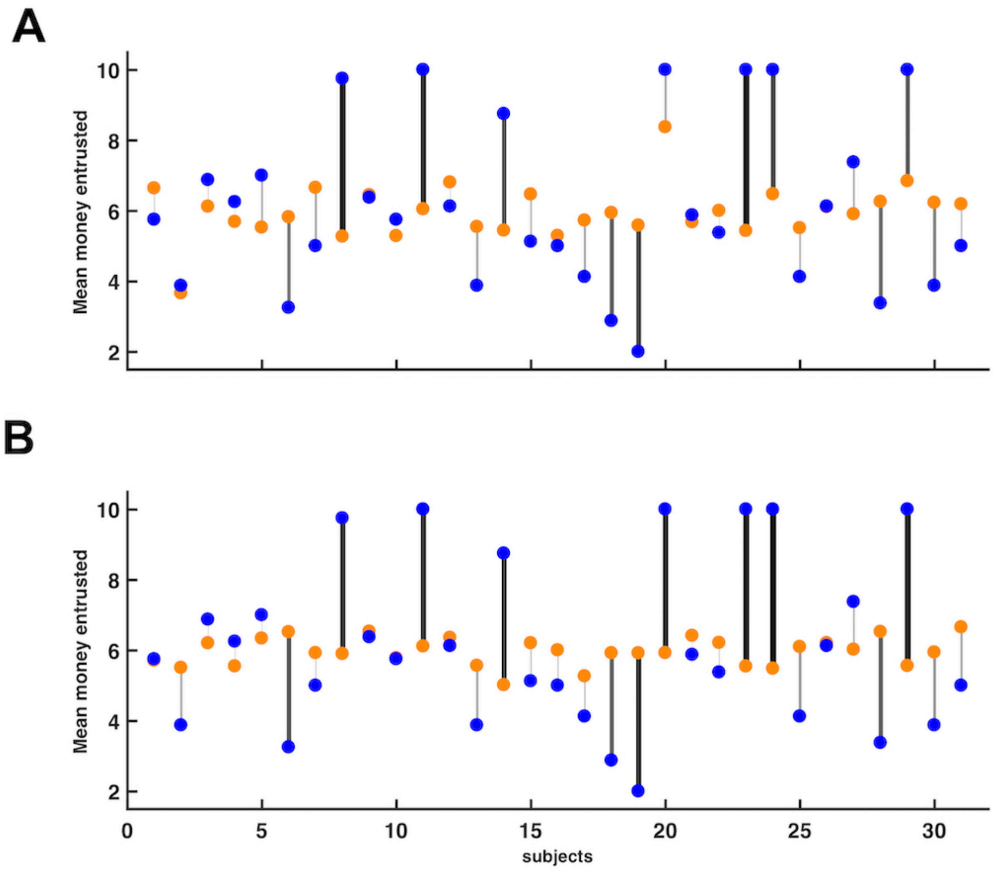


Fig. S4. Model-based predictions of economic trust decisions.

To test whether neural signatures of trustworthiness predict individual trusting behavior, multivariate regression analyses were performed with neural signal from the trustworthiness decoding map (**A**) and the value decoding map (**B**) to predict averaged individual trust in the trust game. Depicted in blue is the individual observed trust, in orange the predicted trust. Black lines connecting blue and orange dots represent model's prediction errors with thicker and darker lines reflecting bigger errors.

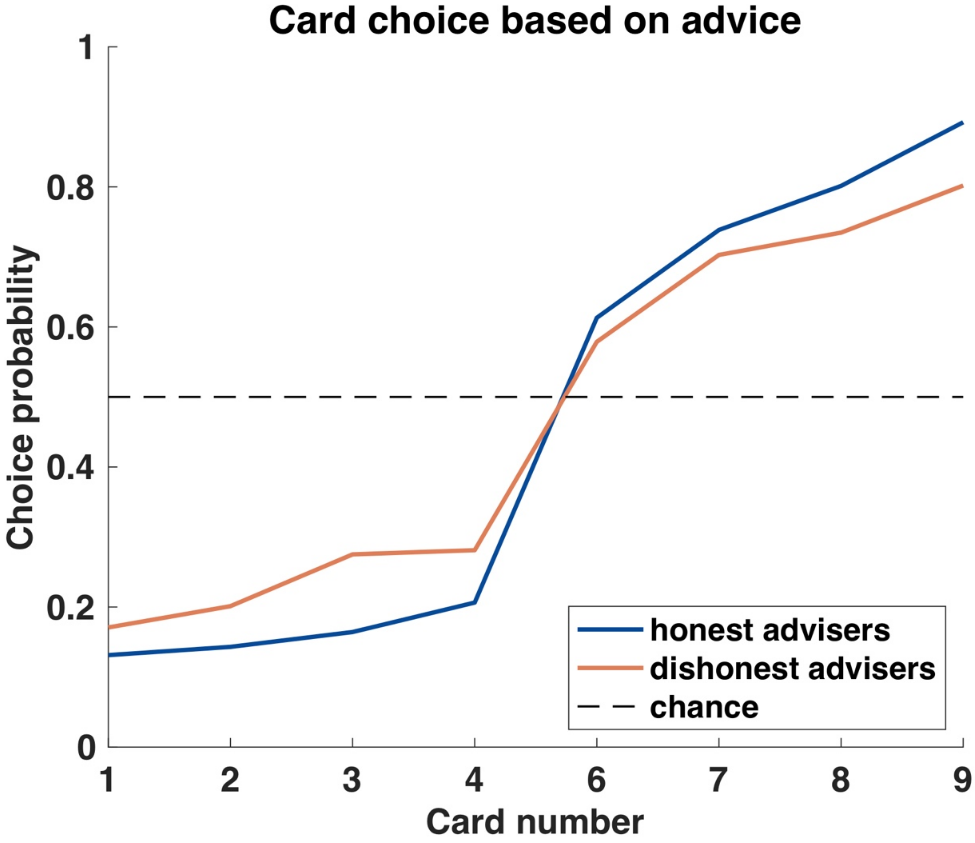


Fig. S5. Card-choice probability analysis.

Advice-taking behavior in the take advice game (TAG) was operationalized as the probability of choosing a card given the informativeness of the advice received. The optimal strategy in the game would be to choose more frequently a card when the adviser communicated that a number bigger than five is on that card but choose the other card when the adviser communicated that a number smaller than five is on that card. Moreover, as we manipulated the advisers' honesty, participants should have employed the optimal card-choice strategy differently for honest and dishonest advisers. In particular, they should have used this strategy more loosely for dishonest advisers compared to honest advisers. Analyses of card choice probabilities confirmed our hypotheses.

Study 3

For copyright reason the original publication is not included in this PDF.

The final version of the article has been accepted on November 22, 2019, and is now in print

in *Journal of Experimental Psychology: General*

with the title provided below.

Please access the published version via the DOI provided below.

Bellucci G, Park S Q, *Honesty biases trustworthiness impressions*, Journal of Experimental Psychology: General

DOI: http://dx.doi.org/10.1037/xge0000730

Study 4

The article in this PDF is a pre-print and not the copy of record.

The article has been submitted and is currently under review for publication

in *Psychopharmacology*, *Springer*, with the title provided below

Bellucci G, Münte T F, Park S Q, *Effects of a dopamine agonist on trusting behaviors in females*, Psychopharmacology, (under review, submitted on July 2019, revision submitted on November 2019)

Title

Effects of a dopamine agonist on behavioral trust in females

Author names and affiliations

Gabriele Bellucci[1,2,3*], Thomas F. Münte[4,5], Soyoung Q. Park[1,2,6,7]

[1] Department of Psychology I, University of Lübeck, 23562 Lübeck, Germany

[2] Department of Education and Psychology, Freie Universität Berlin, Germany

[3] Decision Neuroscience and Nutrition, German Institute of Human Nutrition (DIfE), Potsdam-Rehbruecke, Germany

[4] Department of Neurology, Universitätsklinikum Schleswig-Holstein, 23538 Lübeck, Germany

[5] Department of Psychology II, University of Lübeck, 23562 Lübeck, Germany

[6] Charité-Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Neuroscience Research Center, 10117, Berlin, Germany

[7] Deutsches Zentrum für Diabetes, 85764, Neuherberg, Germany

*Corresponding author

Gabriele Bellucci, MA, MSc
Decision Neuroscience and Nutrition
German Institute of Human Nutrition (DIfE)
Potsdam-Rehbruecke, Germany
gbellucc@gmail.com

Manuscript Information

Number of pages: *15*

Number of figures: *5*

Number of words for Abstract/Introduction/Discussion: *225/675/1169*

Key words: trusting behavior, trust game, investment game, attractiveness, trustworthiness, dopamine, hormonal contraceptive

Abstract

Trust is central to bonding and cooperation. In many situations, individuals need to trust others based exclusively on subjective, first impressions of the other's trustworthiness. Previous studies have shown that trusting behaviors elicit activations in dopaminergic brain regions and that subjective impressions of others can be formed from social information from faces (e.g., facial trustworthiness and attractiveness). However, the effects of dopamine agonists on trusting behaviors based on others' facial trustworthiness are yet unknown. Using a double-blind, placebo-controlled, within-subject design, we here administrated pramipexole (a D2/D3 dopamine agonist) to 28 healthy female participants before playing a one-shot trust game in the role of investor. To induce different trusting behaviors, facial trustworthiness of the partners' face was manipulated with minimal variations in facial attractiveness. Our results show that by minimizing attractiveness information in faces, it is possible to isolate the contribution of facial trustworthiness to behavioral trust. Notably, even though pramipexole did not alter participants' trustworthiness impressions, trusting behavior was significantly impacted by pramipexole intake. Importantly, these pramipexole's effects on impression-based trust were mediated by hormonal contraceptive use. In particular, trust increased in women using hormonal contraceptive but decreased in non-users after pramipexole intake. This study fills an important gap in the experimental literature on trust and its underlying neural mechanisms, pointing to peculiar cognitive and neural dynamics underlying trusting behaviors based on subjective, trustworthiness impressions.

Introduction

Trust and trustworthiness are key features of interpersonal social interactions, as they foster cooperation and facilitate binding and social integration [1,2]. Research on trust has so far provided increasing evidence that trusting behaviors rely on two cognitive mechanisms. On one hand, individuals ground their trust decisions in subjective impressions about the partner's trustworthiness that are formed rapidly and effortlessly [3]. On the other hand, trust can be rooted in dynamically updated beliefs about the partner's trustworthiness based on previous experience with the partner during repeated social interactions [4].

Previous neuroimaging studies have indicated that both impression- and experience-based trust activate dopaminergic brain regions such as the striatum and medial prefrontal cortex [1,5,6]. However, it is an open question whether the engagement of dopamine plays a key role in trust as such, or it rather reflects other cognitive mechanisms laterally associated with trusting behavior. In repeated trusting interactions (e.g., in the multi-round trust game, TG), previous work has indicated a time-dependent shift in the neural patterns underlying experience-based trust with higher-order brain regions mainly engaged at the beginning of a social exchange and dopaminergic brain structures recruited in later stages [1,7]. These results indicate that the engagement of dopaminergic brain regions in experience-based trust might reflect reinforcement learning mechanisms related to belief update about the other's social character, which has recently been suggested also by meta-analyses on trusting behaviors in single and repeated interactions [8,9].

Similarly, the role of dopamine in impression-based trust is yet to be clarified. Previous studies on subjective impressions about others' trustworthiness have largely focused on impressions formed from faces [10]. However, faces entail other types of social information than facial trustworthiness, such as facial attractiveness, which not only guides trusting behaviors [11-13] but also evokes activations in dopaminergic brain regions [14,15]. As facial attractiveness is closely related to facial trustworthiness and, to our knowledge, has not properly controlled for in previous research [10,16,17], the observed neural responses underlying impression-based trust in dopaminergic regions might be as likely evoked by facial trustworthiness as by facial attractiveness.

Thus, given that the engagement of dopaminergic brain structures in repeated trusting interactions might be related to reward anticipation or reinforcement learning processes [13,18] and that the role of dopamine in trust based on facial trustworthiness impressions might be confounded by other types of social information from faces [11,12], the link between trust and the dopaminergic system remains to date an open question.

In this study with a double-blind, within-subject, placebo-controlled design, we investigated for the first time whether administration of a dopamine agonist (DA, i.e., pramipexole) impacts impression-based trust. To limit as much as possible the engagement of other cognitive mechanisms during trust, participants played in the role of investor the one-shot TG and their trust was manipulated by presentation of faces that varied in their facial trustworthiness. As impressions of facial trustworthiness are not affected by dopaminergic modulation [19], we expected that our pharmacological intervention would not impact subjective trustworthiness impressions. Thus, we assumed that our manipulation of behavioral trust would not be confounded by possible effects of pramipexole on trustworthiness impressions. Moreover, we employed the one-shot TG, because it reliably induces impression-based trusting behaviors that are not confounded by learning mechanisms, as participants in the one-shot TG interact only once with their trustees without feedback information about the trustees' behavior. Further, to minimize facial attractiveness confounds, we chose faces that maximally differed on the trustworthiness dimension with minimal variations on the attractiveness one.

Pramipexole, a D2/D3 DA preferentially targeting brain areas in the striatum and medial prefrontal cortex [20,21] was administered. Given this restricted focus of pramipexole's action and given the fact that those target brain regions coincide with trust-related brain areas, pramipexole is a good candidate to study pharmacologically-induced variations of behavioral trust. Further, given sex differences in the dopaminergic system and its pharmacological modulation [22-25], we recruited only female participants. Finally, as social behaviors have previously been shown to differ as a function of hormonal contraceptive use in women [26,27], we also assessed the hormonal contraceptive use in our female sample.

Materials and Methods

Subjects. Thirty participants took part in the experiment. Due to technical problems in data collection, 2 participants had to be excluded leaving a final sample of 28 healthy, female subjects (22.11±2.25 years, mean±SD). All subjects had normal or corrected-to-normal vision. The study was approved by the local Medical Ethics Committee of the University of Lübeck. All subjects provided written consent for participation. We recruited only female participants, due to previous research showing sex differences in dopamine function [22], receptor availability [25] and modulation by pharmacological intervention [23,24]. Fifteen of our participants used hormonal contraceptives (53.4% of our sample). In line with studies showing that the dopaminergic system is modulated by the estrous cycle, this information was entered as regressor into each regression model [28,29]. The testing time of day was held constant (1 p.m.) across sessions for each participant, to control for circadian variability in hormone release [30].

Experimental procedure. Subjects were invited to the lab on two different days with a gap of at least one and maximum of two weeks to participate in a double-blind, within-subject experiment. This gap was scheduled to allow for the dopamine wash-out phase. Upon arrival, subjects randomly received either 0.5 mg of a DA (i.e., pramipexole) or a placebo. To counteract the common nausea effects following pramipexole administration, the drug/placebo intake was accompanied by 10 mg domperidon. Three hours after drug/placebo administration [20,21], participants underwent a functional MRI session to test DA effects on neural dynamics [31]. After the MRI session, participants underwent a battery of tasks at a lab computer, among which there was the one-shot TG (**Fig. 1A**).

....

Insert Figure 1 about here

....

On both sessions, subjects performed multiple rounds of the one-shot TG in the role of investor (each round with a different trustee). In each trial, participants were first presented with a picture of a trustee and decided how much they wanted to share with the depicted person on a scale from 0 (sharing nothing) to 10 monetary units (sharing the entire initial endowment) (**Fig. 1B**). Participants were informed that each monetary unit (MU) corresponded to 30 Cents and that thus the total amount of money they were endowed with was 3€. Further, they were told that every shared amount of money would be tripled, and the trustee had to decide whether to share any portion of the tripled amount of money back. After the TG, participants rated the trustworthiness (**Fig. 1C**) and attractiveness (**Fig. 1D**) of each face on a 7-point Likert-scale. Attractiveness and trustworthiness ratings were randomly presented in two separate blocks.

Experimental procedures on the first and second session were exactly the same. Stimuli were presented using Psychtoolbox 3 (http://psychtoolbox.org) on MALTAB 2016b (https://www.mathworks.com).

Stimulus Material. Trustees' pictures were selected from a dataset of 98 different pictures of faces from participants who participated in previous experiments of the lab and gave written consent that these stimuli can be used in further experiments. These pictures were rated in an on-line survey for attractiveness (1 = really unattractive; 7 = really attractive), trustworthiness (1 = really untrustworthy; 7 = really trustworthy) and facial expression (-3 = negative; 3 = positive) on a 7-point Likert-scale by an independent sample (*N* = 60, 38 females). All faces of the dataset had an emotionally neutral expression (-0.02±0.7). Of all pictures, 12 (6 female) were selected that did not significantly differ on the attractiveness dimension ($F_{(2,9)} = 1.53$; $p = .268$), but varied maximally on the trustworthiness dimension ($F_{(2,9)} = 85.04$; $p < .0001$). We categorized the stimuli in three different categories according to the rating: untrustworthy (3.27±0.16), trustworthy (5.0±0.27) and average trustworthy (4.05±0.10). We confirmed that each of these trustworthiness categories differed significantly from the other (trustworthy vs. untrustworthy: $t_{(1,3)} = 10.69$, $p = .0018$; trustworthy vs. average trustworthy: $t_{(1,3)} = 8.22$, $p = .0038$; averaged trustworthy vs. untrustworthy: $t_{(1,3)} = 10.61$, $p = .0018$).

Analyses. We fitted generalized mixed-effects regression models to the three dependent variables under examination, namely, 1) trusting behavior in the one-shot TG, 2) trustworthiness and 3) attractiveness evaluations of faces, which were collected to check that our manipulation of facial trustworthiness perceptions was successful and independent of facial attractiveness perceptions. The best model was selected among a series of models that vary in their complexity from the simplest model testable given our design and with no interaction effects to increasingly complex models with more interaction effects. The simplest model was a model containing five fixed-effects regressors, namely, a dummy variable coding for treatment (P, pramipexole/placebo), trustworthiness dimensions (Tr, trustworthy faces, average faces and untrustworthy faces), gender of pictures' faces (G, male/female), a dummy variable coding for hormonal contraceptive use (HC, contraceptive users/non-users) and a regressor coding for session order (Sess, first/second session). In each model, random-effects structure was kept maximal with by-subject random intercepts and slopes for each main effect and interaction effect, and with by-item random intercepts [32]. By-item random slopes were further added for the trustworthiness dimensions and gender of pictures to account for the non-independence due to the repeated presentation of these categories across observations [33].

This way, we ended up with a total of 18 mixed-effects regression models that were separately applied to each dependent variable of this study, i.e., trusting behavior in the TG, trustworthiness rating and attractiveness ratings (**Tab. S1**). Model selection was based on a model comparison approach using the Akaike Information Criterion (AIC), which estimates the goodness of fit of a model based on its likelihood and complexity (by penalizing for the number of parameters to estimate). As expected, different models were best explaining participants' behavior and subjective impressions. The winning model for trusting behavior (T) was a model with one interaction effect between treatment and hormonal contraceptive use, as follows:

$$T_t = \beta_0 + \beta_1 P_t + \beta_2 G_t + \beta_3 Tr_t + \beta_4 HC_t + \beta_5 Sess_t + \beta_6 P * HC_t. \qquad (1)$$

The winning model for trustworthiness impressions (TI) was a model with one interaction effect between gender and trustworthiness dimensions, as follows:

$$TI_t = \beta_0 + \beta_1 P_t + \beta_2 G_t + \beta_3 Tr_t + \beta_4 HC_t + \beta_5 Sess_t + \beta_6 G * Tr_t. \qquad (2)$$

Finally, the winning model for attractiveness impressions (AI) was a model with one interaction effect between hormonal contraceptive use and trustworthiness dimensions, as follows:

$$AI_t = \beta_0 + \beta_1 P_t + \beta_2 G_t + \beta_3 Tr_t + \beta_4 HC_t + \beta_5 Sess_t + \beta_6 HC * Tr_t. \qquad (3)$$

To test the degree to which trustworthiness information influenced participants' trust decisions independently of attractiveness information, we ran a linear regression analysis on the individual level with mean trustworthiness and attractiveness ratings as predictor of average trust in the TG. All analyses were run in MATLAB 2016b. For mixed-effects regression models, the function *fitglme* was used. For linear regression models the function *fitlm* was used.

Results

Subjective impressions impact on trusting behavior. Results from the TG show that participants, in the role of investor, trusted their partners based on subjective impressions about the partner formed on the basis of the partner's facial trustworthiness. Thus, trustworthy-looking trustees were entrusted with more money in the TG (β = 0.60, standard error (SE) = 0.13, 95% confidence interval (CI) = [0.35, 0.85], p < .00001; **Fig. 2A & Tab. S2**). Further, participants' trustworthiness perceptions were also significantly different between male and female trustees (β = 0.34, SE = 0.13, 95% CI = [0.07, 0.60], p = .013; **Tab. S3**). In particular, trustworthy-looking females were perceived as more trustworthy than males, while untrustworthy-looking males were perceived as more trustworthy than females (**Fig. 3A**). On the contrary, even though females were perceived as slightly more attractive than males (β = 0.30, SE = 0.15, 95% CI = [0.008, 0.60], p = .044; **Tab. S4**), attractiveness perceptions did not differ across trustworthiness dimensions (β = -0.02, SE = 0.13, 95% CI = [-0.27, 0.24], p = .904; **Tab. S4**), thereby validating our stimulus material (**Fig. 2B-C & Fig. 3B**). Finally, an individual regression analysis with both trustworthiness and attractiveness as predictors revealed that trustworthiness significantly predicted trust (β = 1.85, SE = 0.71, 95% CI = [0.39, 3.31], p = .015), but not attractiveness (β = 0.99, SE = 0.70, 95% CI = [-0.45, 2.42], p = .170). These results suggest that in single interactions, trusting behaviors can be guided exclusively by subjective impressions about the other's trustworthiness independently of other social information, such as facial attractiveness.

..

Insert Figure 2 & 3 about here

....

Pramipexole modulation of trusting behavior. Further, pramipexole administration appeared to influence behavioral trust (**Fig. 4A**). In particular, we observed reduced trusting behavior in the pramipexole session compared to the placebo session (β = -0.98, SE = 0.38, 95% CI = [-1.74, -0.23], p = .010; **Tab. S2**), that is, participants entrusted overall less money with their partners after pramipexole administration. Interestingly, this decrease in trusting behavior after pramipexole administration was not reflected by any modulation of subjective impressions about the partner's trustworthiness (β = -0.13, SE = 0.10, 95% CI = [-0.32, 0.06], p = .175; **Tab. S3**) or attractiveness (β = -0.006, SE = 0.12, 95% CI = [-0.23, 0.22], p = .959; **Tab. S4**). Notably, this modulation of trusting behavior was mediated by hormonal contraceptive use (β = 1.32, SE = 0.49, 95% CI = [0.36, 2.29], p = .007; **Tab. S2**). In particular, pramipexole increased trusting behavior in participants who used hormonal contraceptives, whereas it decreased trusting behavior in those who did not (**Fig. 4B**). No contraceptive use effects were found for subjective impressions about the partner's trustworthiness (β = -0.002, SE = 0.16, 95% CI = [-0.32, 0.31], p = .990; **Tab. S3**). On the contrary, we observed an effect of contraceptive use on subjective attractiveness impressions, which interacted with facial trustworthiness dimensions (β = 0.34, SE = 0.16, 95% CI = [0.02, 0.66], p = .038; **Tab. S4**). In particular, trustworthy-looking partners were perceived as more attractive by participants who used hormonal contraceptives (**Fig. 5**). These results indicate a complex interaction between the dopaminergic system and trusting behavior that is modulated by widely used common contraceptive methods in females.

....

Insert Figure 4 & 5 about here

....

Discussion

In this study, we investigated for the first time the DA effects on behavioral trust. We first isolated the contribution of trustworthiness impressions on trusting behavior. We then observed how pharmacological modulation of a DA (i.e., pramipexole) impacts impression-based trust. In particular, administration of pramipexole decreased trust in unknown others. Notably, this decrease of trust was particularly prominent in women who did not use hormonal contraceptives. On the contrary, in contraceptive users, pramipexole intake increased trust in unknown others.

We first replicated previous findings that sensitivity to facial trustworthiness is not affected by dopamine modulation. In particular, a previous study has shown that dopamine depletion does not alter trustworthiness perceptions of others despite successful neural modulation of the brain's reward system [19]. Analogously, although we previously reported that pramipexole successfully modulated neural activity in dopaminergic brain regions in the same sample [31], no significant modulations of facial trustworthiness were observed after pramipexole administration. Hence, results from both our and previous studies suggest that other neural dynamics and brain regions than dopaminergic areas likely underlie first subjective impressions of others' social character.

Moreover, we disentangled for the first time the contribution of trustworthiness information to trusting behavior from attractiveness evaluations. In particular, our results suggest that by reducing attractiveness information in faces, it is possible to isolate the influence of facial trustworthiness on trust. An interesting hypothesis for future studies would be to investigate whether different types of social information from faces make individuals trust others for different motives. For instance, trustworthy-looking individuals may be trusted because they are likely to be good cooperators [34], whereas, attractive others may be trusted simply due to a "beauty premium" [35]. Hence, trust in an unknown other might rely on the ability to form impressions about the other's social character (e.g., trustworthiness) to infer their cooperative intentions and overcome betrayal aversion [36-38].

Finally, our results indicate a successful impact of pramipexole on impression-based trust. In particular, pramipexole reduced trust in unknown others in female participants who did not use hormonal contraceptives. Consistently with the absence of any DA impact on trustworthiness impressions, such effects on trust were observed across facial trustworthiness dimensions. The underlying dynamics of these DA effects on trust may be multiple. The decrease of impression-based trust following pramipexole administration may be due to the effects of pramipexole as DA on dopaminergic brain structures. Such DA effects may reduce participants' sensitivity to social contact and feedback. Dopamine has previously shown to mediate socially-relevant behaviors such as approach strategies and mate preferences [39-41]. As trust signals the willingness to establish a potentially long-lasting relationship advantageous for future cooperation [2,42,43], administration of a DA drug might saturate the human need to belong, limiting one's willingness to relate to and connect with others [44].

However, the same decrease of impression-based trust may be explained by a dopamine antagonist effect of pramipexole as well. In particular, pramipexole acts on both D2 and D3 autoreceptors [20,21,45-47]. D3 autoreceptor activation has been observed to inhibit the reward-related phasic firing of dopaminergic neurons [48]. Such inhibition of dopamine activity has further been hypothesized to be reflected by reduced functional MRI activations in brain structures rich in D3 autoreceptors [21]. A dopamine antagonist effect on dopaminergic brain structures might also have a negative impact on trust. For instance, mental diseases attributed to dopamine dysfunction are characterized by social impairments and social avoidance [49-51]. Hence, pramipexole administration might have silenced the ability to form and maintain satisfying social relationships, reducing an individual's willingness to trust.

Importantly, the effects of pramipexole administration on trust in women using hormonal contraceptives were reversed. That is, pramipexole increased impression-based trust in unknown others in female participants who used hormonal contraceptives at the time of the study. Previous work has shown that hormonal contraceptive use alters both neural dynamics and behavior in women. Neuroimaging studies have provided preliminary evidence of altered neural reward processing [52] and differing functional connectivity in higher-order brain areas pivotal to social cognition [53] in women using hormonal contraception as compared to naturally cycling women. Behavioral studies have pinpointed differences in women using hormonal contraceptives in mate-choice behaviors and attraction to other-sex features, such as male face, voice and odor [27,54,55]. Notably, contraceptive use contributes to romantic relationship outcome, with important implications for personal satisfaction and quality of life [56].

In particular, contraceptive use has been observed to shift women's mating preferences to less masculine features (indicative of low testosterone levels), so that women using hormonal contraceptive prioritize traits such as wealth and intelligence in mate choice [57,58]. Since contraceptive use enhances preferences for safety and future security, women using hormonal contraceptive might also exhibit stronger attraction to trustworthy partners. Indeed, in our sample, women using hormonal contraceptive perceived trustworthy faces as more attractive, although facial attractiveness was manipulated to be comparable across trustworthiness dimensions. Pramipexole administration might hence boost such preferences in women using hormonal contraceptive, leading to a more pronounced trusting behavior.

A couple of limitations have to be addressed. First, due to gender differences in pharmacological interventions using dopamine drugs [23,24], we tried to avoid gender variability in the DA effects on trust by limiting our sample to female participants. However, this choice reduces the generalizability of our results. Thus, future studies need to replicate our results in a bigger and more heterogenous sample. Second, interpretation of our results is limited by the absence of data on contraceptive type used by our female sample. Different types of hormonal contraceptives may interact in different ways with pharmacological modulations of brain dynamics [53]. Hence, future studies are needed to replicate our findings controlling for variables that might confound a pharmacological intervention. Third, a previous study has found weak, but significant evidence on the effects of endogenous sex hormones on interpersonal trust during the preovulatory phase in a sample of 12 naturally cycling women [30]. As we could not control in this study for menstrual cycle phases, future studies are needed to check whether our results hold also after controlling for sex hormones in naturally cycling women. Finally, the use of more suitable techniques such as positron emission tomography might help provide better insights into the neural relationships between dopamine and trust, for instance, by further collecting data on the binding profile of the administrated drug.

In conclusion, we provided first pharmacological evidence on the effects of a DA drug on impression-based trust. By controlling for variables that may have confounded results of previous studies, we demonstrated that facial trustworthiness uniquely affects trusting behavior and that pramipexole alters behavioral trust across facial trustworthiness dimensions with no impact on subjective, trustworthiness impressions. Notably, DA effects on trust were mediated by hormonal contraceptive use. This finding indicates complex neural dynamics underpinning social behaviors, which likely involve the interplay of different neuromodulators and brain systems. Thus, these findings importantly contribute to fill an epistemological gap in the current literature, potentially directing the current research toward a new path of investigations aiming at unearthing the complex cognitive and neural dynamics that bring about social behaviors.

Acknowledgments

This work was supported by the German Research Foundation Grants INST 392/125-1 (SFB TRR (07)) and PA 2682/1-1 and PA 2683/2-1 (FOR 2698/1, TP3) to S.Q.P. and SFB TR134 C1 to T.F.M. This study was funded by the grant from the German Ministry of Education and Research (BMBF) and the State of Brandenburg (to S.Q.P.; DZD, FKZ grant 82DZD00302).

Author Contribution

G.B., T.F.M. and S.Q.P. designed the study; G.B. programmed the task, collected and analyzed the data; G.B. and S.Q.P. wrote the manuscript with edits from T.F.M.; all authors agreed on the last draft of the paper.

Conflict of Interest

None declared.

Captions to Figures

Fig. 1. Procedures. A. Schematic representation of the one-round trust game (TG). In the one-round TG, the investor received an initial endowment that she could share with a second player, i.e., the trustee. If the investor decided to trust, the amount was tripled and passed on to the trustee who could decide whether to reciprocate by sending back part of the tripled amount received. **B.** In the TG, participants played one round with each trustee whose picture was presented on the screen. Participants made their decisions on a scale from 0 (sharing nothing) to 10 monetary units (sharing the entire initial endowment). No time limits were given for the decisions and presentation of trustees' pictures was separated by an interstimulus interval (ISI) of 0.5s. Finally, after the TG, participants rated the trustworthiness (**C.**) and attractiveness of the trustees (**D.**). Trustworthiness and attractiveness ratings were presented in randomized order.

Fig. 2. Trust Game and Ratings Results. Facial trustworthiness significantly impacted trusting behavior in the trust game (**A.**). Moreover, participants' trustworthiness ratings (**C.**) indicate that our manipulation of their subjective, trustworthiness impressions was successful independently of their attractiveness impressions (**B.**).

***p < .001; ns., nonsignificant.

Fig. 3. Ratings Results. Trustworthiness ratings show that participants' subjective impressions about the other's trustworthiness interacted with the face's gender, with trustworthy-looking female faces being perceived as more trustworthy than trustworthy-looking male faces and the opposite effect for untrustworthy-looking faces (**A.**). Female faces were further perceived as more attractive than male faces (**B.**).

* p < .05

Fig. 4. Pramipexole's Effects on Trust. Across facial trustworthiness, participants trusted the partner significantly less after pramipexole intake (**A.**). Moreover, such dopamine agonist effects on behavioral trust interacted with participants' contraceptive use, with increased trust after pramipexole intake in women using hormonal contraceptives and decreased trust in non-users (**B.**).

* *p* < .05

Fig. 5. Interaction between Attractiveness Impressions and Contraceptive Use. Although attractiveness impressions did not differ across facial trustworthiness, women using hormonal contraceptive perceived trustworthy-looking faces are as more attractive than non-users. The opposite effect was found for untrustworthy-looking faces.

* *p* < .05

References

1 Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M, et al. Neural correlates of trust. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(50):20084-9.

2 Hillebrandt H, Sebastian C, Blakemore SJ. Experimentally induced social inclusion influences behavior on trust games. Cogn Neurosci. 2011;2(1):27-33.

3 Todorov A, Pakrashi M, Oosterhof NN. Evaluating Faces on Trustworthiness After Minimal Time Exposure. Social Cognition. 2009;27(6):813-33.

4 Hula A, Vilares I, Lohrenz T, Dayan P, Montague PR. A model of risk and mental state shifts during social interaction. PLoS Comput Biol. 2018;14(2):e1005935.

5 Lauharatanahirun N, Christopoulos GI, King-Casas B. Neural computations underlying social risk sensitivity. Frontiers in human neuroscience. 2012;6:213.

6 Stanley DA, Sokol-Hessner P, Fareri DS, Perino MT, Delgado MR, Banaji MR, et al. Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2012;367(1589):744-53.

7 Bellucci G, Hahn T, Deshpande G, Krueger F. Functional connectivity of specific resting-state networks predicts trust and reciprocity in the trust game. Cogn Affect Behav Neurosci. 2019;19(1):165-76.

8 Bellucci G, Chernyak SV, Goodyear K, Eickhoff SB, Krueger F. Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. Hum Brain Mapp. 2017;38(3):1233-48.

9 Bellucci G, Feng C, Camilleri J, Eickhoff SB, Krueger F. The role of the anterior insula in social norm compliance and enforcement: Evidence from coordinate-based and functional connectivity meta-analyses. Neurosci Biobehav Rev. 2018;92:378-89.

10 Todorov A, Olivola CY, Dotsch R, Mende-Siedlecki P. Social attributions from faces: determinants, consequences, accuracy, and functional significance. Annu Rev Psychol. 2015;66:519-45.

11 Wilson RK, Eckel CC. Judging a book by its cover: Beauty and expectations in the trust game. Political Research Quarterly. 2006;59(2):189-202.

12 Stirrat M, Perrett DI. Valid facial cues to cooperation and trust: male facial width and trustworthiness. Psychological science. 2010;21(3):349-54.

13 Chang LJ, Doll BB, van 't Wout M, Frank MJ, Sanfey AG. Seeing is believing: trustworthiness as a dynamic belief. Cogn Psychol. 2010;61(2):87-105.

14 Aharon I, Etcoff N, Ariely D, Chabris CF, O'Connor E, Breiter HC. Beautiful faces have variable reward value: fMRI and behavioral evidence. Neuron. 2001;32(3):537-51.

15 Winston JS, Strange BA, O'Doherty J, Dolan RJ. Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nature neuroscience. 2002;5(3):277-83.

16 Sofer C, Dotsch R, Wigboldus DH, Todorov A. What is typical is good: the influence of face typicality on perceived trustworthiness. Psychological science. 2015;26(1):39-47.

17 Todorov A. Evaluating faces on trustworthiness: an extension of systems for recognition of emotions signaling approach/avoidance behaviors. Ann N Y Acad Sci. 2008;1124:208-24.

18 van 't Wout M, Sanfey AG. Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. Cognition. 2008;108(3):796-803.

19 Zebrowitz LA, Boshyan J, Ward N, Hanlin L, Wolf JM, Hadjikhani N. Dietary dopamine depletion blunts reward network sensitivity to face trustworthiness. J Psychopharmacol. 2018;32(9):965-78.

20 Ishibashi K, Ishii K, Oda K, Mizusawa H, Ishiwata K. Binding of pramipexole to extrastriatal dopamine D2/D3 receptors in the human brain: a positron emission tomography study using 11C-FLB 457. PLoS One. 2011;6(3):e17723.

21 Riba J, Kramer UM, Heldmann M, Richter S, Munte TF. Dopamine agonist increases risk taking but blunts reward-related brain activity. PLoS One. 2008;3(6):e2479.

22 Castner SA, Xiao L, Becker JB. Sex differences in striatal dopamine: in vivo microdialysis and behavioral studies. Brain Research. 1993;610(1):127-34.

23 Soutschek A, Burke CJ, Raja Beharelle A, Schreiber R, Weber SC, Karipidis II, et al. The dopaminergic reward system underpins gender differences in social preferences. Nature Human Behaviour. 2017;1(11):819-27.

24 Munro CA, McCaul ME, Wong DF, Oswald LM, Zhou Y, Brasic J, et al. Sex differences in striatal dopamine release in healthy adults. Biol Psychiatry. 2006;59(10):966-74.

25 Pohjalainen T, Rinne JO, Nagren K, Syvalahti E, Hietala J. Sex differences in the striatal dopamine D2 receptor binding characteristics in vivo. Am J Psychiatry. 1998;155(6):768-73.

26 Birnbaum GE, Zholtack K, Mizrahi M, Ein-Dor T. The Bitter Pill: Cessation of Oral Contraceptives Enhances the Appeal of Alternative Mates. Evolutionary Psychological Science. 2019.

27 Alvergne A, Lummaa V. Does the contraceptive pill alter mate choice in humans? Trends in Ecology & Evolution. 2010;25(3):171-79.

28 Becker JB, Perry AN, Westenbroek C. Sex differences in the neural mechanisms mediating addiction: a new synthesis and hypothesis. Biol Sex Differ. 2012;3(1):14.

29 Jacobs E, D'Esposito M. Estrogen shapes dopamine-dependent cognitive processes: implications for women's health. J Neurosci. 2011;31(14):5286-93.

30 Ball A, Wolf CC, Ocklenburg S, Herrmann BL, Pinnow M, Brune M, et al. Variability in ratings of trustworthiness across the menstrual cycle. Biol Psychol. 2013;93(1):52-7.

31 Bellucci G, Münte TF, Park SQ. Resting-state dynamics as a neuromarker of dopamine administration in healthy female adults. J Psychopharmacol. 2019:269881119855983.

32 Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang. 2013;68(3).

33 Clark HH. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior. 1973;12:335-59.

34 Dunbar RIM. Gossip in evolutionary perspective. Review of General Psychology. 2004;8(2):100-10.

35 Mobius MM, Rosenblat TS. Why Beauty Matters. American Economic Review. 2006;96(1):222-35.

36 Frith CD, Frith U. Interacting Minds--A Biological Basis. Science. 1999;286(5445):1692-95.

37 Aimone JA, Houser D. Harnessing the benefits of betrayal aversion. Journal of Economic Behavior & Organization. 2013;89:1-8.

38 Bohnet I, Zeckhauser R. Trust, risk and betrayal. Journal of Economic Behavior & Organization. 2004;55(4):467-84.

39 Aragona BJ, Liu Y, Yu YJ, Curtis JT, Detwiler JM, Insel TR, et al. Nucleus accumbens dopamine differentially mediates the formation and maintenance of monogamous pair bonds. Nature neuroscience. 2006;9(1):133-9.

40 Fisher H, Aron A, Brown LL. Romantic love: an fMRI study of a neural mechanism for mate choice. J Comp Neurol. 2005;493(1):58-62.

41 Enter D, Colzato LS, Roelofs K. Dopamine transporter polymorphisms affect social approach-avoidance tendencies. Genes Brain Behav. 2012;11(6):671-6.

42 Balliet D, Van Lange PA. Trust, Punishment, and Cooperation Across 18 Societies: A Meta-Analysis. Perspect Psychol Sci. 2013;8(4):363-79.

43 Romano A, Balliet D, Yamagishi T, Liu JH. Parochial trust and cooperation across 17 societies. Proceedings of the National Academy of Sciences of the United States of America. 2017.

44 Baumeister RF, Leary MR. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull. 1995;117(3):497-529.

45 Hall H, Halldin C, Dijkstra D, Wikström H, Wise LD, Pugsley TA, et al. Autoradiographic localisation of D 3 -dopamine receptors in the human brain using the selective D 3 -dopamine receptor agonist (+)-[3 H]PD 128907. Psychopharmacology. 1996;128(3):240-47.

46 Gurevich E, Joyce JN. Distribution of Dopamine D3 Receptor Expressing Neurons in the Human Forebrain Comparison with D2 Receptor Expressing Neurons. Neuropsychopharmacology.

1999;20(1):60-80.

47 Murray AM, Ryoo HL, Gurevich E, Joyce JN. Localization of dopamine D3 receptors to mesolimbic and D2 receptors to mesostriatal regions of human forebrain. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(23):11271-5.

48 Sokoloff P, Diaz J, Le Foll B, Guillin O, Leriche L, Bezard E, et al. The Dopamine D3 Receptor: A Therapeutic Target for the Treatment of Neuropsychiatric Disorders. CNS & Neurological Disorders - Drug Targets. 2006;5(1):25-43.

49 Fernandez-Theoduloz G, Paz V, Nicolaisen-Sobesky E, Perez A, Buunk AP, Cabana A, et al. Social avoidance in depression: A study using a social decision-making task. J Abnorm Psychol. 2019;128(3):234-44.

50 Cacioppo JT, Norris CJ, Decety J, Monteleone G, Nusbaum H. In the eye of the beholder: individual differences in perceived social isolation predict regional brain activation to social stimuli. J Cogn Neurosci. 2009;21(1):83-92.

51 Caceda R, Moskovciak T, Prendes-Alvarez S, Wojas J, Engel A, Wilker SH, et al. Gender-specific effects of depression and suicidal ideation in prosocial behaviors. PloS one. 2014;9(9):e108733.

52 Bonenberger M, Groschwitz RC, Kumpfmueller D, Groen G, Plener PL, Abler B. It's all about money: oral contraception alters neural reward processing. Neuroreport. 2013;24(17):951-5.

53 Petersen N, Kilpatrick LA, Goharzad A, Cahill L. Oral contraceptive pill use and menstrual cycle phase are associated with altered resting state functional connectivity. Neuroimage. 2014;90:24-32.

54 Roberts SC, Gosling LM, Carter V, Petrie M. MHC-correlated odour preferences in humans and the use of oral contraceptives. Proc R Soc B. 2008;275:2715-22.

55 Feinberg DR, DeBruine LM, Jones BC, Little AC. Correlated preferences for men's facial and vocal masculinity. Evolution and Human Behavior. 2008;29(4):233-41.

56 Roberts SC, Klapilova K, Little AC, Burriss RP, Jones BC, DeBruine LM, et al. Relationship satisfaction and outcome in women who meet their partner while using oral contraception. Proc Biol Sci. 2012;279(1732):1430-6.

57 Little AC, Jones BC, Penton-Voak IS, Burt DM, Perrett DI. Partnership status and the temporal context of relationships influence human female preferences for sexual dimorphism in male face shape. Proc Biol Sci. 2002;269(1496):1095-100.

58 Gangestad SW, Garver-Apgar CE, Simpson JA, Cousins AJ. Changes in women's mate preferences across the ovulatory cycle. J Pers Soc Psychol. 2007;92(1):151-63.

Fig. 1

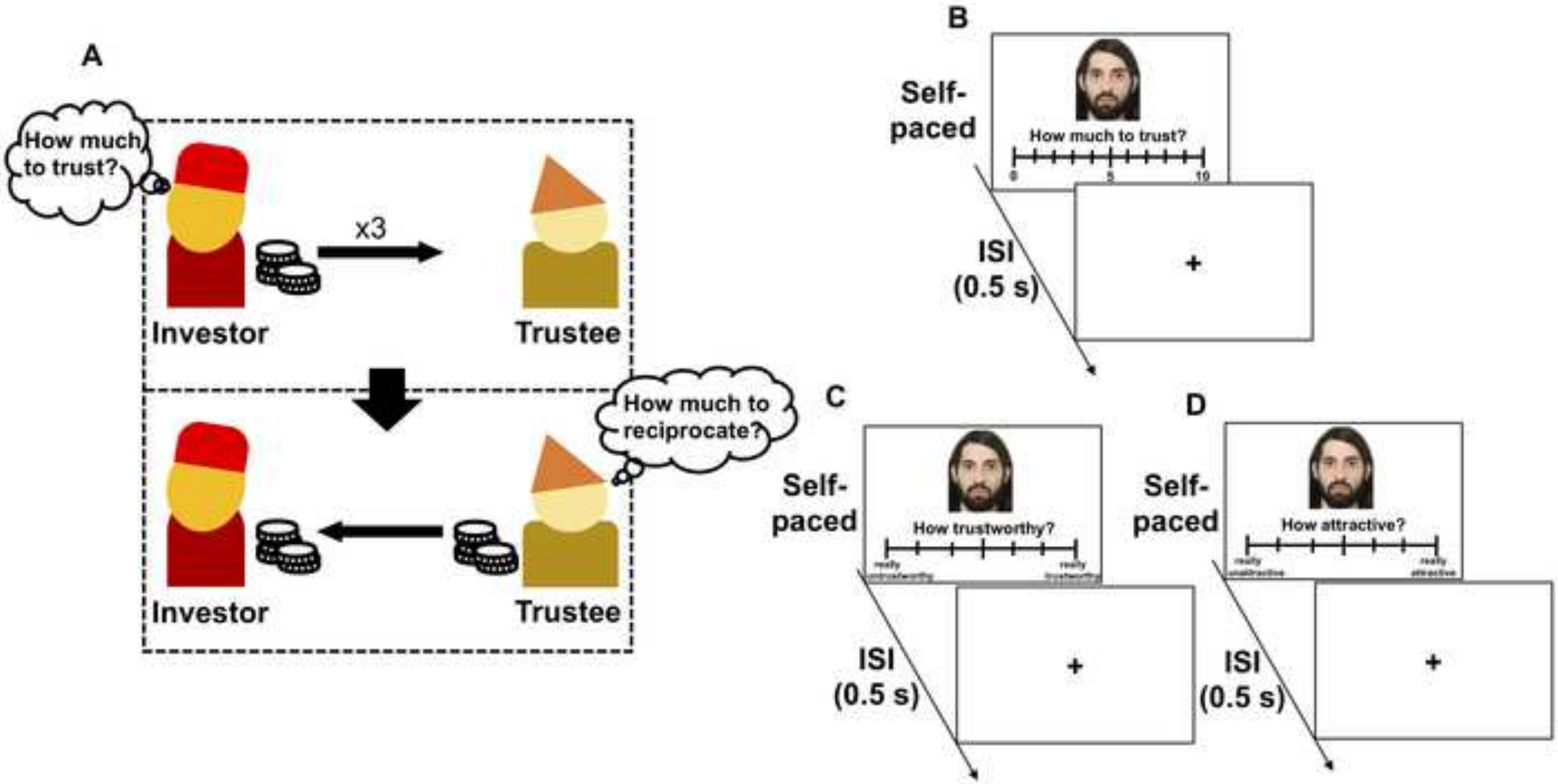

A
How much to trust?
x3
Investor
Trustee
How much to reciprocate?
Investor
Trustee
B
Self-paced
How much to trust?
0
5
10
ISI (0.5 s)
+
C
Self-paced
How trustworthy?
really untrustworthy
really trustworthy
ISI (0.5 s)
+
D
Self-paced
How attractive?
really unattractive
really attractive
ISI (0.5 s)
+

Fig. 2

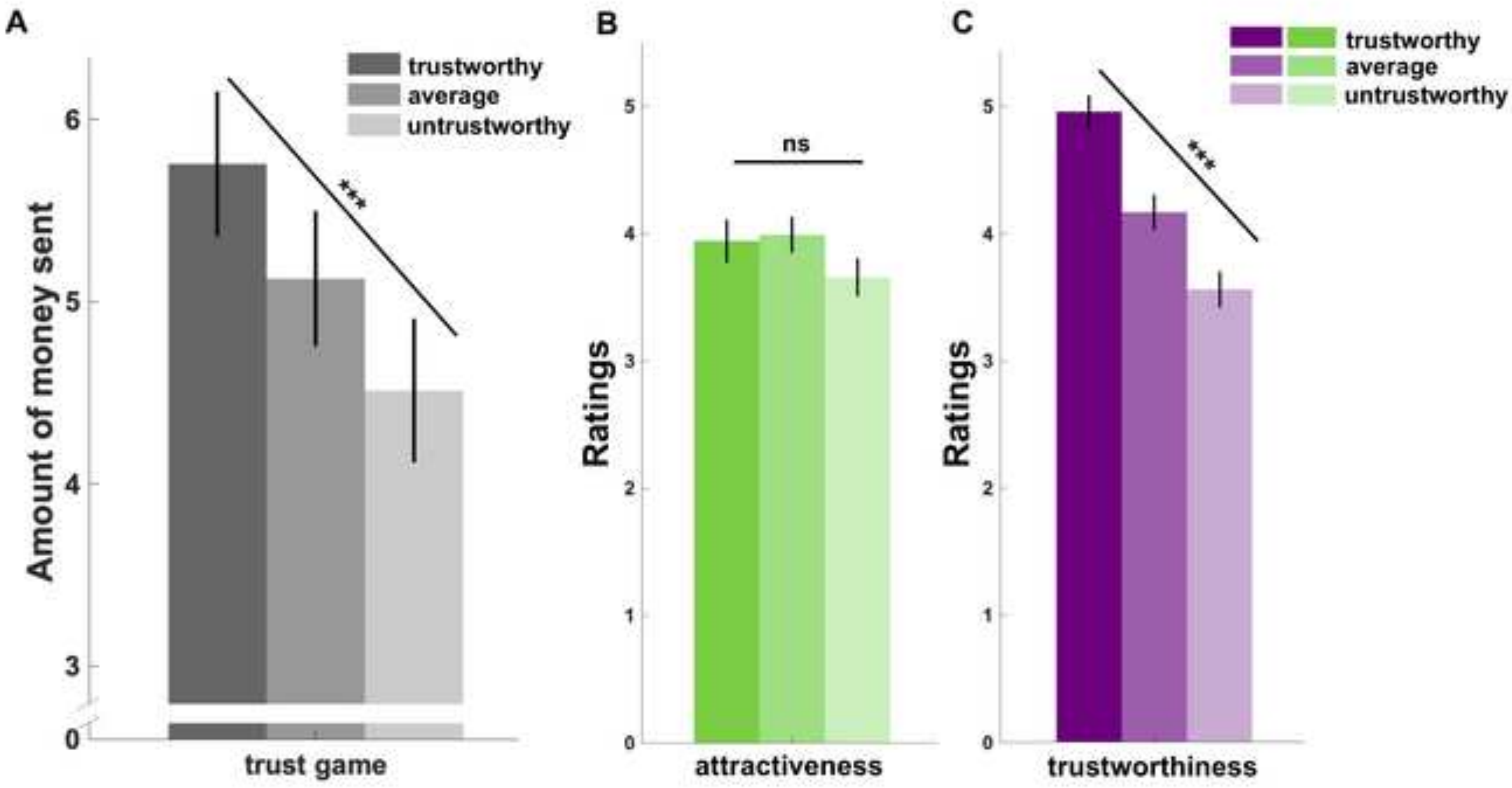

A
Amount of money sent
6
5
4
3
0
trust game
trustworthy
average
untrustworthy
***
B
Ratings
5
4
3
2
1
0
ns
attractiveness
C
Ratings
5
4
3
2
1
0
trustworthy
average
untrustworthy
***
trustworthiness

Fig. 3

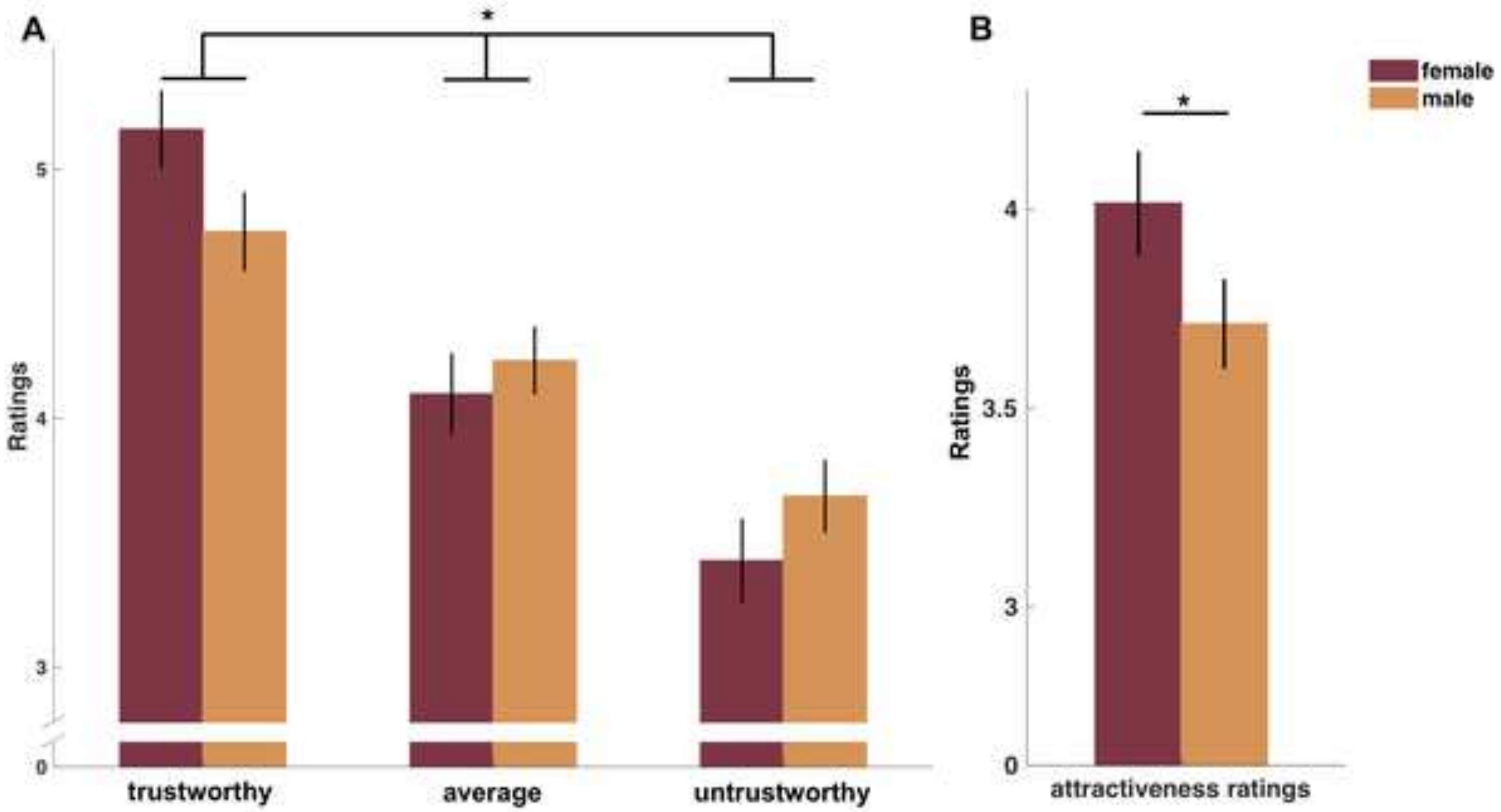

A
*
5
4
3
0
Ratings
trustworthy
average
untrustworthy
B
*
4
3.5
3
0
Ratings
attractiveness ratings
female
male

Fig. 4

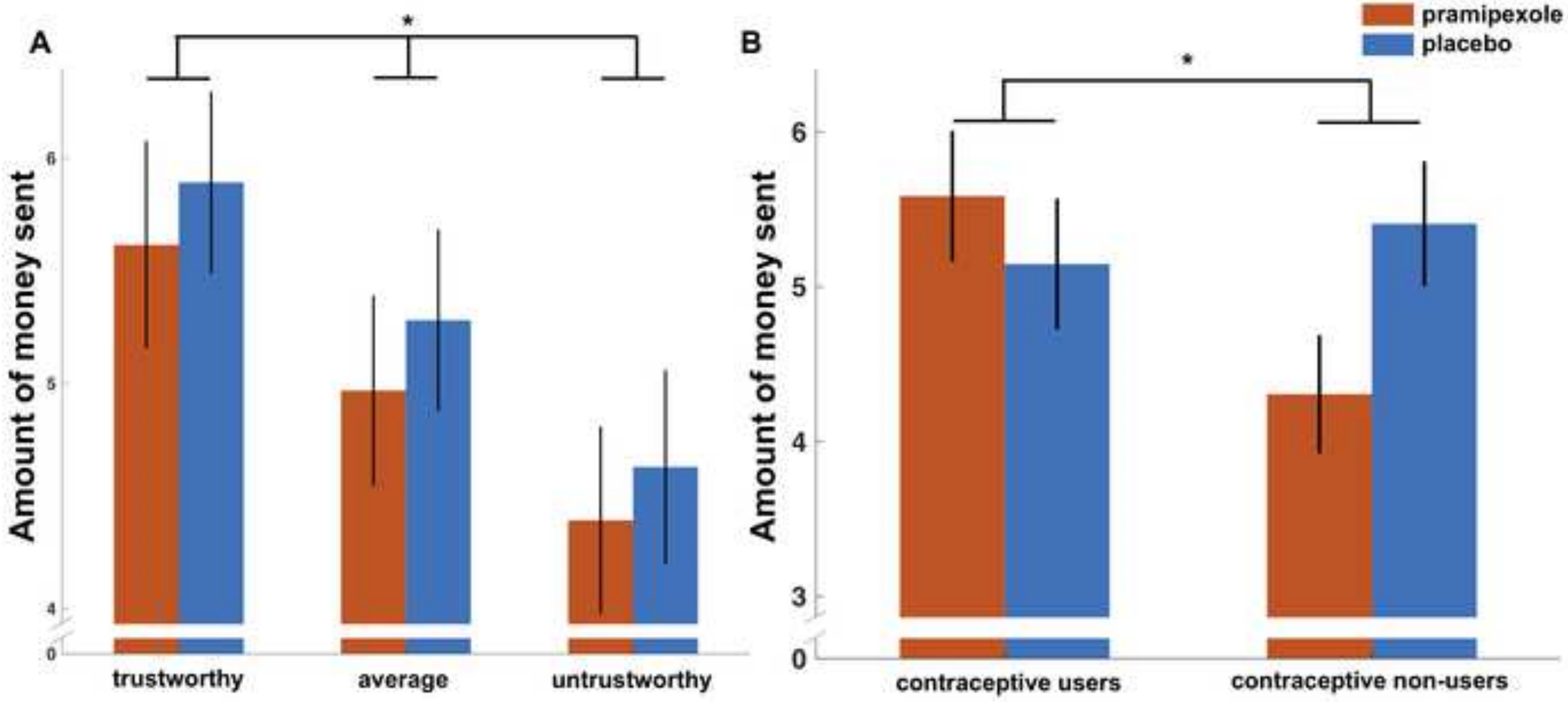
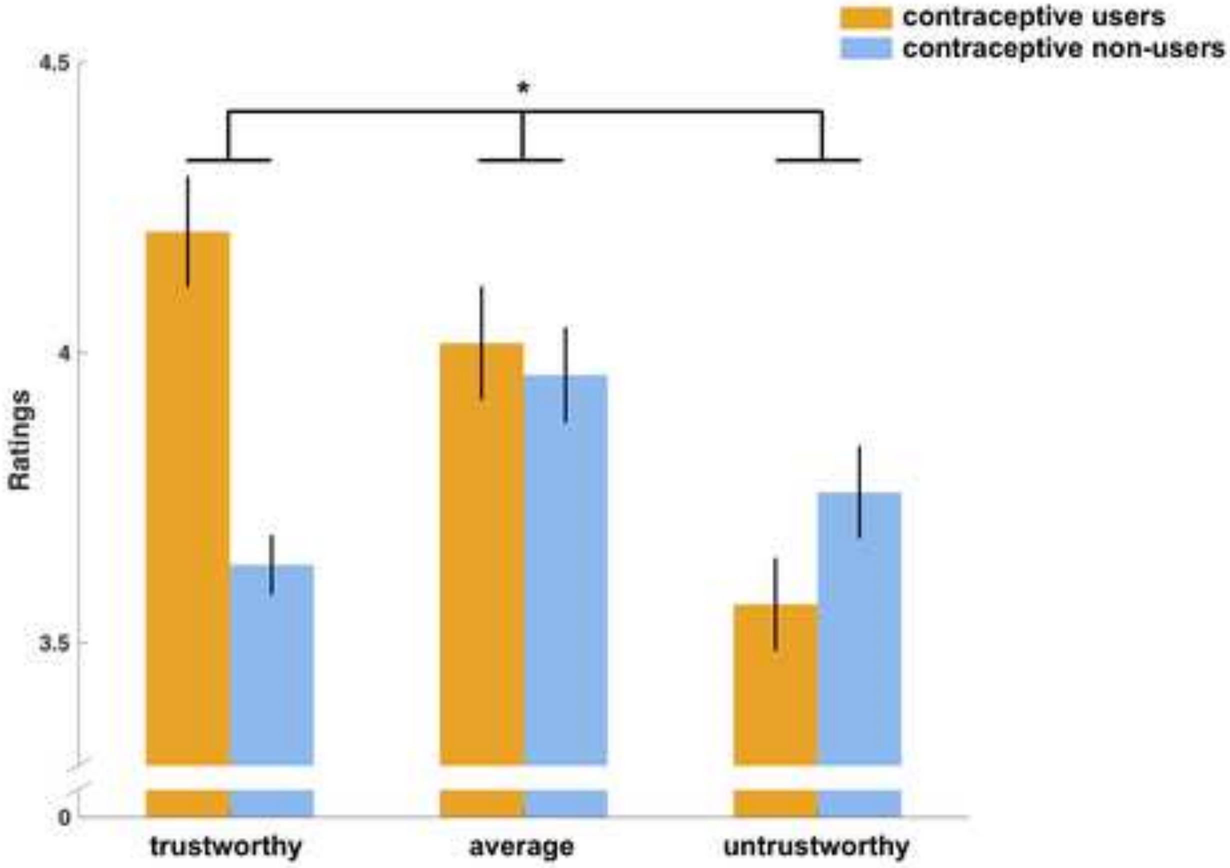

A
*
Amount of money sent
6
5
4
0
trustworthy
average
untrustworthy
B
pramipexole
placebo
*
Amount of money sent
6
5
4
3
0
contraceptive users
contraceptive non-users

Fig. 5


🟧 contraceptive users
🟦 contraceptive non-users
*
Ratings
4.5
4
3.5
0
trustworthy
average
untrustworthy

Tab. S1. Model selection

Model	N° fixed-effects parameters	Model AIC		
		Trust Game	Trustworthiness	Attractiveness
1	6	2,498	2,261	2,432
2	7	**2,495**	2,262	2,434
3	7	2,499	2,262	**2,430**
4	7	2,499	2,261	2,433
5	7	2,507	2,265	2,440
6	7	2,508	2,274	2,445
7	7	2,505	**2,260**	2,434
8	8	2,511	2,275	2,449
9	8	2,519	2,283	2,456
10	9	2,516	2,285	2,458
11	10	2,524	2,287	2,463
12	10	2,504	2,276	2,450
13	10	2,505	2,273	2,442
14	10	2,509	2,264	2,435
15	13	2,518	2,288	2,461
16	13	2,515	2,280	2,454
17	13	2,515	2,276	2,448
18	15	2,518	2,293	2,467

Each model contained a fixed-effects intercept in addition to the fixed-effects regressors of interest (included in the count). AIC, Akaike Information Criterion. In bold is depicted the winning model.

Tab. S2. Trust in the Trust Game

Regressors	***β* (SE)**
Intercept	4.64 (1.27)***
Treatment	-0.98 (0.38)*
Gender	0.05 (0.20)
Trustworthiness levels	0.60 (0.13)***
Contraceptive use	-0.46 (0.78)
Session	-0.30 (0.78)
Treatment * Contraceptive use	1.32 (0.49)**

β coefficients (standard errors, SE) from the winning mixed-effects regression model for participants' trusting behavior in the trust game.

*p < .05; **p < .01; ***p < .001

Tab. S3. Trustworthiness perceptions

Regressors	***β* (SE)**
Intercept	4.16 (0.49)**
Treatment	-0.13 (0.10)
Gender	-0.66 (0.29)*
Trustworthiness levels	0.20 (0.21)
Contraceptive use	-0.002 (0.16)
Session	-0.20 (0.18)
Gender * Trustworthiness levels	0.34 (0.13) *

β coefficients (standard errors, SE) from the winning mixed-effects regression model for participants' ratings of trustee's trustworthiness.

**p* < .05; ***p* < .001

Tab. S4. Attractiveness perceptions

Regressors	***β* (SE)**
Intercept	3.78 (0.46)**
Treatment	-0.006 (0.12)
Gender	0.30 (0.15)*
Trustworthiness levels	-0.02 (0.13)
Contraceptive use	-0.60 (0.35)
Session	-0.28 (0.19)
Trustworthiness levels * Contraceptive use	0.34 (0.16)*

β coefficients (standard errors, SE) from the winning mixed-effects regression model for participants' trusting behavior in the trust game.

**p* < .05; ***p* < .001

Appendix

Anlage A — Curriculum vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

Anlage B –Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt,

- dass ich die vorliegende Arbeit eigenständig und ohne unerlaubte Hilfe verfasst habe,
- dass Ideen und Gedanken aus Arbeiten anderer entsprechend gekennzeichnet wurden,
- dass ich mich nicht bereits anderwärtig um einen Doktorgrad beworben habe und keinen Doktorgrad in dem Promotionsfach Psychologie besitze, sowie
- dass ich die zugrundeliegende Promotionsordnung vom 08.08.2016 anerkenne.

Ort, Datum Unterschrift

Anlage C

Erklärung gemäß § 7 Abs. 3 Satz 4 der Promotionsordnung über den Eigenanteil an den veröffentlichten oder zur Veröffentlichung vorgesehenen eingereichten wissenschaftlichen Schriften im Rahmen meiner publikationsbasierten Arbeit

I.	Name, Vorname:	Bellucci, Gabriele
	Institut:	German Institute of Human Nutrition
	Promotionsfach:	Psychologie
	Titel:	Master of Arts (MA), Master of Science (MSc)

II. Nummerierte Aufstellung der eingereichten Schriften (Titel, Autoren, wo und wann veröffentlicht bzw. eingereicht):

1. **Bellucci G**, Münte T F, Park S Q, *Resting-state dynamics as a neuromarker of dopamine administration in healthy female adults*, J Psychopharmacol, doi: 10.1177/0269881119855983, 2019

2. **Bellucci G**, Molter F, Park S Q, *Neural representations of honesty predict trust*, Nature Communications, 2nd round of revision, decision letter on minor revision received on August 9, 2019

3. **Bellucci G**, Park S Q, *Honesty biases trustworthiness impressions*, Journal of Experimental Psychology: General, submitted on March 31, 2019

4. **Bellucci G**, Münte T F, Park S Q, *Effects of a dopamine agonist on trusting behaviors in females*, Psychopharmacology, submitted on July 18, 2019

III. Darlegung des eigenen Anteils der Schriften:

Die Bewertung des Eigenanteils richtet sich nach der Skala: “vollständig – überwiegend – mehrheitlich – in Teilen” und enthält nur für den jeweiligen Artikel relevante Arbeitsbereiche.

Zu II.1.: Konzeption (überwiegend), Versuchsdesign (überwiegend), Programmierung (vollständig), Datenerhebung (in Teilen), Datenauswertung (vollständig), Ergebnisdiskussion (überwiegend), Erstellen des Manuskriptes (überwiegend).

Zu II.2.: Konzeption (überwiegend), Versuchsdesign (vollständig), Programmierung (vollständig), Datenerhebung (überwiegend), Datenauswertung (vollständig), Ergebnisdiskussion (überwiegend), Erstellen des Manuskriptes (überwiegend).

Zu II.3.: Konzeption (überwiegend), Versuchsdesign (vollständig), Programmierung (vollständig), Datenerhebung (vollständig), Datenauswertung (vollständig), Ergebnisdiskussion (überwiegend), Erstellen des Manuskriptes (überwiegend).

Zu II.4.: Konzeption (überwiegend), Versuchsdesign (vollständig), Programmierung (vollständig), Datenerhebung (in Teilen), Datenauswertung (vollständig), Ergebnisdiskussion (überwiegend), Erstellen des Manuskriptes (überwiegend).

IV. Die Namen und Anschriften nebst E-Mail oder Fax der jeweiligen Mitautorinnen oder Mitautoren:

zu II.1.: Thomas F Münte: Department of Neurology, Universitätsklinikum Schleswig-Holstein, 23538 Lübeck, Germany; Department of Psychology II, University of Lübeck, 23562 Lübeck, Germany.

Soyoung Q Park: Decision Neuroscience and Nutrition, German Institute of Human Nutrition (DIfE), Potsdam-Rehbruecke, Germany; Charité-Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Neuroscience Research Center, 10117, Berlin, Germany; Deutsches Zentrum für Diabetes, 85764, Neuherberg, Germany.

zu II.2.: Felix Molter: WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin.

Soyoung Q Park: s.o.

zu II.3.: Soyoung Q Park: s.o.

zu II.4.: Thomas F Münte: s.o.

Soyoung Q Park: s.o.

Ich bestätige die von Gabriele Bellucci unter III. angegebene Erklärung:

Name: Soyoung Q Park Unterschrift:____________________________

Name: Felix Molter Unterschrift:____________________________

Name: Thomas F Münte Unterschrift:____________________________