Psychological and Neural Mechanisms of Trust Formation from a Computational Modeling Perspective: The Investor' s Perspective in Trust Games

Authors: Gao Qinglin, Zhou Yuan, Zhou Yuan

Date: 2020-09-27T00:00:00+00:00

Abstract

Interpersonal trust pervades various aspects of social interaction and constitutes a crucial cornerstone for facilitating and sustaining cooperation. Previous researchers have employed the trust game paradigm to investigate theoretical models, biological substrates, and influencing factors of interpersonal trust. In recent years, researchers have begun applying computational models to trust game data analysis to elucidate the psychological mechanisms underlying interpersonal trust behavior, and have integrated computational models with neuroimaging techniques to enhance understanding of the neural mechanisms underlying trust behavior. Current research utilizing computational models within the trust game paradigm primarily focuses on the scientific question of "how trust is formed" ; future directions should involve further developing computational modeling approaches, integrating non-invasive brain stimulation techniques, and applying these methods to psychiatric populations to deepen our understanding of the psychological and neural mechanisms of normal and aberrant trust formation.

Full Text

Psychological and Neural Mechanisms of Trust Formation from a Computational Modeling Perspective: The Investor' s Viewpoint in the Trust Game

Qinglin Gao[1],[2], Yuan Zhou[1],[2]
[1]Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[2]Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract
Interpersonal trust permeates all aspects of social interaction and serves as a crucial foundation for promoting and maintaining cooperation. Previous researchers have employed the trust game paradigm to investigate the theoretical models, biological underpinnings, and influencing factors of interpersonal trust. In recent years, computational modeling has been increasingly applied to trust game data analysis, enabling deeper exploration of the psychological mechanisms underlying trust behavior. Integrating computational modeling with neuroimaging techniques has further enhanced our understanding of the neural mechanisms that support trust. Current applications of computational modeling to the trust game have primarily addressed the fundamental question of "how trust is formed." Future research should advance computational modeling methods, combine them with non-invasive brain stimulation techniques, and apply them to clinical populations to elucidate the psychological and neural mechanisms of both normal and pathological trust formation.

Keywords: interpersonal trust; trust game; computational modeling; functional magnetic resonance imaging

Trust functions as a lubricant in economic and social life (Snijders & Keren, 2001) and as a bonding agent that maintains social relationships (Wilson & Eckel, 2006). As one of the most complex social skills, interpersonal trust plays a vital role in social interactions (Fett et al., 2014). Although definitions of interpersonal trust vary, its core essence refers to a psychological state in which individuals, based on positive expectations of others' behavior (e.g., anticipating cooperation in situations that could involve either cooperation or competition), willingly place themselves at risk (Krueger et al., 2007; Rotter, 1967). This definition highlights two key features: holding positive expectations about others' intentions and exposing oneself to risk or disadvantage. Within the framework of economic game theory, economists have abstracted interpersonal trust from its complex definitional context while preserving its essential characteristics, operationalizing it as the Trust Game (TG) paradigm, which has been widely adopted in trust research. Previous investigations have examined the biological foundations of interpersonal trust at multiple levels, including neurotransmitters and hormones (molecular level), decision-making and reasoning (cognitive level), and brain regions and networks (functional level), proposing various theoretical models to understand the psychological and neural mechanisms underlying trust (Krueger & Meyer-Lindenberg, 2019; Riedl & Javor, 2012; Tzieropoulos, 2013; 陈欣, 叶浩生, 2009; 陈瀛等, 2020; 史燕伟等, 2015; 张宁等, 2011; 张蔚等, 2016). While these studies have provided answers to the question of "why and when people choose to trust or distrust" (张蔚等, 2016), traditional methods cannot address the question of "how trust is formed." The development of the repeated trust game paradigm and the application of computational modeling methods have made it possible to answer this latter question.

In recent years, computational modeling has been increasingly applied in decision-making research due to its rigorous quantitative approach and its

ability to reveal the hidden dynamic psychological processes underlying behavior and brain activity. This approach offers a novel framework for deepening our understanding of the psychological mechanisms and neural foundations of behavior (Montague et al., 2012; Read Montague, 2018). Moreover, this data-driven quantitative approach not only allows for model evaluation but also enables model comparison to determine which model better explains and predicts psychological phenomena (Cheong et al., 2017). Game paradigms are commonly used in computational modeling research. In repeated games, individuals must infer their opponents' mental states, and such inferences are recursive—that is, they involve cyclical causal relationships. This recursiveness constitutes the core concept underlying many computational models, leading to increasing applications of computational modeling to various game paradigms, including the trust game, to explore the internal mechanisms of diverse psychological phenomena (Ray et al., 2009). This paper first introduces the trust game paradigm and the concept of computational modeling, then reviews applications of computational modeling in behavioral and neuroimaging studies of interpersonal trust, focusing on the question of "how trust is formed." We synthesize current research progress on the psychological and neural mechanisms of trust formation and, finally, identify limitations in existing research and propose new directions for further investigation.

2. Single-Round and Repeated Trust Games

The trust game is the most commonly used paradigm for studying interpersonal trust. In the classic trust game (Kreps, 1990), two participants assume the roles of investor and trustee, each starting with an identical amount of money. First, the investor decides whether to trust (send all money to the partner) or distrust (keep all money). If the investor chooses to trust, the invested amount is multiplied (typically tripled) and given to the trustee. If the investor chooses not to invest, the game ends and both parties retain their original amounts. Next, the trustee decides whether to reciprocate (return half of the money) or not reciprocate (keep all money). If the trustee reciprocates, both parties end up with double their original amount; if not, the trustee receives triple the original amount while the investor receives nothing. This paradigm allows trust to be quantified as the investor' s decision and trustworthiness as the trustee' s decision. Berg et al. (1995) modified this to create the standard trust game paradigm, which differs from the classic version in that investors and trustees can voluntarily decide how much money to give or return, rather than being limited to all-or-nothing decisions. This modification enables measurement of varying levels of trust and reciprocity. Using this paradigm, Berg et al. (1995) found that people choose to trust and reciprocate even in one-shot interactions with strangers, a result replicated in numerous subsequent studies (Declerck et al., 2013; Johnson & Mislin, 2011). Researchers have also developed variants of this paradigm for specific experimental purposes, such as allowing one minute of verbal communication before decisions, providing investors with promises from trustees about return amounts, playing against real or computer-simulated part-

ners, and playing against trustees of different social status levels, to investigate factors influencing trust behavior in the trust game (Ben-Ner et al., 2011; Blue et al., 2020; Ma et al., 2015; Tzieropoulos, 2013).

In the classic trust game, interactions between the same pair of players are single-round, whereas real-life social interactions rarely occur only once. Researchers therefore developed the Repeated Trust Game (RTG) paradigm, in which the same pair of players engages in multiple consecutive trust games and can adjust their decisions based on immediate feedback (Figure 1 [Figure 1: see original paper]). Unlike single-round trust games, both parties in repeated trust games face the risk that the other may not return money. Consequently, not only does the investor' s trust behavior depend on the trustee' s return amounts, but the trustee must also consider the investor' s behavior (which is not the case in single-round games). This leads to different behavioral patterns in repeated versus single-round games. Research has found that trustees return more money in repeated trust games than in single-round games to encourage greater investments from investors (Cochard et al., 2004), and that participants' trust and reciprocity decisions show a monotonic decreasing trend with an endgame effect—where trust and reciprocity decisions drop sharply as the game nears its conclusion (Anderhub et al., 2002; Keser, 2003). Compared to single-round trust games, repeated trust games involve multiple cognitive processes including learning, reasoning, and strategy updating, providing a more ecologically valid experimental paradigm for studying trust formation and making it possible to introduce reinforcement learning and other computational models in social interaction contexts (Anderhub et al., 2002; King-Casas et al., 2005).

[Figure 1: see original paper] Schematic diagram of the repeated trust game paradigm

3.1 Overview of Computational Modeling

Computational modeling uses abstract mathematical expressions to characterize the dynamic processes of learning and decision-making in human social interactions (Hackel & Amodio, 2018), enabling the depiction of hidden dynamic psychological processes underlying behavior or brain activity (Montague, 2018). Computational models can investigate the dynamic processes of psychological phenomena based on behavior and explore the neural mechanisms underlying these phenomena when combined with brain imaging techniques. A rapidly developing approach involves combining computational modeling with brain imaging, such as model-based functional Magnetic Resonance Imaging (fMRI). fMRI measures brain activity evoked by experimental stimuli by assessing changes in Blood Oxygen Level Dependent (BOLD) signals, where increased BOLD signal in a brain region indicates activation. In traditional fMRI studies, researchers typically correlate BOLD signals with behavioral measures such as accuracy and reaction time to establish associations between behavioral tendencies and brain function (Engelmann, 2010). In contrast, model-based fMRI studies can extract internal variables that cannot be directly observed from experimental

paradigms (e.g., reward prediction errors, learning rates) from behavioral data through model-based calculations, simulating the complex cognitive processes underlying behavioral phenomena. These variables or model parameters are then correlated with experimentally evoked BOLD signals, establishing links among behavior, cognition, and brain function to better understand the neural mechanisms of behavior (Charpentier & O' Doherty, 2018; O' Doherty et al., 2007).

Current computational models used in interpersonal trust research can be divided into two categories: outcome-based models and intention-based models (McCabe et al., 2003). Outcome-based models posit that in trust games, inferring intentions is less important than the feedback outcomes individuals obtain from interactions—that is, people primarily focus on their own monetary gains. In contrast, intention-based models emphasize that inferring the partner' s intentions is more critical in decision-making processes, with individuals making decisions based on their partner' s intentions. The primary outcome-based model used in trust games is the reinforcement learning model (Cisler et al., 2015; Fouragnan, 2013; Radell et al., 2016), while the main intention-based model is the Bayesian model (Jung et al., 2017; Moutoussis et al., 2014; Ray et al., 2009). Studies using reinforcement learning models have primarily addressed how prior trustworthiness promotes trust formation, whereas Bayesian models in trust research have mainly focused on how inference of others' intentions facilitates trust formation. Since only the investor' s behavior reflects trust decisions in the trust game, and current computational modeling studies of trust behavior analyze only the investor' s actions, this paper reviews computational modeling research on trust behavior from the investor' s perspective.

3.1.1 Reinforcement Learning Models

Reinforcement Learning (RL) models are the most commonly used computational models for investigating psychological and neural mechanisms in economic decision-making, addressing how people learn from feedback through repeated interactions with their environment (Read Montague, 2018). These models assume that the interaction between an individual and the environment follows a Markov Decision Process (MDP), which includes environmental states (S), individual actions (A), and transition probabilities (P) and rewards (R) that link them. States refer to the individual' s current situation, which determines available actions, while transition probabilities indicate the likelihood of moving from one state to another after taking a particular action (Puterman, 1995). As shown in Figure 2 [Figure 2: see original paper], at time t, an agent perceives the current state St and receives reward Rt, then takes action At. This action leads to state St+1 and reward Rt+1 at time t+1, with the probability of environmental state transition given by P(St+1|St, A) (Fouragnan, 2013). RL models posit that individuals learn the relationship between actions and feedback outcomes in different environmental states, updating expected utility values for actions based on prediction errors (the gap between expected and observed values), and

making adaptive decisions to maximize rewards. The learning rate parameter reflects the weight individuals assign to outcome feedback, measuring the speed of updating expected utility values, with higher values indicating greater weight on feedback and faster updating (Claus & Boutilier, 1998).

RL models are divided into model-free and model-based categories based on whether they incorporate prior models (Montague et al., 2012). Model-free RL theory suggests that individuals make decisions based on "trial-and-error" principles, relying only on previously learned outcomes—similar to stimulus-response habitual behavior. The most commonly used model is the Rescorla–Wagner (RW) model. Model-based RL theory, in contrast, posits that individuals form an internal model of the external environment based on feedback, which serves as an internal representation of the world and enables goal-directed behavior (Daw & Doya, 2006). The key difference lies in the presence of an internal model; model-based RL processes feedback more flexibly and enables faster adaptation to environmental changes.

In the context of trust games, the feedback signal for investors comes from whether trustees return money and how much they return. Under model-free RL assumptions, investors would make decisions based solely on observed trustworthiness levels regardless of external cues about the trustee' s reliability. In model-based RL, investors are assumed to form prior expectations about trustees' trustworthiness based on reputation cues and then update subsequent prediction errors based on these prior expectations (Fouragnan, 2013).

3.1.2 Bayesian Models

[Figure 2: see original paper] Framework diagram of the reinforcement learning model
Source: Fouragnan (2013)

Reinforcement learning models are based on the classical economic assumption of perfect rationality, yet numerous findings have violated this assumption (Fehr & Schmidt, 2005). Moreover, RL models assume that individuals must acquire all possible states in the environment through MDPs, whereas in real social interactions, the environment is uncertain and only partially observable. Researchers have therefore proposed Bayesian models based on Partially Observable Markov Decision Processes (POMDP). These models assume bounded rationality: before social interaction, individuals hold preferences about environmental states, which constitute prior beliefs. During interaction, individuals update these prior beliefs based on environmental feedback, resulting in posterior beliefs that guide adaptive decision-making. Such models typically represent beliefs using probability distributions (Kaelbling et al., 1995). As shown in Figure 3 [Figure 3: see original paper], prior beliefs are represented by prior probability distribution Pr before processing external information, while posterior beliefs are represented by posterior probability distribution P after information processing. At time t, posterior belief formation is based on the prior belief Pr, the observed interac-

tion behavior set Ot, and the reward set Rt at that moment. The agent takes action At based on posterior beliefs, which in turn leads to interaction observation set Ot+1 and reward set Rt+1 at time t+1. Based on Ot+1 and Rt+1, prior and posterior beliefs are further updated, enabling new actions at t+1 according to the updated posterior beliefs. The key difference between Bayesian inference models and RL models is that the latter iterate value functions over time, while the former iterate belief distributions (prior and posterior beliefs) over time (Friston et al., 2013).

When states in an MDP are belief states that are uncertain and only partially observable, they can be represented as Partially Observable Markov Decision Processes (Khalvati et al., 2019). Researchers have further proposed Interactive POMDP (IPOMDP), where each individual' s decision-making process follows a standard POMDP—essentially making IPOMDP a collection of POMDPs. Repeated trust games can be viewed as two individuals' IPOMDPs, where both parties' states depend on each other' s decisions and their internal models of each other' s intentions (Hula et al., 2015).

[Figure 3: see original paper] Framework diagram of the Bayesian model
Source: Friston et al., 2013

Bayesian models are primarily used to study how people make decisions based on intention inference under uncertainty. Some researchers have introduced "Theory of Mind" —the ability to infer one' s own and others' intentions—into Bayesian models (Ray et al., 2009). They argue that in game tasks, people need Theory of Mind for strategic reasoning about players' intentions and behaviors (Gonzalez & Chang, 2019; Ong et al., 2019). In trust games, Theory of Mind manifests as both parties' inferences about each other' s type: investors infer the trustee' s type, trustees infer the investor' s type, and investors infer how they are perceived by trustees (Rusch & Gläscher, 2019). Researchers have classified individuals into different thinking-depth groups based on the levels of inference involved and used parameters to measure individual thinking depth (Ray et al., 2009; Xiang et al., 2012). Friston et al. (2013) approached the problem from environmental/cognitive uncertainty, introducing the free-energy principle into Bayesian models to propose active inference for simulating decision-making in trust games (Moutoussis et al., 2014). Unlike Theory of Mind-based Bayesian models, active inference models use parameters measuring the precision of individuals' own strategies rather than the depth of strategic thinking.

3.2 Behavioral and Neuroimaging Studies Based on Reinforcement Learning Models

Reinforcement learning models help researchers better understand how investors make decisions based on environmental information in repeated trust games and the neural mechanisms underlying these decisions. Environmental information includes prior trustworthiness information obtained before decision-making and trustworthiness information acquired during interaction (Fareri et al., 2012,

2015; Fouragnan, 2013).

3.2.1 Behavioral Studies

Current behavioral research has used RL models to investigate how healthy participants make trust decisions when prior trustworthiness information is available. Chang et al. (2010) provided prior reputation cues about facial trustworthiness (high/medium/low) and compared three model-based RL models to examine how prior trustworthiness influences trust establishment. These models included a gain-loss theory-based model (people prefer risk avoidance over gain acquisition), a confirmation bias theory-based model (people weight information consistent with advice more heavily than inconsistent information), and the authors' proposed dynamic belief iteration model. The dynamic belief iteration model posited that prior information influences trust behavior throughout the entire trust game process, with participants forming beliefs about the likelihood of reciprocity based on prior information and updating these beliefs iteratively based on actual experience. Results showed that the dynamic belief iteration model best predicted how prior trustworthiness influences trust formation, suggesting that trust is established through dynamic belief iteration based on prior trustworthiness.

In another study, participants first learned about trustees' character traits (good/neutral/bad) through a ball-tossing game before completing a repeated trust game as investors (Fareri et al., 2012). Using a gain-loss theory-based model-based RL model, the authors found that initial social impressions learned from direct interaction interact with subsequent feedback signals. Social impressions influence trust behavior during interaction, while feedback outcomes in turn affect initial social impressions, which are iteratively updated throughout repeated interactions. Fareri et al. (2015) extended this work by examining how opponents with different levels of intimacy affect trust behavior in repeated trust games. By comparing friends, strangers, and computer opponents, they investigated how prior trustworthiness levels influence trust behavior. Results showed that participants exhibited smaller prediction errors when interacting with opponents with higher prior reputations. Additionally, Radell et al. (2016) used the same experimental design and RW RL model to examine how inhibited personality types (those tending toward avoidance in social situations) make trust decisions toward opponents of different trustworthiness levels. They found that inhibited participants showed less trust toward moderately trustworthy opponents compared to non-inhibited participants, due to lower initial trust values for moderately trustworthy opponents. This suggests that individuals prone to social avoidance interpret neutral or ambiguous information more negatively.

Beyond studying how prior reputation cues affect trust behavior, researchers have also compared decision-making processes with and without prior information. Fouragnan (2013) proposed two ways investors obtain prior trustworthiness information: a prior condition where opponents' trustworthiness is disclosed beforehand, and a no-prior condition where investors interact directly with op-

ponents without prior information. By comparing model-free and model-based RL models, the study investigated the psychological mechanisms underlying trust behavior toward high/low trustworthiness opponents under both conditions. The belief adaptation model was found to best explain investors' trust behavior in repeated trust games. This model posits that prior trustworthiness information serves as a social signal that influences not only initial decision values but also subsequent iterative decision functions based on reciprocity feedback experience. People form beliefs about opponents' trustworthiness levels (Trustworthiness belief, TW) based on prior information, and these beliefs are updated iteratively in the utility function alongside monetary feedback outcomes, similar to bonus rewards. The study found that in both prior and no-prior conditions, people first form beliefs about opponents' trustworthiness based on monetary feedback and then adjust their decisions accordingly to make adaptive investment behaviors. The key difference is that prior trustworthiness information alters investors' initial expectations about opponents' reliability (Fouragnan, 2013). Other research using RL models has found that trust behavior in repeated trust games reflects a mutual learning process between both parties, where people make decisions based on feedback from multiple interactions and show heightened sensitivity to negative outcomes, rapidly adjusting their decisions after negative feedback to produce adaptive behavior (Haiyan, 2018).

In summary, behavioral studies reveal that trust in repeated trust games is a continuous learning process through which people evaluate outcomes from multiple interactions to learn about others' reputation levels and decide whether to trust them. RL theory can effectively reveal this dynamic trust establishment process.

3.2.2 Neuroimaging Studies

Using functional magnetic resonance imaging, researchers have further explored the neural mechanisms through which prior trustworthiness promotes trust formation. Fareri et al. (2012) investigated how prior trustworthiness obtained before decision-making influences trust behavior and reward-related brain activity. Participants first played a ball-tossing game with computer-simulated opponents of three different trustworthiness levels (high/medium/low) to learn initial impressions. In the subsequent repeated trust game, participants acted as investors playing against these opponents (whose actual behavior was random and independent of the ball-tossing game). Using the RW RL model to analyze behavior and its neural correlates, the study found that belief updating was faster when experienced opponent behavior matched prior impressions compared to mismatched cases. When facing positive/negative feedback, striatal and anterior cingulate activation increased compared to neutral feedback, and learning rate parameters from the model correlated significantly with BOLD signal changes in these regions. This indicates that BOLD signals in reward circuitry brain regions reflect prediction error signals used to update beliefs at the behavioral level, demonstrating that these regions are responsible for belief

updating through prediction errors in gain/loss contexts. These results suggest that initial impressions learned from direct social interaction are continuously updated through reinforcement learning mechanisms, particularly when information is consistent.

Fouragnan (2013) compared trust decisions and corresponding brain activation between prior and no-prior trustworthiness conditions to investigate the neural basis of how prior trustworthiness influences trust decisions. Results showed that striatal activation correlated significantly with RL model estimates of behavior only in the no-prior condition, with prior trustworthiness disrupting this correlation. In the prior condition, negative striatal activation in response to trust violations correlated with learning rates in the RL model, but this correlation was absent in the no-prior condition. Participants continued to rely on prior information even when experience contradicted it. Compared to the no-prior condition, negative activation in the caudate nucleus was stronger when cooperative opponents violated trust in the prior condition. Prior information enhanced connectivity between the striatum and ventrolateral prefrontal cortex, modulating tolerance for violations, which correlated negatively with retaliation rates. Additionally, prior trustworthiness affected initial trust decisions, reflected in prefrontal cortex activation. Fareri et al. (2015) further investigated the neural basis of how intimacy with opponents influences trust behavior by incorporating social value reward signals into the RL model. Participants played as investors against friends, strangers, and computers. The model posited that feedback outcomes included not only monetary rewards but also social value reward signals, represented by participants' initial trustworthiness ratings of opponents. Results showed that participants extracted social value reward signals from feedback based on intimacy levels, with these signals correlating significantly with activation in the ventral striatum and medial prefrontal cortex, indicating that people make trust decisions based on social value reward signals during repeated social interactions.

These neuroimaging studies not only validate findings from behavioral computational modeling research but also reveal that prior trustworthiness influences initial trust decisions, identifying the neural basis of dynamic trust iteration. The striatum and anterior cingulate in reward circuitry reflect belief updating through prediction errors; prefrontal cortex activation reflects the influence of prior trustworthiness on initial trust decisions; and dynamic connectivity between the striatum and prefrontal cortex reflects regulation of trust behavior during the game.

3.3 Behavioral and Neuroimaging Studies Based on Bayesian Models

Applications of Bayesian models to trust games have focused on understanding how investors make trust decisions based on intention inference in repeated trust games and the underlying neural mechanisms.

3.3.1 Behavioral Studies

Ray et al. (2009) introduced Theory of Mind into Bayesian models, establishing a belief hierarchy model for trust behavior in repeated trust games within the IPOMDP framework. This model assumes that players know their own cooperative/non-cooperative type but not their opponent' s type, making it a dynamic game of incomplete information. Players' prior beliefs about opponents' trustworthiness are updated in a Bayesian manner based on observed behavior, and players' own actions also influence their beliefs about opponents' trustworthiness. This process involves a limited hierarchy of beliefs: what type the investor thinks the trustee is, what type the trustee thinks the investor thinks they are, and so on. Games reach a subjective Bayes-Nash Equilibrium (BNE) when players form conclusions about opponents' trustworthiness through multiple interactions. The model' s innovation lies in introducing strategic thinking levels into the IPOMDP framework to explain how social utility, strategic level, and prior beliefs influence trust behavior. Through model inversion, participants can be classified into different strategic thinking levels based on their dynamic trust behavior during experiments, with higher strategic thinking levels associated with greater investment frequencies. This model provides a new approach for studying individual differences in interpersonal trust.

Hula et al. (2015) used partially observable Monte Carlo planning (POMCP) algorithms to investigate repeated trust games within the IPOMDP framework. Results showed that investors form beliefs about opponents' trustworthiness after approximately 10 game rounds and subsequently make stable investment decisions. Therefore, investors' behavior during the first 10 rounds can be used to infer optimal parameter values in their internal subjective models, ensuring that these behaviors reflect decisions made under their internal models. This algorithm also enables inference of participants' ability to infer others' intentions through model inversion.

Friston et al. (2013) developed an active inference Bayesian model for social decision-making within the IPOMDP framework based on Bayesian theory. This model introduces parameters for the precision of prior beliefs and proposes using the free-energy principle to update posterior beliefs. Moutoussis et al. (2014) applied this model to trust games, combining utility functions, prior beliefs, and outcomes to model the evolution of trust as game rounds increase, finding that investors form beliefs about opponents' trustworthiness after approximately 10 interactions. Schwartenbeck et al. (2015) experimentally demonstrated that decision theory under active inference better predicts human economic decision-making than utility maximization theory.

Researchers have also applied Bayesian models to practical problems. Jung et al. (2017) constructed a medical trust game to study placebo analgesic effects, establishing a Bayesian framework that creates a likelihood relationship between pain intensity and pain ratings, with pain ratings corresponding to posterior distributions and placebo effects representing the gap between ascending sen-

sory signals and descending pain predictions. People' s subjective pain ratings can thus be inferred from posterior distributions in the Bayesian model. By comparing Bayesian and linear regression models, the study found that prior expectations influence pain perception and that Bayesian models can predict pain ratings in medical trust games.

In summary, behavioral studies based on Bayesian inference reveal that in repeated trust games, people form beliefs about opponents' trustworthiness after approximately ten interactions and make decisions accordingly. Individuals differ in their ability to infer others' intentions, exhibiting varying depths of thinking during games.

3.3.2 Neuroimaging Studies

Xiang et al. (2012) used functional magnetic resonance imaging to investigate whether the ability to infer others' intentions could serve as objective biomarkers for deviations in trust game behavior. Using a Theory of Mind-based Bayesian model, they characterized participants' thinking depth with model parameters and classified them into high/medium/low thinking-depth groups. Results showed that low thinking-depth participants exhibited stronger striatal activation than high and medium thinking-depth participants, while high thinking-depth participants showed stronger activation in the temporoparietal junction (TPJ)—a region associated with Theory of Mind—than medium and low thinking-depth groups. This suggests that low thinking-depth participants are more sensitive to feedback outcomes and adjust their behavior primarily based on these outcomes, whereas high thinking-depth participants make decisions primarily through intention inference.

Nihonsugi et al. (2015) combined fMRI, transcranial direct current stimulation (tDCS), and computational modeling to investigate whether intention inference and feedback processing represent two separable neural systems in trust decision-making. They developed a model combining guilt aversion, inequity aversion, and utility functions from RL models. By correlating guilt sensitivity and inequity sensitivity parameters from this model with imaging results, they found that right dorsolateral prefrontal cortex (DLPFC) activation was associated with intention-based economic decision-making, while ventral striatum and amygdala activation were associated with feedback-based economic decision-making. Selective stimulation of DLPFC enhanced intention-based decision-making. These results indicate that right DLPFC plays an important role in processing intention-based cooperative behavior. Overall, Nihonsugi et al. (2015) proposed that repeated trust games involve two separable neural systems: one for inferring others' intentions and another for making decisions based on feedback outcomes. Through multiple interactions, people learn opponents' trustworthiness levels from reward signals (striatal activation) and intention inferences (DLPFC and cingulate activation), making adaptive decisions based on these processes (with BOLD signals in these regions correlating significantly with prediction errors from the model). Moreover, prior trustworthiness strengthens the connection

between these two systems.

These neuroimaging studies have not only identified the neural basis of individual differences in trust decisions among people with different thinking depths but also revealed the existence of two separable neural systems for intention inference and feedback-based decision-making in repeated trust games.

4. Limitations and Future Directions

In summary, computational modeling-based behavioral and neuroimaging studies have identified the psychological and neural mechanisms through which prior trustworthiness and intention inference promote trust formation, providing deeper understanding of "how trust is formed." However, several limitations remain, and several directions warrant further exploration.

4.1 Development of Computational Models

Current computational models applied to trust games primarily include reinforcement learning models and Bayesian models. RL models assume perfect rationality, positing that individuals update expected values of actions based on current prediction errors and use learning rates to measure the weight assigned to feedback outcomes (with learning rates showing individual differences). Since its proposal, this model has been widely applied in various learning tasks, and its combination with neuroimaging has yielded important discoveries about brain reward functions (Jaafra et al., 2019; Lee et al., 2012). However, the objectively adaptive learning process of RL models is difficult to apply in real life, where situations are uncertain and action utility values are unknown and must be inferred (Mathys et al., 2011).

Bayesian models, in contrast, assume bounded rationality and combine Bayesian theory with conditional probability to link individuals' beliefs with their actions, effectively characterizing decision-making under uncertainty (Mathys et al., 2011). Active inference models have been applied in various research domains (Friston et al., 2016; Parr & Friston, 2017; Smith et al., 2019), with simulation studies identifying specific parameters affecting different behaviors, such as strategic depth, decision uncertainty, and prior beliefs (Smith et al., 2019). However, only one study has applied this model to trust games (Moutoussis et al., 2014). Future research should apply this model more extensively to simulate trust formation processes, identify key parameters influencing trust formation, and design functional imaging tasks to examine the neural basis of these parameters, thereby establishing the psychological and neural mechanisms underlying individual differences in trust formation.

Researchers have also attempted to develop other types of computational models. For example, Mathys et al. proposed the Hierarchical Gaussian Filter (HGF) model, which combines Bayesian probability-based uncertainty characterization with RL model' s approach to individual differences in updating. The model' s update equations are similar to RL models (driven by prediction errors) but

differ in using individuals' trade-off between strategy precision as the learning rate to characterize decision-making under environmental and perceptual uncertainty. This model has been successfully applied in social exchange scenarios requiring intention inference (Diaconescu et al., 2014). Additionally, some studies have combined RL utility functions with Bayesian probability-based uncertainty characterization to propose the Fehr–Schmidt inequality aversion model (FS model). This model includes parameters for individual differences in learning rates and inequality aversion, as well as parameters for strategic depth in intention inference and planning horizon, providing a comprehensive simulation of decision-making in trust games. By comparing parameters between groups receiving high-quality early education versus those who did not, one study found that individuals who received high-quality early education planned more steps ahead in trust games and other social interactions, indicating long-term beneficial effects of early education on social decision-making (Luo et al., 2018).

Future research should flexibly select and develop various models for trust game studies to deepen understanding of the dynamic trust iteration process and promote comprehension of the psychological and neural mechanisms underlying individual differences in trust games.

4.2 Causal Brain-Behavior Research Based on Computational Models

Although computational modeling-based neuroimaging research can characterize the dynamic neural activity patterns underlying specific cognitive processes in both temporal and spatial dimensions, establishing links between cognition and brain function, current studies cannot address causal relationships between brain and behavior. Computational modeling studies of abnormal decision-making in brain-damaged patients (Gu et al., 2015) are valuable for inferring the unique roles of specific brain regions in cognitive processes but are difficult to replicate. The emergence of non-invasive brain stimulation techniques such as Transcranial Magnetic Stimulation (TMS) and tDCS provides opportunities to investigate causal relationships between brain and decision-making behavior (荣悦彤等, 2019). For example, Zheng et al. (2017) used tDCS to enhance right DLPFC excitability and found no effect on trust behavior in trust games. Only one study has combined model-based fMRI with tDCS, finding that intention inference and feedback processing in trust decisions involve two separable neural systems (Nihonsugi et al., 2015). Future research should combine non-invasive brain stimulation, fMRI, and computational modeling within the trust game framework to further reveal causal relationships between the psychological mechanisms and neural foundations of trust formation.

4.3 Interpersonal Trust Research in Psychiatric Populations

Recent developments in computational neuroscience have promoted the application of computational models in clinical research, giving rise to a new field: computational psychiatry (Huys et al., 2011; Montague et al., 2012; Stephan & Mathys, 2014). Computational psychiatry uses model-based quantitative met-

rics to infer hidden causes of abnormal behavior and neural activity, thereby explaining psychopathology.

Previous studies have found that psychiatric patients exhibit abnormal trust behavior in trust games. For example, patients with Borderline Personality Disorder (BPD) and children with Autism Spectrum Disorder (ASD) show reduced trust behavior (King-Casas et al., 2008; Knoch et al., 2009; Maurer et al., 2018), while adolescent depression patients show excessive trust and adult depression patients show reduced trust (Mellick et al., 2019; Wehebrink et al., 2018). However, these studies only identified abnormal trust behavior without clarifying the psychological processes and neural mechanisms underlying abnormal trust decisions. Only one computational modeling study has examined trust games in psychiatric populations. Using a Theory of Mind-based Bayesian model, this study found that BPD patients as investors showed different thinking-depth distributions compared to healthy controls, suggesting that neural response patterns corresponding to thinking depth derived from such models could serve as objective markers for identifying abnormal trust behavior (Xiang et al., 2012). Future research should investigate abnormalities in trust formation processes in psychiatric populations from a computational psychiatry perspective. Combining trust games, computational modeling, and neuroimaging can not only deepen our understanding of normal trust formation (Sanfey, 2007) but also provide new perspectives for studying social dysfunction in mental disorders.

References

Anderhub, V., Engelmann, D., & Güth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization*, 48(2), 197-216.

Ben-Ner, A., Putterman, L., & Ren, T. (2011). Lavish returns on cheap talk: Two-way communication in trust games. *Journal of Socio-Economics*, 40(1), 1-13.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122-142.

Blue, P. R., Hu, J., Peng, L., Yu, H., Liu, H., & Zhou, X. (2020). Whose promises are worth more? How social status affects trust in promises. *European Journal of Social Psychology*, 50(1), 189-206.

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

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience*, 13(6), 637-647.

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models

in social neuroscience. *Computational Models of Brain and Behavior*, 229-244.

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

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752), 2.

Cochard, F., Nguyen Van, P., & Willinger, M. (2004). Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization*, 55(1), 31-44. doi:10.1016/j.jebo.2003.07.004

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2), 199-204.

Declerck, C. H., Boone, C., & Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain and Cognition*, 81(1),

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 Computational Biology*, 10(9).

Engelmann, J. B. (2010). Measuring Trust in Social Neuroeconomics: a Tutorial. *Hermeneutische Blätter*, 1(2), 225-242.

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

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience*, 35(21), 8170-8180. doi:10.1523/JNEUROSCI.4775-14.2015

Fehr, E., & Schmidt, K. M. (2005). The economics of fairness, reciprocity and altruism–Experimental evidence and new theories. *Economics*, 20, 51.

Fett, A. K., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2014). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience*, 9(4), 395-402. doi:10.1093/scan/nss144

Fouragnan, E. (2013). The neural computation of trust and reputation. *Biochemistry*, 52(29), 4941-54.

Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Review*, 68,

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers Humman Neuroscience*, 7, 598. doi:10.3389/fnhum.2013.00598

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2),

Gonzalez, B., & Chang, L. J. (2019). Computational models of mentalizing. *PsyArXiv*. doi:doi:10.31234/osf.io/4tyd9

Gu, X., Wang, X., Hula, A., Wang, S., Xu, S., Lohrenz, T. M., . . . Montague, P. R. (2015). Necessary, yet dissociable contributions of the insular and ventromedial prefrontal cortices to norm adaptation: Computational and lesion evidence in humans. *Journal of Neuroscience*, 35(2), 467-473. doi:10.1523/JNEUROSCI.2906-14.2015

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, 24, 92-97.

Haiyan, L. (2018). Dynamic trust game model between venture capitalists and entrepreneurs based on reinforcement learning theory. *Cluster Computing*. doi:10.1007/s10586-017-1666-x

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, 11(6).

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? , *Neural Networks*, 24(6), 544-551.

Jaafra, Y., Laurent, J. L., Deruyver, A., & Naceur, M. S. (2019). Reinforcement learning for neural architecture search: A review. *Image and Vision Computing*, 89, 57-66.

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

Jung, W. M., Lee, Y. S., Wallraven, C., & Chae, Y. (2017). Bayesian prediction of placebo analgesia in an instrumental learning model. *PLoS One*, 12(2), e0172609. doi:10.1371/journal.pone.0172609

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1995). Partially observable Markov decision processes for artificial intelligence. Paper presented at the International Workshop on Reasoning with Uncertainty in Robotics.

Keser, C. (2003). Experimental games for the design of reputation management systems. *IBM Systems Journal*, 42(3), 498-506.

Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., & Rao, R. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, 5(11), eaax8783.

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.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, 321(5890), 806-810. doi:10.1126/science.1156902

Knoch, D., Schneider, F., Schunk, D., Hohmann, M., & Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences*, 106(49),

Kreps, D. M. (1990). Corporate culture and economic theory. *Perspectives on Positive Political Economy*, 90, 109-110.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., . . . Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences*, 104(50), 20084-20089.

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust Trends drawn from neuroscience, psychology, and economics. *Neurosciences*, 42(2), 92-101.

Lee, D., Seo, H., & Jung, M.(2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, 35, 287-308.

Luo, Y., Hétu, S., Lohrenz, T., Hula, A., & Ramey, C. (2018). Early childhood investment impacts social decision-making four decades later. *Nature Communications*, 9(1).

expectation modulates feedback-related negativity in the Trust Game. *Plos One*, 10(2), e0119129-.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5, 39.

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

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

Mellick, W., Sharp, C., & Ernst, M. (2019). Depressive adolescent girls exhibt atypical social decision-making in an iterative trust game. *Journal of Social and Clinical Psychology*, 38(3), doi:10.1521/jscp.2019.38.2.224

Montague, P. R. (2018). Computational Phenotypes Revealed by Interactive Economic Games. In A. Anticevic & J. D. Murray (Eds), *Computational psychiatry: Mathematical modeling of mental illness* (pp. 273-292): Academic Press

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1), 72-80.

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R., & Friston, K. (2014). A formal model of interpersonal inference. *Frontiers in Human Neuroscience*, 8, 160.

Nihonsugi, T., Ihara, A., & Haruno, M. (2015). Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *Journal of Neuroscience*, 35(8), 3412-3419. doi:10.1523/JNEUROSCI.3885-14.2015

O' Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104(1), 35-53.

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap. *Topics in Cognitive Science*, 11(2), 338-357.

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, 2017, 7(1):1-21.

Premack, D., & Woodruff, G. (1978). Does a chimpanzee have a theory of mind. *Behavioral & Brain Sciences*, 1(4), 515-526.

Puterman, M. L. (1995). Markov decision processes: Discrete stochastic dynamic programming. *Journal of the Operational Research Society*, 46(6), 792-792.

Radell, M. L., Sanchez, R., Weinflash, N., & Myers, C. E. (2016). The personality trait of behavioral inhibition modulates perceptions of moral character and performance during the trust game: Behavioral results and computational modeling. *PeerJ*, 4, e1631. doi:10.7717/peerj.1631

Ray, D., King-Casas, B., Montague, P. R., & Dayan, P. (2009). Bayesian model of behaviour in economic games. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou. (Eds), *Advances in neural information processing systems 21* (pp. 1345-1352).

Riedl, R., & Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics*, 5(2), 63-91. doi:10.1037/a0026318

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

Rusch, T., & Gläscher, J. (2019). Classification of Theory of Mind tasks and their computational models. *PsyArXiv*, 7. doi:10.31234/osf.io/uf85z

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science*, 318.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Kronbichler, M., & Friston, K. (2015). Evidence for surprise minimization over value maximization

in choice behavior. *Scientific Reports*, 5, 16575. doi:10.1038/srep16575

Smith, R., Khalsa, S. S., & Paulus, M. P. (2019). An Active Inference Approach to Dissecting Reasons for Nonadherence to Antidepressants. *Biological Psychiatry:Cognitive Neuroscience and Neuroimaging*. doi:10.1016/j.bpsc.2019.11.012

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019). Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews*, 107, 473-491. doi:10.1016/j.neubiorev.2019.09.002

Snijders, C., & Keren, G. (2001). Do You Trust? Whom Do You Trust? When Do You Trust? , *Advances in Group Processes*, 18(18), 129-160.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, 25, 85-92.

Tzieropoulos, H. (2013). The Trust Game in neuroscience: A short review. *Social Neuroscience*, 8(5), 407-416. doi:10.1080/17470919.2013.832375

Wehebrink, K. S., Koelkebeck, K., Piest, S., de Dreu, C. K. W., & Kret, M. E. (2018). Pupil mimicry and trust - Implication for depression. *Journal of Psychiatric Research*, 97, 70-76. doi:10.1016/j.jpsychires.2017.11.007

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-

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 Computtational Biology*, 8(12), e1002841. doi:10.1371/journal.pcbi.1002841

Enhancing the activity of the DLPFC with tDCS alters risk preference without changing interpersonal trust. *Frontiers in Neuroscience*, 11, 52.

Note: Figure translations are in progress. See original paper for figures.