CAN YOU TRUST A RAT? USING ANIMAL MODELS TO INVESTIGATE THE NEURAL BASIS OF TRUST-LIKE BEHAVIOR

Marijn van Wingerden
*Social Rodent Lab, Heinrich-Heine-University*

Wouter van den Bos
*Center for Adaptive Rationality, Max Planck Institute for Human Development*

The ability to form and maintain long-term trusting relationships with others arguably has enabled humans to develop high levels of cooperation within social groups. Trust, as operationalized in the economic trust game, varies greatly between individuals and under different contexts, probably reflecting contributions from different component processes. From a mechanistic point of view, progress has been made to delineate neural substrates and neurochemical pathways involved in trusting behavior. However, many questions regarding the exact neural implementation of trust remain unanswered. To make progress on these outstanding questions, it would be helpful to be able to turn to an animal model of trust, in which the full range of neurobiological manipulations and readout could be employed. The fundamental question therefore is: can trust be adequately modeled in laboratory animals, such as rodents? Here we present a breakdown of trusting behavior into its component processes such as social recognition, and reward contingency learning, and discuss which of these components could be translated to an animal experiment. We finally present a pilot experiment that indicates that rodents can learn reward probability contingencies that depend on social cues, and that this probability estimation bias persisted even when the actual outcome distributions between social partners was equated, suggesting a rudimentary capability for trust-like behavior in rodents.

This work was supported by grants from the Volkswagen Stiftung (grant 85646 WB & MW, grant 88216 MW). The authors wish to thank Ms. Meng May Wang for help with the rodent experiment, and the other members of the Social Interaction network (Christoph Teufel, Marion Godman, Mog Stapleton, and Vivian Bohl) for discussions.

Address correspondence to Marijn van Wingerden, Institute for Experimental Psychology, Heinrich-Heine-University, Universitätsstrasse 1, D-40225 Düsseldorf, Germany; E-mail: Marijn.Wingerden@uni-duesseldorf.de, or to Wouter van den Bos, Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany; E-mail: vandenbos@mpib-berlin.mpg.de.

© 2015 Guilford Publications, Inc.
A social organism of any sort whatever, large or small, is what it is because each member proceeds to his own duty with a trust that the other members will simultaneously do theirs.

—William James (1897, p. 24)

Despite their bad reputation as all-round nasty animals, bringers of disease, and the disputable honor of having their name used as a label to indicate untrustworthy people, rats are actually highly social animals (Whishaw & Kolb, 2004). Born in litters of about ten (Chahoud & Paumgartten, 2009), infant rats are in constant contact with conspecifics and display a wide range of social behaviors throughout development into adolescents, including reciprocal grooming and play fighting (Pellis & Pellis, 1997)—affiliative behaviors that are associated with appetitive 50 kHz ultrasonic vocalizations (Burgdorf et al., 2008; Knutson, Burgdorf, & Panksepp, 1998), thought to reflect a positive emotional state (Panksepp & Burgdorf, 2000). Traditionally, rodents, but rats in particular, have been used extensively to probe the neural mechanisms underlying economic choice and decision-making processes using behavioral, pharmacological, and electrophysiological methods (Fawcett et al., 2014; Kalenscher & van Wingerden, 2011; Wallis, 2012). Recently, investigators have capitalized on these methods and the social nature of the rat to study decision making and social interactions in rodents in a comparative social neuroscience framework (Ben-Ami Bartal, Decety, & Mason, 2011; Hernandez-lallement, Wingerden, Marx, Srejic, & Kalenscher, 2015; Rutte & Taborsky, 2007a; Viana, Gordo, Sucena, & Moita, 2010). For instance, these studies have shown that rats are capable of exhibiting direct and generalized reciprocity (Rutte & Taborsky, 2007a, 2007b) and possess the cognitive capabilities to recognize a tit-for-tat strategy in a rodent Prisoner’s Dilemma analogue (Viana et al., 2010). Based on these results we hypothesized that the basic capacity for trust could be present in rats as well. To adequately test this hypothesis it is necessary to (1) decompose the cognitive processes involved in trust, and (2) to examine which of those could be adequately modeled in an animal experiment.

In the first part of this article we identify the constituent processes of trust and review the evidence of their existence in rats. Finally, we present a novel behavioral paradigm that illustrates how one could isolate some of these processes in rodents, opening up the potential of extending these paradigms to more detailed neurobiological investigation and interventions.

THE CONSTITUENT PROCESSES OF TRUST

OPERATIONALIZATION AND DYNAMICS OF TRUST IN THE LABORATORY

According to most definitions trust entails (1) the willingness to increase one’s vulnerability to another whose behavior is not under one’s control and (2) a certain expectation of positive reciprocity (Colman, 2003; Rousseau, Sitkin, Burt, & Camerer, 1998; Sztompka, 1999). As such, the decision to trust is fundamentally
A decision under uncertainty. Similarly, the reciprocation is necessarily delayed in time, eliciting a need for at least a temporary release from myopic mechanisms that discount the value of future outcomes due to (in evolutionary sense very real) uncertainty with regards to that future. When encountering a novel interaction partner, a decision to trust this partner could thus be framed as a gamble with an uncertain and perhaps delayed outcome.

Within psychology and behavioral economics, the processes related to trust are studied in a paradigm borrowed from game theory called the Trust Game (Berg, Dickhaut, & McCabe, 1995; see Fig. 1). In the trust game two players can share a certain amount of money. The first player has two options: (1) she can choose to divide the money equally between herself and the second player (both get €50), or (2) give it all to the second player with the advantage that the stake increases in value (sum is tripled so new total = €300). The second player then has the choice to reciprocate and share the increased amount of money with the first player (both get €150), or to defect and exploit the given trust by keeping the money for herself.

In contrast with these predictions, experimental data show that most people trust the second player, and also that the second player’s trust is generally reciprocated (Camerer, 2003). In iterative, multiple round trust games, the theoretical
predictions and behavioral strategies change (Axelrod, 2006; Axelrod & Hamilton, 1981). Studies with multiple round games have shown that participants often play what seems like a tit-for-tat strategy (King-Casas et al., 2005; Nowak & Sigmund, 1992; van den Bos, van Dijk, & Crone, 2012; Wedekind & Milinski, 1996).

In sum, trusting involves generating, acting upon, and updating expectations of reciprocity. These processes in their most basic form can be studied in context of the (iterated) trust game and can be found, on a basic level, in nonhuman animals as well. Often these games involve anonymous interactions with unknown others, but more recently they have been extended to study the effects of reputation (Chang, Doll, van’t Wout, Frank, & Sanfey, 2010; Delgado, Frank, & Phelps, 2005) and social relationships (Güroğlu, Van Den Bos, & Crone, 2009; Güroğlu, van den Bos, & Crone, 2014). Next we will review the cognitive processes involved in trust, starting with the generation of the initial expectation of reciprocity.

GENERATING EXPECTATIONS

Numerous studies have shown that the decisions to trust are correlated with expectations of reciprocal behavior (Evans & Krueger, 2009, 2011; Fetchenhauer & Dunning, 2009, 2010; van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2011; van den Bos, Westenberg, van Dijk, & Crone, 2010). This indicates that expectations do matter but also raises the question what these expectations are based upon. When encountering an interaction partner, there are numerous cues, and thus numerous processes, that can be relied on to generate expectations about the probability of future reciprocity of a known partner. First, the identity of the other needs to be established to retrieve information about past interactions. Establishing the identity of the interaction partner entails minimal cognitive skills to distinguish kin from non-kin and to recognize familiar agents. It is known that nonhuman primates are very good at recognizing kin as well as individual group members (e.g., Silk, Alberts, & Altmann, 2004; Tomasello & Call, 1997). Similarly research on rodents has shown that they are good at recognizing both kin and familiar others based on unique “olfactory signatures” (Brown, Singh, & Roser, 1987; Engelmann, Hädicke, & Noack, 2011; Thor & Holloway, 1982). Rats and mice will show a bias toward investigating a novel interaction partner over a partner they have recently interacted with, showing that they act as if they have a social memory for previous interactions (Caldwell, 2012; Lukas & Neumann, 2013). Simple familiarity is already associated with increased levels of trust (Dunn & Schweitzer, 2005), but this may be modulated by further knowledge of the individual or context.

In the context of strategic social interactions, humans often engage in making inferences about the knowledge and intentions of others in order to better predict their behavior, an ability often termed Theory of Mind (ToM). Additionally, people may use self-projection/perspective-taking (putting oneself in the other’s shoes) in order to predict their future behavior (Evans & Krueger, 2009, 2011; van den Bos

1. Note, however, that the studies by Dunning and colleagues also have shown that people’s expectations are significantly lower than the levels of reciprocity encountered in these games.
et al., 2010). However, as far as we know there is currently no evidence that such processes play a role in the social behavior of other animals.

Another very useful source of information used to generate predictions about other’s behavior is to rely on indirect or direct experience with the interaction partner. That is, we may have learned about the trustworthiness of an agent by having observed its reciprocal behavior. This type of indirect experience has already been demonstrated to play a role in 3-month-old infants’ decision how to interact with others. For instance, Hamlin, Wynn, and Bloom (2008, 2010) have demonstrated that 3- and 6-month-old infants have a preference for an agent who acted prosocially over one who acted antisocially toward others. More recently, Tomasello and others have shown similar findings in orangutans and chimpanzees (Herrmann, Keupp, Hare, Vaish, & Tomasello, 2012). Such image scoring behavior, however, has not yet been reported for rodents in the literature.

Of course, another very important mode of indirect experience is gossip (Dunbar, 2004), although this type of verbal transmission of social information is likely only seen in humans. Several studies have shown, using vignettes describing the moral status of the interaction partners, that gossip, or social information, has a large influence on the initial level of trust, but is quickly altered by direct experience of reciprocity (Chang et al., 2010; e.g., Delgado et al., 2005). For instance, in one study the participants made decisions to trust hypothetical trading partners after having read descriptions indicating praiseworthy, neutral, or suspect moral character (Delgado et al., 2005). They played multiple rounds with each of the players, which were pre-programmed to have the same level of reciprocity. Initially, the vignettes about the moral character led to large differences in expectations of trustworthiness, but after 24 rounds the expectations were the same for each of the partners. The learning mechanisms involved in updating of expectations based on direct experience will be discussed separately below (Repeated Interactions: Strategies, Learning, and Memory).

Finally, if there is no direct or indirect experience available, or any cues that would predict trustworthiness, one can derive a prediction based on the general level of reciprocal behavior of the population. This, population level reciprocity, or general reciprocity, is of course not known but inferred from the individual life history of social interactions. As pointed out above, rats seem to be sensitive to the general level of reciprocal behavior in the environment (Rutte & Taborsky, 2007a).

TRUST, RISK, AND UNCERTAINTY

No matter how much information we rely on to predict the future behavior of others, there is always a certain level of uncertainty about what they will do next. Traditionally, uncertainty is classified into two types (Knight, 1921): risk (known variance in an outcome distribution; O’Neill & Schultz, 2010) and ambiguity (an unknown outcome distribution; Burke & Tobler, 2011). Of course most social situations are not examples of pure risk, that is, we normally do not have access to the exact probabilities underlying the behavior of others. However, with repeated
sampling, that is, over many repeated interactions with the same partner, ambiguity might turn into risk (Schultz, O’Neill, Tobler, & Kobayashi, 2011). Thus, to understand trust decisions, we need to understand processes involved in both risky and ambiguous choice (Fairley, Sanfey, Vyrastekova, & Weitzel, 2012).

A wealth of research in the field of behavioral economics has indicated that both humans and animals, when deciding between outcomes that differ in terms of probability (risky choices), exhibit deviations from rational choice, as predicted by normative theory (Expected Utility Theory; von Neumann & Morgenstern, 1944) by showing risk-averse or risk-seeking behavior (Kahneman & Tversky, 1979; Kalenscher & van Wingerden, 2011). In the lab, rodents are mostly risk averse (Cardinal & Howes, 2005). Cumulative Prospect Theory (CPT) is currently the dominant framework that explains this pattern of risk attitudes. The key elements of this theory are (1) a value function that is concave for gains, convex for losses, and steeper for losses than gains, and (2) a nonlinear transformation of the probability scale, which overweights small probabilities and underweights moderate and high probabilities (Tversky & Kahneman, 1992).

If a decision to trust is indeed equivalent to risky decision making, it is well possible that risk attitudes and tendencies to trust are correlated within the population. Alternatively, the two decision-making processes might recruit different cognitive and/or neural systems. For example, it has been shown that probabilistic reasoning is more successful in a social compared to a non-social scenario (Cosmides & Tooby, 1992), so it might be the case that risk attitudes and trusting behavior rely on separable processes and could diverge. Several studies have tried to establish this link directly (comparing monetary gambles and trust game behavior) but with mixed results (see, e.g., Ashraf, Bohnet, & Piankov, 2006; Eckel & Wilson, 2004; Houser, Schunk, & Winter, 2010; Schechter, 2007). Furthermore, the neurochemical substrates of both processes seem to diverge as testosterone increases risk taking (Apicella et al., 2008) but decreases trust (Bos, Terburg, & van Honk, 2010), and oxytocin increases trust but reduces risk taking (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005).

There are several reasons why there is no strong relationship between monetary gambles and trust behavior. First of all, trust choices are often more like ambiguous choices (Fairley et al., 2012). Other than in the canonical risky gambles, people do not have exact point estimates of the probabilities that are given in the monetary gamble paradigms. Instead, one might rely more heavily on previous experience. Interestingly, it has recently been shown that people behave differently when making risky gambles based on experience versus description (Hertwig & Erev, 2009). What these experiments have consistently shown is that when people gain probabilistic information by experience they actually overweight large probabilities and underweight small probabilities (the opposite behavior as predicted by CPT). It may very well be that the choice behavior on such a task is more predictive of trust behavior. Interestingly, in studies with rats the standard way to induce knowledge about probabilities is by experience, given that it is impossible to convey risk probabilities directly.
Finally, there have been numerous studies that have shown that people’s uncertainty attitudes differ across domains and situations, even when using the same assessment method (for a review, see Blais & Weber, 2006). This research has indicated that people have very different risk attitudes for financial versus social risks (Weber, Blais, & Betz, 2002). For instance, people report a greater likelihood to engage in social versus financial risk. Other studies also show that people, who in financial situations are normally ambiguity averse (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005; Levy, Snell, Nelson, Rustichini, & Glimcher, 2010), may be ambiguity seeking when choices are in the domain of their expertise (Heath & Tversky, 1991). In sum, trust decision involves dealing with different levels of uncertainty, and may thus be associated with processes involved in risky and ambiguous choices. The domain specificity of uncertainty attitudes calls for the incorporation of a non-social control condition when assessing the influence of social context on choices under uncertainty.

TEMPORAL DIMENSIONS OF TRUST

Besides wondering if someone will reciprocate, it may also be important to consider when trust will be reciprocated (Stevens & Hauser, 2004). As the payoffs of a fruitful, enduring cooperation are delayed in the future with respect to immediate payoffs, the decision to trust could be subject to delay discounting as well. Delay discounting refers to the finding that humans and all other animals tested exhibit preferences for immediate over delayed outcomes reward (Ainslie, 1974; Kalenscher & van Wingerden, 2011; McDiarmid & Rilling, 1965; Rachlin & Green, 1972; Samuelson, 1937). Delay discounting has been formalized in the normative economic framework of discounted utility theory (Kalenscher & Pennartz, 2008; Koopmans, 1960; Prelec & Loewenstein, 1991; Samuelson, 1937). This theory states that the value of a delayed reward decreases with the duration of the delay according to a discount function. Therefore, in maximizing the utility of her choices, a decision maker should take into account the discounted utility, derived by reducing the utility of each of the alternatives according to this time-dependent discount factor. For example, most people prefer $10 today over $20 in 6 months. Most empirical research on choices over delayed outcomes is consistent with a hyperbolic decay of the value of outcomes as they are delayed into the future proposed (Ainslie, 1975; Green, Myerson, & Calvert, 2010; Mazur, 1984, 1988; Rachlin, Raineri, & Cross, 1991). The tendency of humans and other animals to put a premium on the short-term availability of rewards has been called impulsivity or failure of self-control. However, while humans are known to be able to delay outcomes of longer periods of time, many animal species can wait only a few seconds for delayed benefits, reflected in steeper hyperbolic discount functions (a point raised by Stevens & Hauser, 2004). Note however that these discount rates are not static parameters and are dependent on the state of the agent, the context, and the modality of rewards (van den Bos & McClure, 2013). Also, when humans are asked to decide between immediate or delayed food rewards, using similar designs as employed
with animals such as rats and pigeons, remarkably similar discount rates emerge (Calvert, Green, & Myerson, 2011; Jimura, Myerson, Hilgard, Braver, & Green, 2009).

Importantly, the extent to which an organism discounts future rewards thus puts significant constraints on trusting behavior. An essential feature of trust is that the expected reciprocal behavior will occur somewhere in the future. For example, even though we would be 100% sure someone would reciprocate our trust at some future date, we could decide against it given that the future reward is discounted so much that it is valued less than the immediate reward associated with not trusting. Some authors thus argue that the delays involved in reciprocity are too long to sustain trust in animals with high discount rates (Stevens & Hauser, 2004), while others argue that emotional heuristics could take the place of explicit delay discounting to sustain long-term cooperation (Berra, 2014). To develop a successful animal model of a full Trust Game analogue with delayed outcomes, these design constraints should be taken into account.

REPEATED INTERACTIONS: STRATEGIES, LEARNING, AND MEMORY

Finally, for trusting relationships to emerge we have to understand which processes are involved in repeated interactions. The questions regarding repeated play have been dominated by the search for interactive strategies that could lead to the evolution of stable cooperative populations (Axelrod & Hamilton, 1981). These endeavors resulted in various strategies, the most well known being tit-for-tat (TFT). This decision strategy always cooperates (trusts) on the first encounter and subsequently mimics the others player’s last choice. Over the past decades, similar strategies (e.g., generous TFT, tit-for-two-tats) have been developed to increase robustness of cooperation in different contexts. However, while these decision theories have not focused on the cognitive processes involved in the implementation of these strategies it is clear that all of them rely on a very basic capacity to memorize and recall the outcomes of previous encounters.

An alternative approach to understand behavior in repeated interactions has focused on the role of very basic reinforcement learning (RL) processes (Camerer & Hua Ho, 1999). RL outlines a formal framework of how agents learn to optimize their behavior (i.e., obtain rewards and avoid punishments) through interactions with their environment. A critical aspect of RL is the processing of prediction errors (i.e., the actual minus the expected outcome of an action). Each time an interaction takes place the prediction error will be used to update future expectations. The advantage of the RL process is that it updates expectations based on the outcome of the current interaction only. Thus, as most TFT-like strategies, RL does not rely heavily on memory capacities. However, both learning strategies make different behavioral predictions. For instance, if a partner starts to defect after a long period of cooperation, TFT will immediately stop trusting but RL will still trust for several rounds (depending on the learning rate) given that the expected value of trust is still high.
Interestingly, recent neuroimaging studies have suggested that, in humans, RL mechanisms are involved in iterated games, including the trust game (Delgado et al., 2005; King-Casas et al., 2005; van den Bos, Talwar, & McClure, 2013). That is, these studies have consistently shown the involvement of the ventral striatum and the ventral medial prefrontal cortex (vmPFC) in iterated games. These structures are associated with the generation of prediction errors and the representation of expected value respectively (for a review, see Rangel, Camerer, & Montague, 2008). Animals, including monkeys and rats, have been shown to learn in accordance with models of reinforcement learning and to rely on these same neural structures (for a review, see Schultz, 2006). It is therefore conceivable that rats will also rely on these mechanisms when interacting with conspecifics and might show different learning rates for social versus nonsocial contexts.

EXPERIMENT: RODENT ANALOGUE OF THE TRUST GAME

In sum, is it feasible to study the complex subject of trust in nonhuman animals? Ideally, an animal model would be established in rodents as these animals have been extensively studied in operant choice behavior and other economic choice settings (Kalenscher & van Wingerden, 2011). Such a model would be interesting in itself from a comparative point of view, but importantly also open up the study of trust behavior for invasive methods like single-cell and network electrophysiology, lesion and pharmacological interventions, and causal manipulations such as optogenetic techniques. For example, while fMRI studies show similar activation of reward-related networks when social or non-social outcomes are considered, it is not clear whether this recruits the same neural pathways. Indeed, recent results in nonhuman primates suggest that social and non-social reward processing might activate different neuronal pools, albeit located in the same neural regions (Klein & Platt, 2013).

Clearly, rats do not possess such a rich concept of trust as humans do but we would argue that they have the core abilities related to trust in place: the abilities for social recognition and long-term social memory, for generating predictions, for making choices based on these predictions, and for learning from feedback to update these predictions. We thus attempted to create a rodent analogue of the trust game using a novel operant behavioral task. The general outline is as follows: rats were first trained to discriminate between two odor cues that predicted a large reward with high or low probability. In a second stage, a partner or puppet rat was added to the experimental cage. Now, two novel odors predicted the large and risky reward (LR), or a small and certain reward (SC). Critically, the probability of the LR was signaled by the presence of one of two partner rats (social conditions), or partner puppets (non-social control). As such, a trustworthy partner (TP, rat or puppet) predicted an 80% chance of large reward, whereas a non-trustworthy partner (NTP, rat or puppet) predicted a 20% chance. With these tests, we ascertained whether rats were able to fulfill the basic requirements for trusting behavior: the discrimination of different outcome probabilities and the adjustment of the
propensity for risky choice (analogue to trust decision) dependent on the context cues predicting those probabilities (analogue to [learned] trustworthiness of the partner).

The face validity of this approach was ensured by copying the basic experimental design of the “first move” in a trust game: an actor chooses to trust (select a larger, risky outcome with stable/known variance) or not to trust (select a smaller, certain outcome). The reciprocity rate (Second Move: the probability of the second player to share: here either resulting in a large payback or 0) was set as a fixed sequence of pseudo-randomized responses, in order to study the First Move in absence of within-session variance in the Second Move. In the design chosen here, no actual move (i.e., observable operant response) of the partner (rat or puppet) is required. Thus, “reciprocate” and “defect” decisions are only revealed at the moment of reward payout (either large reward or omission). However, as the majority of studies employing the trust game are performed in computerized forms, where the participants are only shown the “selected” outcome for the current partner, often just depicted by a snapshot or avatar, and the actual choices of the partner are sometimes determined by software, the setup of the rodent design is not that different. As requiring movement from the partners would introduce observable differences in the agency of partner (voluntary vs. motorized movement), we chose to restrict this exploratory study to the First Move only. Further suggestions on studying a second move and reciprocity, perhaps based on social motivations, are discussed below.

Based on averaged sucrose solution access, a rational decision maker should prefer the odor cue linked to the risky option in the presence of a TP, and the odor cue linked to the safe option in the presence of an NTP. However rats (Cocker, Dinelle, Kornelson, Sossi, & Winstanley, 2012; Cocker & Winstanley, 2015), like humans (Kahneman & Tversky, 1979; Keren & Wagenaar, 1987), but unlike certain primates (O’Neill & Schultz, 2010), are mostly risk averse. As such, we expected to find a modulation of risky choice by partner identity, but not necessarily risk-seeking (i.e., LR choice > 50%) behavior.

Construct validity of the animal model would be strengthened if a difference in choice behavior could be shown between the social and nonsocial conditions, allowing to disentangle the specific social reinforcement learning (i.e., Trustworthiness feedback) component from general (non-social) reinforcement learning strategies. Such a divergence was tested in probe sessions, where the presence of both partner cues (TP and NTP) in both contexts (social and non-social) now predicted a 50% chance of LR delivery. If actor rats would disregard the partner cues, but only track actual reward delivery probabilities, no difference in risky choice would be expected for the probe sessions. If, however, rats continued to use the social and/or non-social cues as a predictor for the probability of LR reward delivery, a carryover effect of the previously learned probabilities on risky choice could be expected. Similar carryover effects have been demonstrated in humans (Chang et al., 2010; Fareri, Chang, & Delgado, 2012). We hypothesized that rats, like humans, would show this carryover effect only in the social context condition. This leads to the following predictions:
let the trust contrast \( \text{TC}_{(x,y)} = \text{TP}_{(x,y)} - \text{NTP}_{(x,y)} \)  
(1)

for each of the combinations of \( x \) (social/non-social) and \( y \) (learn/probe) and

\[
\Delta \text{TP}_{(x)} = \text{TP}_{(\text{probe}, x)} - \text{TP}_{(\text{learn}, x)} ;
\]

\[
\Delta \text{NTP}_{(x)} = \text{NTP}_{(\text{probe}, x)} - \text{NTP}_{(\text{learn}, x)}
\]

(2a)

(2b)

(1) rats, regardless of whether they were confronted with partners or puppets, will choose the LR cue more when partnered with their TP than when partnered with their NTP during the learning phase:

\[
\text{TC}_{(\text{learn}, \text{social})} > 0 \text{ and } \text{TC}_{(\text{learn}, \text{non-social})} > 0
\]

(2) no differences between learning in the social or non-social condition:

\[
\text{TC}_{(\text{learn}, \text{social})} - \text{TC}_{(\text{learn}, \text{non-social})} \approx 0
\]

(3) during the probe phase, rats will retain discrimination between former TP and former NTP in the social condition but not in the non-social condition:

\[
\text{TC}_{(\text{probe}, \text{social})} > 0 \text{ and } \text{TC}_{(\text{probe}, \text{non-social})} \approx 0
\]

(4) carryover is specific to the social condition:

\[
\text{TC}_{(\text{probe}, \text{social})} > \text{TC}_{(\text{probe}, \text{non-social})}
\]

(5) carryover in the social condition is present for both partners:

\[
\Delta \text{TP}_{(\text{social})} \approx 0 \text{ and } \Delta \text{NTP}_{(\text{social})} \approx 0
\]

(6) carryover in the non-social condition could be affected for both partners:

\[
\Delta \text{TP}_{(\text{non-social})} < 0 \text{ and } \text{or } \Delta \text{NTP}_{(\text{non-social})} > 0
\]

(7) leading to inequality between social and non-social conditions, that is, the difference of the differences unequal to zero.

\[
(\Delta \text{TP}_{(\text{social})} - \Delta \text{NTP}_{(\text{social})}) > (\Delta \text{TP}_{(\text{non-social})} - \Delta \text{NTP}_{(\text{non-social})})
\]

**MATERIALS AND METHODS**

**SUBJECTS**

Twenty adult male Long-Evans actor rats (28 postnatal days upon arrival, Centre d’Elevage Janvier, France) were used in this experiment, of which 4 were picked to become interaction partners and the remaining 16 were used as actors in the behavioral testing. All animals were socially housed (\( N = 4 \) per cage) in Makrolon©
Type IV cages with sawdust bedding under environmentally controlled conditions (7:00 a.m./7:00 p.m. reversed light/dark cycle, temperature 20 ± 2°C, relative humidity 55 ± 5%) with access to water *ad libitum*. The partner rats were housed together in a single cage. Both groups of rats were prevented from interaction prior to the experiment. Focus rats weighed 184–205 g at the start of training and were maintained at 90% of their free feeding weight. All experimental procedures comply with the German Animal Welfare Act and were approved by the local institutional animal care committee.

APPARATUS

Rats were tested in four operant chambers (ENV-009; Med-Associates Inc., St. Albans, VT, U.S.; Fig. 2A) enclosed in ventilated, sound- and light-attenuating experimental cubicles. Two loudspeakers attached above the cubicle entrance delivered white noise. Chambers (30.5 × 39.4 × 29.2 cm) were equipped with a stainless steel grid floor, a transparent polycarbonate door and rear panel (front and back), and two stainless steel walls (left and right) that hold modular components within vertical aluminum channels. A 100 mA house light was located at the top center of the left wall. A transparent perforated polycarbonate partition divided each chamber into two equally sized compartments, while allowing visual, olfactory, and auditory exchange between the subdivisions. In each compartment, a sucrose solution delivering sipper tube (ENV-252M) mounted outside could be made available in a reward port, equipped with a photobeam for access detection. A reward cue light (ENV-22M) was placed 11 cm above the reward delivery port. Every compartment held two nose-poke devices embedded cue lights and olfactory presentation capabilities (ENV-257-NPP) on the right wall. Nosepoke devices and reward delivery port were situated 2.5 cm above ground level. Regulation of odor delivery followed a system of solenoid valves and flow meters (ENV 275-5; van Duuren et al., 2007).

AUTOSHAPING AND TRIAL STRUCTURE

After an initial habituation period, during which the animals were allowed to explore the apparatus for 10 minutes in two daily sessions, autoshaping of the operant procedures began. Rats were solitarily trained in the actor compartment (labeled “A” in Fig. 2A) to make a nosepoke, sample the odor, make a reward poke, and consume the reward. A 10% sucrose reward solution was used throughout the experiment. A trial started with illumination of the nosepoke cue lights. Rats were shaped to make a successful nosepoke (lasting for at least 1,250 ms, divided in a 500 ms pre-stimulus period and a 750 ms odor stimulation phase) within 10 seconds after cue light illumination. Upon a successful nosepoke, the reward light was illuminated. If the rat made a successful reward poke (lasting at least 1000 ms), the reward bottle was lowered and available for reward delivery for a specified time. Note that the spatial location and odor cue (for example, jasmine
or nutmeg odor) for both options was different and counterbalanced across sessions, while the reward delivery site was the same for both outcomes. Nosepoke events were captured with infrared photo beams. A computer equipped with the software program MED-PC IV controlled the chambers and recorded the data.

**TASK DESIGN: PROBABILITY DISCRIMINATION**

After autoshaping, all actor rats performed a probability discrimination task. In this task, we tested in $N = 4$ sessions whether our rats were able to distinguish between two odors linked to a low (LP, 20%) and high (HP, 80%) probability of reward delivery (3 s sucrose solution access, the alternative outcome being reward omission). Each session consisted of 40 pseudo-randomized forced choice trials followed by 40 free choice trials. During free choice trials, both nosepoke cue lights were turned on to signal presence of choice options. Choosing to nose-poke in one...
port deactivated the other nosepoke port. Successful probability discrimination would be ascertained by testing the percentage of choices for HP against indifference (50%).

**TASK DESIGN: PROBABILITY MODULATION BY CONTEXT CUE, LEARNING PHASE**

After the probability discrimination task, the actor rats were split in two groups that would be partnered with rats (social context cues) or rat puppets (non-social context cues). Now, the context cue was placed in the second compartment (labeled “B” in Fig. 2A), while the operant devices in the partner compartment were disabled (Fig. 2B). Each actor rat was paired 9 times in \( N = 18 \) daily sessions with their TP (predicting 80% probability of reward on the LR option) and NTP (predicting only 20% probability), alternating day by day (Fig. 2C) in the learning phase. We chose an interleaved design to arrive at an equal distribution of partner trustworthiness conditions along the general task learning curve, and, in turn, to maximally challenge the animals to learn and update the value of the probabilistic reward, as it changes value between each session. Each session consisted of 40 forced choice trials to sample the current probability distributed (high or low probability of reward) associated with the odor linked to the risky outcome, followed by 60 free choice trials. The odors linked to the LR and SC rewards, respectively, were always the same and counterbalanced across rats. Only the actual probability of LR delivery differed based on the context cue. The nosepoke hole (left, right) linked to each odor was counterbalanced within rats across sessions.

**TASK DESIGN: PROBABILITY MODULATION BY CONTEXT CUE, PROBE PHASE**

In the final probe phase, the alternating pairing with TP and NTP partners (puppets or rats) was continued for 2 sessions with each partner. However, the forced choice trials were omitted and rats just completed \( N = 60 \) free choice trials. The same odors linked to the LR and SC rewards were used as in the learning phase, but now the context cue was made irrelevant and actual large reward delivery probability was set in all four sessions at 50%.

**DATA INCLUSION AND ANALYSIS**

Choices were extracted as %choice of HP (probability discrimination) or LR (learning and probe phases) cue per session, counting the free choice trials only. Only sessions in which rats completed > 50% of free choice trials were included. Only rats that completed >15 out of 18 learning sessions according to this criterion were included in the final analysis. This excluded three rats from the social context group (now \( N = 5 \)), and one rat from the non-social context group (now \( N = 7 \)).
The minimum of valid sessions for a condition was now reduced to 7. To balance the number of sessions that were entered into the final analysis, we selected the last 7 completed sessions per condition per rat. Choice data was tested in two ways: First, using parametric ANOVAs with trustworthiness (TP vs. NTP) as within-subjects factor and partner context (social/partner vs. non-social/puppet) as between-subjects factor. Second, to overcome the difference in statistical power for the within-subject and between-subject comparisons, by a bootstrap resampling procedure (van Wingerden et al., 2012). In this procedure, \( N = 5000 \) paired or unpaired draws are made, with resampling, from the population of raw scores under investigation (i.e., \( N = 5 \) \( TP_{probe/social} - NTP_{probe/social} \) vs. \( N = 7 \) \( TP_{probe/non-social} - NTP_{probe/non-social} \) for an unpaired comparison). The subsequent distributions of these draws are normally distributed and can be compared directly with a Z-test. Significance of the difference is thus assessed by taking, for unpaired comparisons:

\[
Z_{A-B} = \frac{\mu_A - \mu_B}{(\sigma_A + \sigma_B)/2} \tag{3a}
\]

and, for paired comparisons:

\[
Z_{A-B} = \frac{\mu_{(A-B)}}{\sigma_{(A-B)}} \tag{3b}
\]

that is, the difference in means normalized to the average standard deviation (Eq. 3a) or the mean of the difference divided by the standard deviation of the difference (Eq. 3b). The resulting Z-score is then compared to Z-levels corresponding to certain \( p \)-values, such as \( |Z| > 1.96 \) for \( p < 0.05 \).

In addition, contrast scores \([probe(A) - learning(A) for all types or sessions: TP-social, TP_non-social, NTP_social and NTP_non-social]\) were calculated and tested against 0 using a one-sample \( t \)-test or, using the bootstrapped population, using

\[
Z_A = \frac{\mu_A}{\sigma_A} \tag{4}
\]

as an expression of the Z-scored distance to 0.

RESULTS

Initially, rats were trained in four shaping sessions to discriminate two cues linked to a low (LP, 20%) and high (HP, 80%) probability of reward delivery. Unsurprisingly, rats preferred the high-probability outcome at levels above chance, indicating they can discriminate between these two probability levels (Fig. 3). Averaging
%HP choice over the four sessions yielded a score of 61 ± 2.9% (mean ± SEM), significantly different from chance (50%) as established through a one-sample t-test: \( t(11) = 3.83; p = 0.003 \). When the data were split amongst groups that would later be trained in either social or non-social contexts, no differences emerged: \( t(10) = 1.29; p = 0.22 \).

Next, rats were split into two groups, to be trained in a social (rat partners) or non-social (puppet partners) context condition. Rats were put through daily learning sessions, partnered in alternating fashion with either their TP or NTP. Importantly, each partner predicted different levels of probability for the delivery of the LR outcome (TP: 80%, NTP: 20%). Both in the social condition (Fig. 4A) and the non-social condition (Fig. 4B), actor rats over days started to choose the risky outcome more when partnered with the TP as compared to the NTP. A mixed ANOVA with session (1–7) and trustworthiness (TP/NTP) as within-subject factors and context (social/non-social) as between-subject factor yielded a significant main effect for the factor trustworthiness only: \( F(1, 10) = 17.01; p = 0.002; \text{partial } \eta^2: 0.63, \text{power: 0.96} \). There was no significant main effect of session, \( F(6, 60) = 1.50; p = 0.20; \text{partial } \eta^2: 0.13, \text{power: 0.54} \), nor an interaction, \( F(6, 60) = 1.35; p = 0.25; \text{partial } \eta^2: 0.12, \text{power: 0.49} \). The possibility that subjects would be faster to distinguish between partners in the non-social versus social context would be supported by a significant triple interaction. This was not the case: session*trustworthiness*context, \( F(6, 60) = 0.90, p = 0.50; \text{partial } \eta^2: 0.09; \text{power: 0.33} \).

Indeed, when averaged across sessions, a mixed ANOVA with trustworthiness (TP, NTP) as a within-subject factor and context condition (rat, puppet) as a between-subject factor yielded a significant main effect for trustworthiness, \( F(1, 10) = 24.6, p < 0.001, \text{partial } \eta^2: 0.71, \text{power: 0.99} \), but neither a main effect of context, \( F(1, 10) = 0.84, p = 0.40, \text{partial } \eta^2: 0.08, \text{power: 0.13} \), nor a significant interaction (trustworthiness*context), \( F(1, 10) = 0.90, p = 0.37, \text{partial } \eta^2: 0.08, \text{power: 0.14} \). These results suggest that rats were able to learn the contingencies linking partner identity with the current probability associated with the LR choice option, irrespective if the partner was another rat or a rat puppet (Fig. 5A). The bootstrap procedure yielded a similar picture. The trust contrast (TC; see hypotheses section) was significantly above zero for both social, \( TC_{(social, learn)} = 7.94, p < 0.001 \), and non-social, \( TC_{(non-social, learn)} = 3.93, p < 0.001 \). Though the contrast is larger for the social condition, the difference between TCs was not significant (\( Z = -1.88, p = 0.06 \)).

However, it might be that this relative preference for the LR outcome reflects within session learning of the current contingencies. Such learning could take place during the 40 forced choice trials that were imposed at the start of each session, and could be independent of the (non-)social cue present in the adjoining compartment of the operant cage. If, however, actor rats did use the identity of the partner as a predictor to guide their choice either toward or away from the LR option, such a carryover effect could persist even when the actual LR probabilities were equated.

This is exactly what we tested in 2 × 2 subsequent probe sessions, directly following the learning sessions. All actor rats were partnered twice with both their partners again, essentially continuing the day-by-day alternation scheme. In these
probe sessions, the forced choice trials were omitted and just the choice distribution over 60 free choices was assessed. The LR choice option in both the presence of the TP and NTP now yielded the large reward (long sucrose access) on 50% of the trials in both conditions.

A mixed ANOVA for the probe phase with trustworthiness (TP/NTP) as within-subject factor and context (social/non-social) and between-subject factor resulted in a trend for trustworthiness to become significant, $F(1, 10) = 3.6, p = 0.09$, partial $\eta^2: 0.26$, power: 0.40, but no main effect of context, $F(1, 10) = 0.08, p = 0.78$, partial $\eta^2: 0.008$, power: 0.06, or significant interaction (trustworthiness*context), $F(1, 10) = 0.06, p = 0.82$, partial $\eta^2: 0.006$, power: 0.06, could be found (Fig. 5B). However, here, the bootstrap analysis revealed that the TC for the social condition remained significant ($TC_{(social,probe)} = 2.79, p = 0.005$) while the TC for the non-social condition was not ($TC_{(non-social,learn)} = 1.08, p = 0.28$). The difference between TCs was again not significant ($Z = 0.51, p = 0.61$).

The bootstrap analyses indicated that the TCs for all except the non-social–probe group were significantly greater than 0, even at Z-levels adjusted with Bonferroni-correction for multiple testing ($0.05/4 = 0.0125, |Z| > 2.56$; Fig. 6A). Indeed, one-sample t-tests on the TC scores showed a similar pattern, though with more modest levels of significance (Fig. 6B).

These results suggest that actor rats could have used the identity of their social partners in estimating the probability of LR reward, and that this effect persisted even when the actual probability of large reward delivery associated with each partner was equal, but only for the social context. A difference in TC values be-
between learning and probe sessions could arise by a change in %LR choice toward TPs, NTPs, or both. To identify the nature of the difference in TC values between the social and non-social contexts, we calculated difference scores separately for TP and NTP sessions ($\Delta TP/\Delta NTP$, see hypotheses above). We found that $\Delta TP_{(\text{social})}$, $\Delta NTP_{(\text{social})}$, and $\Delta TP_{(\text{non-social})}$ were all negative, though not significantly different from 0, while $\Delta NTP_{(\text{non-social})}$ was the largest and positive difference score. This is illustrated by the individual difference plots, showing the %change in %LR choice from learning to probe phase, for each group (Fig. 7).

**DISCUSSION**

All in all, some of the hypotheses for the emergence of trust-like behavior in a rodent analogue of the Trust Game as postulated above could be confirmed in this pilot study. We found that rats were able to distinguish between TPs and NTPs in both the social and non-social context conditions (1a & 1b), with no difference between the conditions (2). In the probe phase, this pattern was maintained for the social context group (3a), but not for the non-social context group (3b). However, though both groups show a difference in significance, the difference between the groups was not significant (~4). We found that the carryover effect in the social condition was present both to the TP and NTP partners (5); in the non-social condition, however, we found that carryover differed in sign between the TP and NTP.

**FIGURE 4. Evolution of LR choice over sessions.**

(A) Actor rats that were partnered with rats started to choose the large risky (LR) outcome more often when partnered with the trustworthy (TP, solid lines) as compared to the non-trustworthy (NTP, dashed line) partner. (B) As in (A), but now for actor rats partnered with puppet partners. There was a main effect for the factor trustworthiness, $F(1, 10) = 17.01; p = 0.002$; partial $\eta^2$: 0.63, power: 0.96. There was no significant main effect of session, $F(6, 60) = 1.50; p = 0.20$; partial $\eta^2$: 0.13, power: 0.54, nor an interaction, $F(6, 60) = 1.35; p = 0.25$; partial $\eta^2$: 0.12, power: 0.49. Error bars indicate SEM.
partners, though not significantly (~6). This suggests that rats in the non-social context condition converged in their choice allocation more strongly to the actual probability of large reward delivery. A direct comparison of the difference between the social and non-social contexts, however, was not significant (~7), though a trend could be observed. Individual analysis showed that the largest change between learning and probe phases occurred in the %LR choice in the presence of the NTP in the non-social condition.

**COMPARISON TO THE TRUST GAME FOR HUMANS**

As discussed in the introduction to the experiment, the lack of agency on behalf of the partner is a salient divergence from human studies on trust, even though this agency of the human partner is sometimes only imagined (and, even, intentionally suggested while actually absent in computerized trust game partners). It is unlikely that the actor rats perceive the partner as the agent responsible for
delivering reward in the current design, perhaps also in part because they are not reinforced in this design. Rather, partners should be seen as a contextual modulating cue with its own reinforcement rule ("In the presence of this partner, choose cue A"), that is, represent a state space that promotes model-based learning (Abe, Seo, & Lee, 2011; McDannald et al., 2012) of the best course of action for the "First Move." It is currently unclear whether rats show aversion to inequity, as saliently present in the "defect" choice option, to the same degree as humans (Crockett, Clark, Tabibnia, Lieberman, & Robbins, 2008; Güroğlu, van den Bos, Rombouts, & Crone, 2010) or monkeys (Brosnan & De Waal, 2003), but if they do, the additional negative feedback from receiving zero while the other receives everything would be expected to change the choice structure. However, it is conceivable that inequity aversion shows a similar social/non-social salience difference. In that case, adding the Second Move could strengthen the observed differences in Trust Contrast between rats in social or non-social contexts.

A second notable difference is the inclusion of forced choice trials in the learning phase. Forced choice trials are typically used in animal models to instruct the

---

**FIGURE 6.** Trust contrast for all groups.

(A) Bootstrapped TC scores (%LR choice with TP – %LR choice with NTP) plotted as Z-levels per group (mean/ std). White bars: learning phase; grey bars: probe phase. Solid horizontal line: p = 0.025, |Z| = 1.96; dashed line: Bonferroni-correction to p = 0.025/4; |Z| = 2.50. (B) Raw TC box plot displays 25%-75% percentile of TC scores, medians (solid horizontal lines), minima and maxima (whiskers). TC values were tested against zero for significance (one-sample t-test vs. 0; *p < 0.05, **p < 0.01). Actors partnered with rats retained Trust Contrast values above 0 from learning to probe phase, while the TC score for actors partnered with puppets during the probe phase was not significantly different from 0.
animals about the payoff matrix and the probabilities associated with each action, while human participants can be cued in advance on the payoff matrix—and separately can be induced to a belief (Delgado et al., 2005) about the probabilities of reciprocation.

Third, we have used a payoff structure that is symmetric with regards to the SC and LR options: if a rat would consistently choose the SC option, it would get as much reward in the end (all sessions 100% * 1.5 s sucrose) as consistently choosing the LR option (50% of session 80% large, 50% of sessions 20% large; average 50% large; large = 2x small). This is different from the usual Trust Game choice structure (Fig. 1), where the larger reward > 2x the small reward. In the current regime of reward sizes, this leads to risk aversion (Kahneman & Tversky, 1979) in our rats (Fig. 4). Interestingly, in monkeys, small rewards as typically used in experiment with many trials, lead to risk-seeking choices (O’Neill & Schultz, 2010), though the same monkeys become risk-averse if the size of rewards increases. It would be possible to determine for each rat the subjective “Certainty Equivalent” (Rudorf, Preuschoff, & Weber, 2012), that is, the indifference point at which the large reward size is equivalent at 50% probability to the small reward (Cardinal & Howes, 2005), to make the differences in trustworthiness more salient, however, we are not aware of any studies in humans that have employed such a calibration procedure.

Apart from these differences, we feel that the current design mimicked choice differences as seen in humans for the contrast between programmed trustworthy and untrustworthy partners adequately (Montague, Lohrenz, & Dayan, 2015; van den Bos et al., 2012). We will now turn to a discussion of the implementation of
the Second Move in future designs to approach a full Trust Game as played with humans.

FROM FIRST MOVE TO TRUST GAME

Follow-up studies could employ an extension of the design in which the partner has to perform some action to be perceived as having agency in delivering the outcome, perhaps while still keeping this a “forced” choice on the part of the partner so as not to perturb the level of reciprocated trust associated with the partner. This would be akin to playing a trust game with a convincing, yet computerized, partner. However, as our results illustrate, the social/non-social distinction is an important control condition, and equating the level of agency with respect to the partner between the social and non-social condition in making the “Second Move” might be difficult. A spatial version of this design, for example, might instantiate a non-social slot machine apparatus with different mechanical “responses” in different spatial choice compartments. Indeed, we have recently launched a rodent version of the prosocial choice test using such a spatial choice paradigm (Hernandez-lallement et al., 2015). A spatial rodent trust game task would have the additional benefit that outcomes for “reciprocate” and “defect” would be spatially distinct, and thus perhaps more salient. However, introducing an additional spatial component complicates the analysis of neuronal responses, perhaps detracting from the efficiency of such a version of a rodent trust game.

CONCLUSION

In sum, we believe that it is well possible to approach the dynamics of some basic forms of the trust game in this rodent version of the trust game. Such an animal model opens up important avenues of comparative research (Kalenscher & van Wingerden, 2011), allowing a range of neurobiological interventions, as well as electrophysiological measurement and electrical or optical causal interventions in the neural systems thought to be involved in human trust. As such, it will provide an important tool to better understand the precise neural implementation of trust decisions and trust-like behavior.
REFERENCES


belief. Cognitive Psychology, 61(2), 87-105. doi: 10.1016/j.cogpsych.2010.03.001


tory rats and mice using the nonconditioned social discrimination procedure. Nature Protocols, 6(8), 1152-1162. doi: 10.1038/nprot.2011.353


Güroğlu, B., van den Bos, W., & Crone, E. A. (2014). Sharing and giving across adolescence: An experimental study ex-
Can You Trust a Rat?


and ambiguity. *Journal of Neurophysiology, 103*(December 2009), 1036-1047. doi: 10.1152/jn.00853.2009


Prelec, D., & Loewenstein, G. (1991). Decision making over time and under uncertain-

*management Science, 37*(7), 770-786.


