Trends in Neurosciences

CellPress
REVIEWS

Opinion

Toward a Model of Interpersonal Trust Drawn from Neuroscience, Psychology, and Economics

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

Trust pervades nearly every social aspect of our daily lives, and its disruption is a significant factor in mental illness. Research in the field of neuroeconomics has gained a deeper understanding of the neuropsychoeconomic (NPE) underpinnings of trust by combining complementary methodologies from neuroscience, psychology, and economics. However, a coherent model of trust that integrates separate findings under a conceptual framework is still lacking. Here, we sketch out an integrative NPE model that explains how the interactions of psychoeconomic components engage domain-general large-scale brain networks in shaping trust behavior over time. We also point out caveats of current research approaches and outline open questions that can help guide future transdisciplinary investigations for a better understanding of the neuropsychology of trust.

Highlights

Human societies are unique in the level to which interpersonal trust penetrates every facet of our private and public social lives.

Theoretical and empirical work has made tremendous strides over the last decade in investigating the neuropsychology of interpersonal trust, but a conceptual framework integrating separate research findings into a neuropsychological model of trust is still lacking.

A neuropsychoeconomic framework – combining complementary methodologies from the fields of economics, psychology, and neuroscience – can help to assimilate findings across behavioral, psychological, and neural levels.

An integrative model of interpersonal trust is proposed that explains how the interactions of psychoeconomic components engage domain-general large-scale brain networks in shaping trust behavior over time.

As the transdisciplinary trust research matures, the proposed framework and model might help to guide future investigations to overcome current research limitations toward a better understanding of the neuropsychological underpinnings of trust.

To Trust or Not to Trust: That Is the Question

Trust is a crucial component of cooperative, mutually beneficial interpersonal relationships, penetrating all human **social interactions** (see Glossary) across all facets of private and public social lives. When we trust each other, society is more inclusive and open, economic development is furthered, and feelings of well-being flourish [1]. At the same time, however, trust relationships are unstable and portray a **social dilemma**. Trusting another person is associated with **uncertainty**, which gives rise to the prevalence of deceivers and cheaters in human society. Further, in mental illness such as schizophrenia or borderline personality disorder, the ability to develop and maintain trust is often impaired [2,3].

Scholars from a range of academic fields, including economists, psychologists, and more recently neuroscientists, have investigated interpersonal trust both theoretically and empirically. Although a plethora of definitions for the concept of trust exists, the identification of common psychological elements across definitions allows formulating a working definition of this phenomenon [4]. Interpersonal trust encompasses one's willingness to accept vulnerability based on the expectation regarding the behavior of another party that will produce some positive outcome in the future. The neuropsychological mechanisms of interpersonal trust have been investigated over the last decade, but an overarching conceptual framework that integrates separate findings into a neuropsychological model of trust is still missing.

The objective of this Opinion is twofold. First, we present a neuropsychoeconomic (NPE) model of interpersonal trust, which provides a more integrative picture compared with previous relatively descriptive neuroscience reviews [5–8] and functional neuroimaging **coordinate-based meta-analyses** [9,10]. We base our model on a framework that integrates research

[1]School of Systems Biology, George Mason University, Fairfax, VA, USA
[2]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Mannheim, Germany

*Correspondence:
fkrueger@gmu.edu (F. Krueger).

findings from the growing field of neuroeconomics – a joined effort of economists, psychologists, and neuroscientists – applying economic exchange **games** (e.g., trust game) to integrate psychological systems (i.e., **motivation**, affect, and **cognition**) with neuroscience mechanisms (e.g., brain circuits, hormones/neurotransmitters, and genes) [9] (Figure 1). Second, we point out limitations in the current research approaches, and we outline open questions that can help guide future transdisciplinary investigations toward a better understanding of the neuropsychological underpinning of trust, including not only interpersonal but also institutional and intercultural trust.

The NPE Model of Trust

Our model of trust is rooted in an integrative NPE framework – based on methodologies from the fields of behavioral economics, social psychology, and social cognitive and affective neuroscience – to integrate research findings across behavioral, psychological, and neural levels (Figure 2, Key Figure).

Behavioral Level

The trust game – taken from **game theory** – measures fundamental features of trust in reciprocity with real monetary consequences in a laboratory setting, combining the benefits

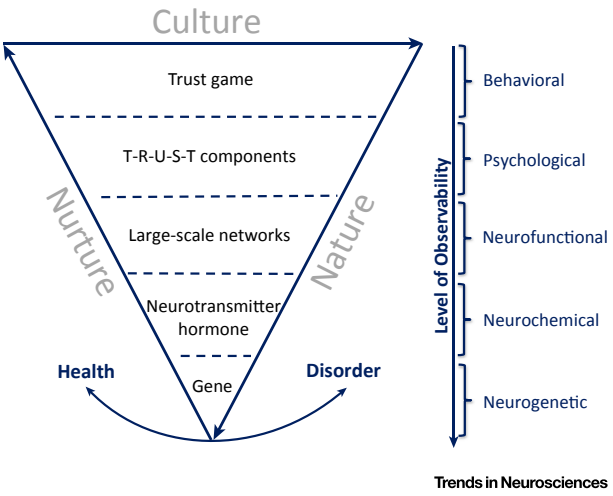

Figure 1. Neuropsychoeconomic Framework of Trust. The synergy of methodologies from economics, psychology, and neuroscience allows integrating knowledge across different levels – economic behaviors (i.e., trust game), psychological systems (i.e., motivation, affect, and cognition), and neural mechanisms (i.e., brain circuits, hormones/neurotransmitter, and genes) – into a framework of interpersonal trust. In a top–down triangle manner, the explanation levels vary from high (behavior) to low (gene) observability and are shaped by a dynamic interplay of culture, nurture, and nature. At the behavioral level, the two-person reciprocal trust game enables to measure both the propensity and dynamics of interpersonal trust behavior. At the psychological level, psychometric and survey measures allow evaluating the psychological systems (i.e., motivation, affect, and cognition) and their linked T-R-U-S-T components (Treachery, Reward, Uncertainty, Strategy, and Trustworthiness). At the neurofunctional level, complementary neuroimaging methods (e.g., functional magnetic resonance imaging, electroencephalography, and focal brain lesions) identify the domain-general large-scale brain networks (i.e., activation and connectivity patterns) shaping the psychoeconomic components of trust behavior. At the neurochemical level, pharmacological manipulations of neuropeptide hormones [e.g., oxytocin (OT)] and steroid hormones (e.g., testosterone) as well as neurotransmitters (e.g., dopamine) reveal the neural signaling pathway mechanisms invoked in trust behavior. At the neurogenetic level, twin and genetic studies looking at individual variations in the human genome and variants of single-nucleotide polymorphisms (e.g., OT receptor gene) explain mechanisms of heritability and genetic variation in producing individual differences in trust behavior. Identifying those patterns of interpersonal trust in healthy people – economic behaviors, psychological systems, and neural mechanisms – can potentially shed light on trust impairment as recognized in the neuropathology of mental disorders.

Glossary

Bounded rationality: limited rationality (e.g., accessible information, cognitive resources, and available time) for making an optimal decision.
Calculus-based trust: trust based on the rational calculation of the costs and benefits of others breaking or maintaining an interdependent relationship.
Central-executive network: a large-scale network of brain regions that form an integrated system for externally directed cognitive functions (e.g., cognitive control, executive functions, and working memory).
Cognition: state of processes such as thinking, planning, and acting.
Cognitive control: processes that allow information processing and behavior to vary adaptively (instead of remaining rigid and inflexible) depending on contextualized goals.
Coordinate-based meta-analysis: analysis of the distribution of coordinates from neuroimaging studies to identify brain regions that are consistently activated during a given experimental task.
Default-mode network: a large-scale network of brain regions that form an integrated system for internally directed cognitive functions (e.g., autobiography, self-monitoring, and social cognition).
Economic rationality: extrinsic motivation to pursue self-regarding interests by cooperating readily when self-interest and collective interest coincide to reap personal benefits from the group.
Electroencephalography: noninvasive neuroimaging method recording electrical signal detected by electrodes placed at different points of the scalp.
Ethnography: the systematic study of peoples and cultures with their customs, habits, and mutual differences, where researchers observe society from the point of view of the people involved in the study.
Functional magnetic resonance imaging: noninvasive neuroimaging method measuring hemodynamic (blood oxygen level dependent) response based on the difference between oxyhemoglobin and deoxyhemoglobin levels in the brain

of quantifiability and replicability across studies [11,12] (Box 1). This two-person reciprocal exchange game represents a social dilemma, where one party (trustor) is willing to be vulnerable to the risk of **treachery** (affect) based on the expectations (cognition) that the action of another party (trustee) will produce some anticipated **reward** (motivation) due to reciprocity in the future.

Psychological Level

Trust, we argue, emerges through the interplay of components represented by the acronym T-R-U-S-T: Treachery, Reward, Uncertainty, Strategy, and Trustworthiness. These components are linked to the following psychological systems: motivation, affect, and cognition. The anticipation of reward (motivation) contrasted with the risk of treachery (affect) creates uncertainty, which is associated with the vulnerability of trusting another person. To reduce uncertainty, two different types of **bounded rationality** (cognition) can be employed: **economic rationality** and **social rationality** [7]. If trust is motivated by extrinsic incentives (i.e., self-regarding interest), it becomes an economically rational choice, pursuing self-interest but trusting readily when self-interest coincides with collective interest (e.g., long-term cooperation, reputation building). The trustor is economically motivated to adopt a **strategy** to reap context-based benefits, thereby removing uncertainty by transforming economic risk of treachery (i.e., losing monetary stakes) to the extrinsically positive expectation of reciprocity. If trust is motivated by intrinsic incentives (i.e., other-regarding interest), it becomes a socially rational choice, contributing to the relationship success and valuing group belonging. The trustor is socially motivated to evaluate **trustworthiness** to promote relationship-based benefits, thereby removing uncertainty by transforming social risk of treachery (i.e., being betrayed by the partner) to intrinsically positive expectations of reciprocity.

Neural Level

Based on a systems neuroscience view [13], trust arises from the interactions of psychological systems (i.e., motivation, affect, and cognition) that engage key regions anchored in domain-general **large-scale brain networks**: **reward network** (RWN), salience network (SAN), **central-executive network** (CEN), and **default-mode network** (DMN). The motivational system of trust involves the RWN to determine the anticipated reward for trusting another person. The affective system of trust engages the SAN to incorporate aversive feelings associated with risk of treachery by another person. The cognitive system of trust involves, on the one hand, the CEN (i.e., **cognitive control** system) to adopt context-based strategies, and on the other hand, the DMN (i.e., **social cognition** system) to evaluate relationship-based trustworthiness for trusting a partner.

RWN – Anticipation of Reward

The motivational system is anchored in the RWN that builds on dopaminergic pathways: The mesolimbic pathway connects the ventral tegmentum area (VTA) in the midbrain to the nucleus accumbens and olfactory tubercle in the ventral striatum (vSTR); the mesocortical pathway – the VTA to the prefrontal cortex (PFC), including the ventromedial PFC (vmPFC); and the nigrostriatal pathway – the substantia nigra in the midbrain to the caudate nucleus and putamen in the dorsal striatum (dSTR) [14].

Both meso-dopaminergic pathways are commonly involved in forming anticipation of reward to forecast positive and negative consequences of available options for guiding adaptive social behavior under uncertainty [15]. As part of the mesocortical pathway, the vmPFC is ideally located – due to its interconnectivity with the vSTR and amygdala – to act as a neural integrator. It has been consistently associated with encoding the expected **utility** of stimuli, combining

arising from changes in local blood flow.

Game: in the context of mathematics – a well-defined object of an abstract version of a real-world decision situation, including players of the game, information and actions available to each player, and payoffs for each outcome at each decision point.

Game theory: branch of applied mathematics providing tools for analyzing situations in which players make strategic decisions.

Identification-based trust: trust based on positive emotions for a deeper understanding and identification with others.

Knowledge-based trust: trust based on acquired knowledge about others' motives, intentions, and behavioral tendencies.

Large-scale brain network: collection of widespread interconnected brain regions across the entire brains that interact to perform circumscribed functions.

Motivation: state in which rewards are sought and punishments are avoided.

Reinforcement learning: learning best-action patterns based on reward or punishment that strengthen a person's future behavior whenever it is preceded by a specific stimulus.

Reward: attractive and motivational property of a stimulus that induces approach/consummatory behavior.

Reward network: a large-scale network of brain regions that form an integrated system for motivation (e.g., desire, craving for a reward), associative learning (e.g., positive reinforcement, classical conditioning), and positive emotions involving pleasure (e.g., joy, euphoria, and ecstasy).

Reward prediction error: phasic activity of dopaminergic neurons in the midbrain, signaling a discrepancy between the predicted and currently experienced reward of an event.

Social cognition: state of processes on how people assimilate, store, and employ information about other people and social situations.

Social dilemma: situation involving a conflict between immediate self-interest and longer-term collective interests of counterparts, where a counterpart's immediate

complex and qualitatively different reward alternatives on a common currency of subjective value [16]. Abnormal judgment and decision making within social contexts are consistently associated with vmPFC damage [17], leading to increased trust behavior [18].

As part of the mesolimbic pathway, the vSTR regulates motivation for rewarding stimuli and facilitates **reinforcement learning** [19,20]. When people learn to trust their partners during repeated interactions, the vSTR encodes a **reward prediction error** signal (i.e., a signal of reciprocity) based on dopamine neurons firing in response to the magnitude in comparing the expected and actual reward [21,22]. This mechanism provides a fundamental brain mechanism by which human trust relationships are initiated and sustained [23].

The nigrostriatal pathway (as part of the basal ganglia motor loop) is involved in the production of movement, thus influencing planning and action selection [24]. The dSTR (i.e., caudate nucleus), as a key brain region of the nigrostriatal pathway, plays a critical role in interaction-based learning when no prior information about the trustee is available [25]. A ventral–dorsal dissociation within the STR has been shown to dissociate the trust and feedback stages during repeated trust games. The dSTR is consistently activated when learning about the partner's reciprocity of trust during the feedback stage, whereas the vSTR is activated when anticipating the reward during the trust decision stage [10].

SAN – Risk of Treachery

The affective system is anchored in the SAN – including crucial regions such as the amygdala, anterior insula (AI), and dorsal anterior cingulate cortex (dACC) – consistently implicated in self-related bottom–up saliency detection for regulating social behavior [13].

The amygdala is necessary for appropriate social functioning [26]. It signals, among other things, the threat of treachery based on encoding emotional salience and promoting social vigilance [27]. Damage to the amygdala leads to increased trust [28,29], supporting its role in evaluating incoming social information, to either enhance trust-related behaviors for positive evaluations or to distrust the individual for negative evaluations – consistent with the literature on the opposite effects of the two hormones OT and testosterone (TE) in balancing trust (Box 2).

Based on a posterior-to-anterior remapping of interceptive signals within the insular cortex, the AI encodes subjective aversive feeling states of unpredictable events – supporting its reliable role in encoding a common currency of aversion [30]. The AI signals aversion of treachery while trusting another partner [31] and its damage results in misplaced trust [32]. Decisions to trust engage the right dorsal AI, whereas decisions to reciprocate engage the right ventral AI [9]. Acting as a hub for multimodal functional integration, the right dorsal AI contributes to dynamically switching between large-scale brain networks, including the CEN (i.e., externally directed cognition) and the DMN (i.e., internally directed cognition) [33].

Lastly, the dACC is persistently implicated in conflict monitoring for both nonsocial and social domains [34] and serves to identify conflicts between brain networks for a successful social adaptation [7]. Trustors, when interacting iteratively with an untrustworthy partner, show higher dACC activity [35], supporting this brain region's role in monitoring the trustor's social dilemma.

CEN – Adoption of Strategy

The cognitive control system is anchored in the CEN – comprising the dorsolateral PFC (dlPFC) and the ventrolateral PFC (vlPFC) – which has been consistently associated with top–down cognitive control in adopting goal-directed behavior under changing contexts [36].

self-interest is tempting, but all counterparts benefit from acting in the longer-term collective interest.
Social interaction: exchange by which people react to others and act based on rules, systems, and institutions.
Social rationality: intrinsic motivation to pursue other-regarding interests by contributing to group success and valuing group belonging.
Strategy: higher-level plan designed to achieve a long-term or overall goal.
Treachery: betrayal or violation of trust by another person.
Trustworthiness: belief in others' perceived ability (e.g., possession of a skill), benevolence (e.g., quality of being kind), and integrity (e.g., quality of having strong moral principles).
Uncertainty: the property of a situation involving insecurity and/or unknown information, such as regarding the prediction of future events.
Utility: subjective value placed on some goods or actions, which emerges through comparing different reward options to generate a decision.

Key Figure

Neuropsychoeconomic Model of Trust

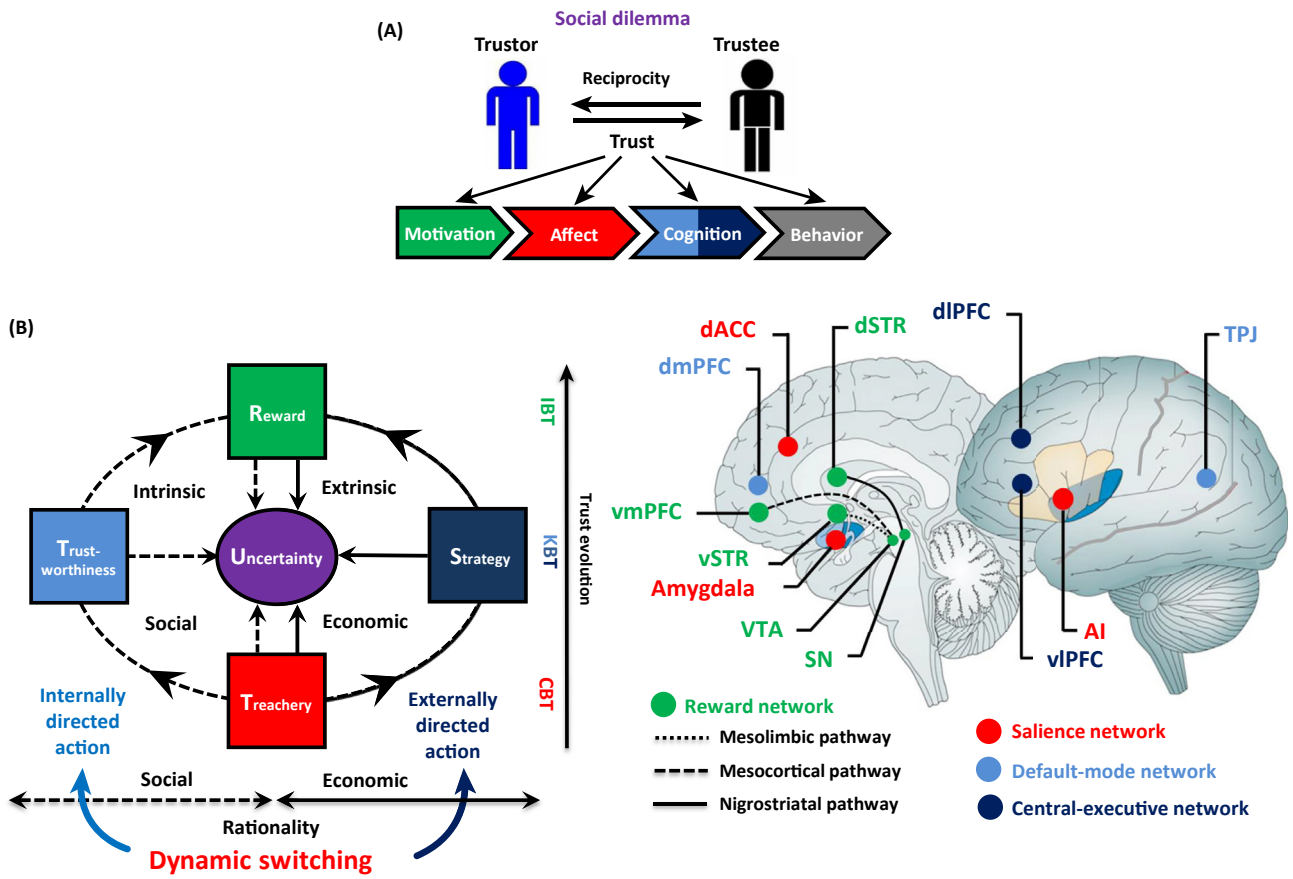

Figure 2. (A) Trust definition. Trust in reciprocity (behavior, gray) represents a social dilemma that encompasses a trustor's willingness to be vulnerable to the risk of treachery (affect, red) based on the expectations (cognition, blue) that the action of a trustee will produce some anticipated reward (motivation, green) due to reciprocity in the future. (B) Trust formation. Trust arises through the interplay of T-R-U-S-T components (treachery, reward, uncertainty, strategy, and trustworthiness) – linked to psychological systems (i.e., motivation, affect, and cognition) – that engage key brain regions (closed circles) anchored in domain-general large-scale brain networks. Trust emergence. The anticipation of reward (green rectangle, motivation, reward network, RWN) contrasted with the risk of treachery (red rectangle, affect, salience network, SAN) creates uncertainty (purple ellipse), which is associated with vulnerability of trusting another person. To remove uncertainty, two types of bounded rationality (cognition) can be employed, where SAN acting as a switch engages either the central-executive network (CEN, externally directed cognition) or the default-mode network (DMN, internally directed cognition). Trustors with extrinsic incentives can adopt a context-based strategy (dark blue rectangle; cognitive control, CEN) to reap personal benefits (i.e., economic rationality), thus removing uncertainty by transforming economic risk of treachery to economically positive expectations of reciprocity (unbroken lines). Trustors with intrinsic incentives can evaluate the relationship-based trustworthiness (light blue rectangle; social cognition, DMN) to contribute to the relationship's success (i.e., social rationality), hence removing uncertainty by transforming social risk of treachery to socially positive expectations of reciprocity (broken lines). Trust evolution. In a calculus-based trust relationship, trustors (encountering ambiguous situations) perform rational calculations of the costs and benefits of creating and sustaining a relationship – preeminently driven by SAN (risk of treachery). In a knowledge-based trust relationship, trustors (facing uncertain situations) acquire knowledge about the contexts and their partners to predict trustees' behaviors accurately to advance their trust relationships – primarily driven by CEN (adoption of strategy) and DMN (evaluation of trustworthiness). In an identification-based trust relationship, trustors (confronted with certain situations) develop a rewarding identification and understanding with trustees to confidently trust them – notably driven by RWN (anticipation of reward). Figure adjusted and reprinted with permission from Macmillan Publishers Ltd [66]. AI, anterior insula; dACC, dorsal anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; dSTR, dorsal striatum; SN, substantia nigra; TPJ, temporoparietal junction; vlPFC, ventrolateral prefrontal cortex; vmPFC, ventromedial prefrontal cortex; vSTR, ventral striatum; VTA, ventral tegmentum area.

Box 1. The Sequential Two-Person Reciprocal Trust Game

In the standard trust game, two players, anonymous to each other, receive an initial endowment for economic exchange and are assigned to the role of either a trustor or trustee [11,12] (Figure I). The game consists of three sequential stages: (i) trust, (ii) reciprocity, and (iii) feedback. During the trust stage, the trustor decides either not to pass an endowment (distrust) or to pass any portion of the endowment (trust) to the trustee. The trustor keeps the remainder of the endowment. The shared money is then multiplied (usually tripled) by the experimenter and passed on to the trustee. During the reciprocity stage, the trustee decides to pass back to the trustor either nothing (treachery) or any portion of the money received (reciprocity). Lastly, during the feedback stage, the trustor learns about the trustee's decision. The amount of money passed by the trustor captures trust, whereas the amount of money reciprocated by the trustee captures trustworthiness. The trustor's final payoff equals the initial endowment minus the transfer to the trustee, plus the back transfer from the trustee. The trustee's final payoff equals the initial endowment plus the tripled transfer of the investor, minus the back transfer to the investor. Cooperation occurs when trustor and trustee act in a manner that mutually benefits both players. When the trustor sends money and the trustee honors the trust by sending some money back, both players end up with a higher monetary payoff than the original endowment. The reciprocation of trust depends on the offset between maximizing the trustee's outcomes relative to the appreciation of the trust that was given by the trustor. The standard economic solution to this game uses backward induction and predicts that a rational and selfish trustee has a strong incentive to keep all the money and repay none to the trustor, therefore, never honors the trust given by the trustor. Realizing this, the trustor should never place trust in the first place and so will invest zero in the transaction. Despite these grim theoretical predictions, most trustors invest more than half of the endowment and trustees often split the sum of money evenly [48]. The one-round version of the game measures trust propensity, whereas the multiround version measures trust dynamics (e.g., building, maintenance). A risk game (i.e., lottery) is often used as a control condition to separate social decision making under nonsocial risk (gamble) from social risk (trust), where the trustor interacts with a computer (i.e., uncertainty of a random process) instead of a human counterpart (i.e., uncertainty of a social partner).

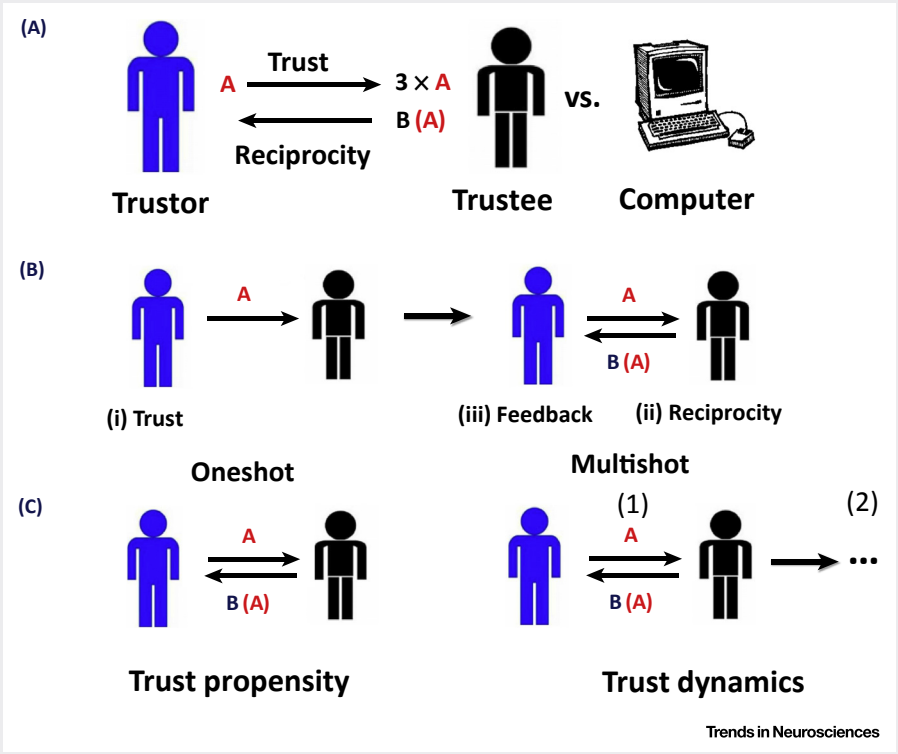

Figure I. Trust Game. (A) Structure, (B) phases, and (C) types of the sequential two-person reciprocal exchange game.

Box 2. Oxytocin and Testosterone Acting as Antagonists of Trust

As a hormonal promoter of trust, the neuropeptide hormone oxytocin (OT) – synthesized in the paraventricular and supraoptic nuclei of the hypothalamus – is released to the brain and the periphery [49]. Exogenous OT administration affects social cognition and behavior, generally regulating lower-level processes that respond to the salience of social cues, anxiety reduction, and facilitation of approach and affiliative behaviors [50]. Studies examined the effects of OT on interpersonal trust via exogenous OT administration, endogenous OT plasma levels, and genetic polymorphisms of the OT receptor gene [51]. For example, OT increases trust behavior while risk preferences remain unchanged [52]. OT can make individuals susceptible to betrayal: Trustors may continue knowingly trusting an untrustworthy person under OT influence, resulting in reduced activity in the salience (amygdala, AI) and reward (striatum) networks [53]. OT, however, does not uniformly facilitate trust; it can also cause distrust, depending on early life experience, social repertoire, social context, and altered functioning of the OT system in psychiatric disorders [50]. Findings are mixed for associating increased trust behavior with endogenous OT plasma levels [54,55], twin and human genome studies [56,57], and genetic polymorphisms of the OT receptor gene [58,59]; thus, higher methodological standards and larger samples are needed to enhance the robustness of OT research [51]. OT's antagonist and inhibitor in the context of interpersonal trust is the steroid hormone TE, dominantly synthesized in the gonads by the Leydig cells in testes in men and by the ovaries in women [60]. TE has a modulatory effect on the brain and social behavior associated with competition and dominance [61]. Exogenous TE administration decreases trust behavior but increases generosity behavior when reciprocating trust [62]. For trust, it mediates antisocial (competitive, potentially aggressive) behavior when social threats need to be considered to better prepare oneself for competition over status and valued resources [63]. For reciprocity, it facilitates prosocial behavior in the absence of social threats when good reputation (or high status) needs to be considered [62]. People are more trusting in iterated than in one-time interactions. However, this effect disappears after TE administration in people with relatively high prenatal TE exposure (as measured via second-to-fourth digit ratio), indicating that TE moderates the effect of the social environment on trust behavior [64]. Acting on vasopressinergic neurons in the amygdala, TE probably reduces trust by inducing a sustained decoupling between the amygdala and vmPFC via prefrontal–dopaminergic mechanisms that result in more vigilant responses of the amygdala to social threats in uncertain situations [65].

The dlPFC provides the cognitive capacity to remove uncertainty by accounting evidence contextually through modulation of bottom–up processes. When no priors about the trustee are available, the dlPFC responds differentially when learning to trust cooperative counterparts compared with individualistic ones [37]. Over the course of iterated interactions, it keeps track of contextually modulated decisions, thereby enhancing goal-directed behavior and improving the long-term outcome.

The vlPFC grants the cognitive capacity to reduce and eliminate uncertainty by discounting evidence contextually through modulation of bottom–up processes. The vlPFC disrupts the impact on learning of the dSTR after violations of trust when priors about the trustee are present [25]. This region maintains choices anchored with reliable prior beliefs [38], and therefore, prevents an unnecessary retaliation after a violation of trust and favors social stability [25].

DMN – Evaluation of Trustworthiness

The social cognition system is anchored in the DMN, including crucial brain regions such as the temporoparietal junction (TPJ) and dorsomedial PFC (dmPFC) [13]. The DMN has been consistently identified in the context of mentalizing about others to facilitate cooperative decision making [39].

The TPJ has been linked to various social cognitive functions, including self-other distinction, perspective taking, and intentional inferences of others [40], making it an essential region for inferring and attributing the intentions of others to evaluate relationship-based trustworthiness. Trustors with higher perspective-taking tendencies show not only greater trust toward others but also reduce their trust more drastically after treachery by their counterparts [41]. TPJ activity increases with age when continuously trusting another person, indicating a higher sensitivity and orientation toward other people's social signals [35]. Sophisticated trustors show higher TPJ activity than naive ones, consistent with the assumption that sophisticated trustors build

better mental models about the intentions of their partners – models that build not only on what trustees reciprocate but also on what they expect from the trustors regarding initial investments [42].

The dmPFC proves to be critical for self-referential processing and forming impressions of others – being activated not only in social 'offline' tasks (e.g., social judgment paradigms) but also in 'online' ones (e.g., interactive games) [40]. This region is engaged in inferring and attributing the traits of others to evaluate a partner's trustworthiness: Higher dmPFC activity is observed when trustors play against human compared with computerized opponents [43]. The dmPFC evaluates partner's trustworthiness not only based on iterative interactions, but also based on priors conveying information about the social characteristics of partners – consistent with its role in ascribing traits to others for anticipating their decisions [25].

Evolution of Trust

The NPE model describes how interpersonal trust evolves through repeated interactions: from **calculus-based trust**, through **knowledge-based trust**, to **identification-based trust** [44]. First, the trust relationship begins with calculus-based trust. Driven primarily by SAN (risk of treachery), trustors encounter ambiguous situations and perform rational calculations of the costs and benefits of creating and sustaining a trust relationship. For example, a shift from the SAN (AI) to RWN (vSTR) activity can be observed when transitioning from one-round to multiround trust game interactions, probably reflecting a shift from the calculus-based trust (guided by uncertainty about the risk of treachery) to identification-based trust (guided by certainty about anticipated reward) [10]. Second, the trust relationship progresses to knowledge-based trust. Driven mainly by CEN (adoption of strategy) or DMN (evaluation of trustworthiness), trustors face uncertain situations and acquire knowledge about the contexts and their partners to predict trustees' behaviors accurately to advance their trust relationships. For example, depending on the trustors' initial type of rationality, that is, social versus economic, DMN (dmPFC) activity reflects whether partners advance from knowledge- to identification-based trust [45]. Indicating this switch, trustors driven more by social than economic rationality show higher DMN (dmPFC) activity during earlier stages – but higher RWN (vSTR) activity during later stages of trust building. Finally, the trust relationship matures to identification-based trust. Driven preeminently by RWN (anticipation of reward), trustors confront certain situations and develop a rewarding identification with trustees and understanding of their motives to trust them confidently.

Concluding Remarks and Future Perspectives

Human societies are unique in the extent to which trust characterizes the development of interpersonal interactions. Theoretical and empirical work has made tremendous strides by applying the trust game to integrate psychological insights with neuroscience mechanisms, and to provide a normative setting for understanding the neuropsychological mechanisms underlying interpersonal trust in laboratory settings. We proposed an NPE model of interpersonal trust that explains how T-R-U-S-T components (Treachery, Reward, Uncertainty, Strategy, and Trustworthiness) engage psychological systems (motivation, affect, and cognition) that recruit domain-general large-scale brain networks (RWN, SAN, CEN, and DMN) in shaping trust over time. However, limitations of current research approaches do exist that future research should address to advance our understanding of the neuropsychology of trust.

One of the remaining challenges is that the trust game only measures a specific type of interpersonal trust where dyads interact in an economic context – reflecting only one aspect of everyday behavior. Not all neuropsychological findings based on the trust game might

necessarily generalize to other real-world scenarios. This is relevant both for different types of 'horizontal' trust interactions (e.g., among family members) and to 'vertical' trust interactions between individuals and those holding institutional positions (e.g., political and organizational ones). Further progress is needed in studying the conjoint psychological function of brain regions working together as large-scale networks in producing trust behavior, beyond mapping psychological functions onto individual brain regions. Emerging evidence, utilizing task-free compared with task-based neuroimaging, indicates that the individual variability in resting-state functional connectivity (based on **functional magnetic resonance imaging** or **electroencephalography**) in large-scale networks (especially DMN) predicts individual differences in trust behavior [46,47].

Furthermore, laboratory and field studies focusing on how culture, nurture, and nature interact to shape trust behavior are necessary. Therefore, larger sample sizes going beyond WEIRD (Western, educated, industrialized, rich, and democratic) populations combined with higher methodological standards are needed to draw relationships from neurotransmitters, hormones, and brain structures to trust behavior. Finally, future investigations about the neuropsychological underpinnings of trust in mental health disorders such as in borderline personality disorder [2] – often characterized by mistrust in social relationships – may help to identify objective biomarkers for disease diagnostic specificity and novel treatment strategies (see Outstanding Questions).

In conclusion, as the transdisciplinary research for neuropsychology of interpersonal trust matures, the framework and model proposed here can hopefully be used to advance our understanding of interpersonal trust and help advocate for a more trusting and inclusive society.

Outstanding Questions

Behavioral level. Can trust be measured with greater ecological validity by combining qualitative methods (e.g., case studies, **ethnography** studies, and interviews) with quantitative methods (e.g., exchange games, scale-based trust surveys, and implicit association of trust assays) to link trust measures from field observations, laboratory experiments, and real-world social interactions?

Psychological level. Apart from 'horizontal' trust among individuals, how do the proposed trust components impact the building, maintenance, and repairing of 'vertical' trust between individuals and those holding institutional positions?

Neurofunctional level. Shifting from standardized univariate analysis techniques (i.e., location-based approach) to more sophisticated multivariate ones (i.e., function-based approach), how is the functional (temporal) and effective (directional) connectivity within and between key regions of domain-general large-scale networks dynamically shaping trust over time organized?

Neurochemical level. What are the causal relationships among exogenous administration and endogenous levels of neuropeptides (e.g., OT, arginine vasopressin), sex hormones (e.g., testosterone, estrogen), and neurotransmitters (e.g., dopamine, serotonin) in modulating interpersonal trust?

Neurogenetic level. In contrast to gene-specific candidate-driven studies focusing on multiple variations of single-nucleotide polymorphisms (e.g., OT receptor gene), can genome-wide association studies identify genome-wide sets of genetic variants associated with interpersonal trust?

Nature/Nurture/Culture. Combining quantitative/molecular genetic studies with trust measures in different cultures, what are the individual and conjointly genetic, environmental, and cultural influences on trust behavior?

Health/Disorder. Applying a computational psychiatry approach, can computational models be built to

References

1. Rothstein, B. and Uslaner, E.M. (2005) All for all – equality, corruption, and social trust. *World Polit.* 58, 41–72
2. Unoka, Z. *et al.* (2009) Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *J. Pers. Disord.* 23, 399–409
3. Fett, A.K. *et al.* (2012) To trust or not to trust: the dynamics of social interaction in psychosis. *Brain* 135, 976–984
4. Seppanen, R. *et al.* (2007) Measuring inter-organizational trust – a critical review of the empirical research in 1990–2003. *Ind. Mark. Manage* 36, 249–265
5. Fehr, E. (2009) On the economics and biology of trust. *J. Eur. Econ. Assoc.* 7, 235–266
6. Tzieropoulos, H. (2013) The trust game in neuroscience: a short review. *Soc. Neurosci.* 8, 407–416
7. Declerck, C.H. *et al.* (2013) When do people cooperate? The neuroeconomics of prosocial decision making. *Brain Cogn.* 81, 95–117
8. Riedl, R. and Javor, A. (2012) The biology of trust: integrating evidence from genetics, endocrinology, and functional brain imaging. *J. Neurosci. Psychol. Econ.* 5, 63–91
9. Bellucci, G. *et al.* (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
10. Bellucci, G. *et al.* (2017) Neural signatures of trust in reciprocity: a coordinate-based meta-analysis. *Hum. Brain Mapp.* 38, 1233–1248
11. Camerer, C. and Weigelt, K. (1988) Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56, 1–36
12. Berg, J. *et al.* (1995) Trust, reciprocity, and social-history. *Games Econ. Behav.* 10, 122–142
13. Bressler, S.L. and Menon, V. (2010) Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* 14, 277–290
14. Ikemoto, S. (2010) Brain reward circuitry beyond the mesolimbic dopamine system: a neurobiological theory. *Neurosci. Biobehav. Rev.* 35, 129–150
15. Knutson, B. *et al.* (2005) Distributed neural representation of expected value. *J. Neurosci.* 25, 4806–4812
16. Hare, T.A. *et al.* (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324, 646–648
17. Koenigs, M. and Tranel, D. (2007) Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *J. Neurosci.* 27, 951–956
18. Moretto, G. *et al.* (2013) Investment and repayment in a trust game after ventromedial prefrontal damage. *Front. Hum. Neurosci.* 7, 593
19. Montague, P.R. and Berns, G.S. (2002) Neural economics and the biological substrates of valuation. *Neuron* 36, 265–284
20. Ruff, C.C. and Fehr, E. (2014) The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* 15, 549–562
21. Delgado, M.R. *et al.* (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* 8, 1611–1618
22. Phan, K.L. *et al.* (2010) Reputation for reciprocity engages the brain reward center. *Proc. Natl. Acad. Sci. U. S. A.* 107, 13099–13104
23. King-Casas, B. *et al.* (2005) Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83

24. Joel, D. and Weiner, I. (2000) Striatal contention scheduling and the split circuit scheme of basal ganglia-thalamocortical circuitry: from anatomy to behaviour. In *Brain Dynamics and the Striatal Complex*, pp. 1–28, CRC Press
25. Fouragnan, E. *et al.* (2013) Reputational priors magnify striatal responses to violations of trust. *J. Neurosci.* 33, 3602–3611
26. Adolphs, R. *et al.* (1998) The human amygdala in social judgment. *Nature* 393, 470–474
27. Engell, A.D. *et al.* (2007) Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* 19, 1508–1519
28. Koscik, T.R. and Tranel, D. (2011) The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia* 49, 602–611
29. van Honk, J. *et al.* (2013) Generous economic investments after basolateral amygdala damage. *Proc. Natl. Acad. Sci. U. S. A.* 110, 2506–2510
30. Namkung, H. *et al.* (2017) The insula: an underestimated brain area in clinical neuroscience, psychiatry, and neurology. *Trends Neurosci.* 40, 200–207
31. Aimone, J.A. *et al.* (2014) Neural signatures of betrayal aversion: an fMRI study of trust. *Proc. Biol. Sci.* 281, 20132127
32. Belfi, A.M. *et al.* (2015) Damage to the insula is associated with abnormal interpersonal trust. *Neuropsychologia* 71, 165–172
33. Uddin, L.Q. (2015) Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* 16, 55–61
34. Bush, G. *et al.* (2000) Cognitive and emotional influences in anterior cingulate cortex. *Trends Cogn. Sci.* 4, 215–222
35. Fett, A.K. *et al.* (2014) Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Soc. Cogn. Affect. Neurosci.* 9, 395–402
36. Miller, E.K. and Cohen, J.D. (2001) An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202
37. Lemmers-Jansen, I.L.J. *et al.* (2017) Boys vs. girls: gender differences in the neural development of trust and reciprocity depend on social context. *Dev. Cogn. Neurosci.* 25, 235–245
38. Souza, M.J. *et al.* (2009) Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. *Neuroimage* 46, 299–307
39. Amodio, D.M. and Frith, C.D. (2006) Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 7, 268–277
40. Van Overwalle, F. (2009) Social cognition and the brain: a meta-analysis. *Hum. Brain Mapp.* 30, 829–858
41. Fett, A.K. *et al.* (2014) Trust and social reciprocity in adolescence – a matter of perspective-taking. *J. Adolesc.* 37, 175–184
42. Xiang, T. *et al.* (2012) Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Comput. Biol.* 8, e1002841
43. McCabe, K. *et al.* (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci. U. S. A.* 98, 11832–11835
44. Lewicki, R. and Bunker, B. (1995) Trust in relationships. *Adm. Sci. Q.* 5, 583–601
45. Krueger, F. *et al.* (2007) Neural correlates of trust. *Proc. Natl. Acad. Sci. U. S. A.* 104, 20084–20089
46. Hahn, T. *et al.* (2015) How to trust a perfect stranger: predicting initial trust behavior from resting-state brain-electrical connectivity. *Soc. Cogn. Affect. Neurosci.* 10, 809–813
47. Bellucci, G. *et al.* (2018) Functional connectivity of specific resting-state networks predicts trust and reciprocity in the trust game. *Cogn. Affect. Behav. Neurosci.* http://dx.doi.org/10.3758/s13415-018-00654-3
48. Johnson, N.D. and Mislin, A.A. (2011) Trust games: a meta-analysis. *J. Econ. Psychol.* 32, 865–889
49. Meyer-Lindenberg, A. *et al.* (2011) Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. *Nat. Rev. Neurosci.* 12, 524–538
50. Bartz, J.A. *et al.* (2011) Social effects of oxytocin in humans: context and person matter. *Trends Cogn. Sci.* 15, 301–309
51. Nave, G. *et al.* (2015) Does oxytocin increase trust in humans? A critical review of research. *Perspect. Psychol. Sci.* 10, 772–789
52. Kosfeld, M. *et al.* (2005) Oxytocin increases trust in humans. *Nature* 435, 673–676
53. Baumgartner, T. *et al.* (2008) Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58, 639–650
54. Zhong, S. *et al.* (2012) U-shaped relation between plasma oxytocin levels and behavior in the trust game. *PLoS One* 7, e51095
55. Zak, P.J. *et al.* (2005) Oxytocin is associated with human trustworthiness. *Horm. Behav.* 48, 522–527
56. Cesarini, D. *et al.* (2008) Heritability of cooperative behavior in the trust game. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3721–3726
57. Reimann, M. *et al.* (2017) Trust is heritable, whereas distrust is not. *Proc. Natl. Acad. Sci. U. S. A.* 114, 7007–7012
58. Krueger, F. *et al.* (2012) Oxytocin receptor genetic variation promotes human trust behavior. *Front. Hum. Neurosci.* 6, 4
59. Apicella, C.L. *et al.* (2010) No association between oxytocin receptor (OXTR) gene polymorphisms and experimentally elicited social preferences. *PLoS One* 5, e11153
60. Zubeldia-Brenner, L. *et al.* (2016) Developmental and functional effects of steroid hormones on the neuroendocrine axis and spinal cord. *J. Neuroendocrinol.* 28, http://dx.doi.org/10.1111/jne.12401
61. Carre, J.M. *et al.* (2014) Testosterone responses to competition predict decreased trust ratings of emotionally neutral faces. *Psychoneuroendocrinology* 49, 79–83
62. Boksem, M.A. *et al.* (2013) Testosterone inhibits trust but promotes reciprocity. *Perspect. Psychol. Sci.* 24, 2306–2314
63. Hareli, S. *et al.* (2009) Emotional versus neutral expressions and perceptions of social dominance and submissiveness. *Emotion* 9, 378–384
64. Buskens, V. *et al.* (2016) Testosterone administration moderates effect of social environment on trust in women depending on second-to-fourth digit ratio. *Sci. Rep.* 6, 27655
65. Bos, P.A. *et al.* (2012) The neural mechanisms by which testosterone acts on interpersonal trust. *Neuroimage* 61, 730–737
66. Blakemore, S.J. (2008) The social brain in adolescence. *Nat. Rev. Neurosci.* 9, 267–277

bridge the explanatory gap between the proposed neuropsychoeconomic basis of trust and the neuropathology that underlies trust impairments in psychiatric diseases?