Building and Dismantling Trust: From Group Learning to Character Judgments

Philip Pärnamets (philip.parnamets@ki.se)
Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.
Department of Psychology, New York University, New York, NY, USA.
Tobias Granwald (tobias@granwald.st)
Andreas Olsson (andreas.olsson@ki.se)
Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.

Abstract

Trust is central to social behavior. In interactions between strangers some information about group affiliation is almost always available. Despite this, how group information is utilized to promote trust in interactions between strangers is poorly understood. Here we addressed this through a two-stage experiment where participants interacted with randomly selected members of two arbitrary groups and learnt their relative trustworthiness. Next, they interacted with four novel individuals from these two groups. Two members, one from each group, acted congruently with their group's previous behavior while the other two acted incongruently. While participants readily learnt the group-level information in the first phase, this was swiftly discounted in favor of information about each individual partner's actual behavior. We fit a reinforcement learning model which included a bias term capturing propensity to trust to the data from the first phase. The bias term from the RL model predicted participants' initial behavior better than their expectations based on group membership. Pro-social tendencies and individuating information can overcome knowledge about group belonging.

Keywords: Trust; reinforcement learning; decision making; morality

Trust

Trust is fundamental to social behavior. If you do not trust a person, you are less likely to want to interact with that person. As such, trust is central not only to maintaining healthy interpersonal relationships but is also foundational for socioeconomic prosperity and the long-term survival of our species. This paper investigates trust in the context of knowledge about an individual's social group.

Due to its importance trust has been investigated in multiple ways. Here we take trust to be reflected in actions of assuming risk by allowing a desirable outcome to be dependent on the (uncontrollable, unknowable) actions of a partner (Mayer, Davis & Schoorman, 1995; Berg, Dickhaut & McCabe, 1995). Trust can be divided into a state and trait component in the trustor (Mayer, Davis & Schoorman, 1995). The trait component reflects the trustor's propensity to trust and may be affected by general beliefs about individuals as well as by what prosocial tendencies are present in the culture or wider context in which trustor's find themselves in. (Peysakhovich & Rand, 2015).

The state component reflects more immediate knowledge about the trustee. For example, when deciding to trust a partner for the first time, multiple sources of information can interact, such as implicit opinions about people belonging to the same ethnic group (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011), subjectively rated trustworthiness based purely on appearances (van 't Wout & Sanfey, 2008) and stories about the partner's moral character (Delgado, Frank, & Phelps, 2005), or other indirect information (Zarolia, Weisbuch & McRae, 2016).

Learning to Trust

Trust decisions in real-life are rarely one-shot interactions. Indeed, the role of trust in grounding lasting relationships is one of the key reasons for the importance of understanding its cognitive mechanisms. Recent work has begun to address how trust evolves during repeated interactions. Outcomes from past interactions lead to expectations concerning how the partner will continue to act. This in turn informs one's decisions of whether to continue trusting an individual or not (King-Casas et al., 2005). Understanding the dynamics of trust, how trust changes in response to incoming information, is therefore central to understanding its underlying psychology.

Studies on repeated trust games have proposed a computational account of how state trust is learnt (e.g. Chang, Doll, van 't Wout, Frank, & Sanfey, 2010; Fareri, Chang, & Delgado, 2015), based on reinforcement learning (RL) models. RL provides a powerful framework for understanding how feedback from the environment is gradually integrated, through prediction error learning, to inform future actions. The assumption is that a person is attempting to maximize their expected utility from a given interaction. Importantly, past research has demonstrated how individuals not only learn the degree to which partners reciprocate, but that this learning is best described using models which include influence of general impressions of trustworthiness (Chang et al., 2010), or a value placed on social interaction (Fareri et al., 2015).

From Group to Character

A limitation of studies like those reviewed above is that they study the dynamic evolution of initial trustworthiness impressions and subsequent trust decisions in interaction with individual partners. However, humans are specialized for group living, and group belonging is a central feature of individuals' identity (Cikara & Van Bavel, 2014). How knowledge about groups influences and affects behavior in repeated trust situations is a surprisingly understudied. Here we attempt to fill this gap by examining how participants first

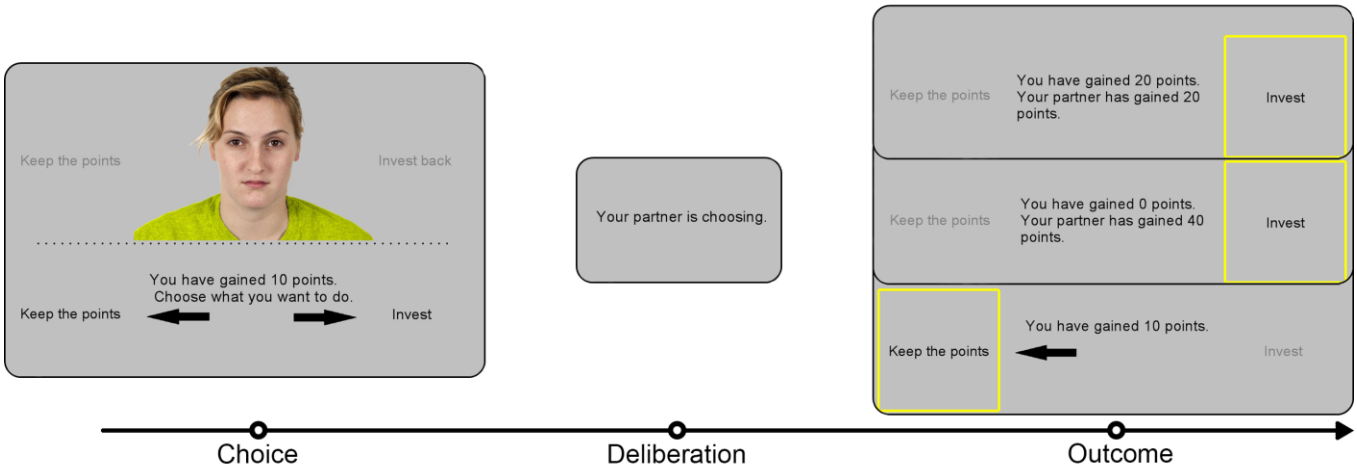

Figure 1. Overview of trial structure. Each trial participants were presented with a partner and given a choice to invest their endowment or keep it. Invested points were quadrupled. If participants invested the partners either reciprocated or kept the transferred points. In the group phase of the experiment, each trial featured a new partner from either the blue or the yellow group. In the character phase of the experiment, each trial featured one of four partners.

learnt about the relative trustworthiness of members of two arbitrary social groups, and then how they used that knowledge when engaging in repeated interactions with novel individuals belonging to either group who behaved as expected or not.

Our aim was to investigate how group information transferred to individual cases, as an experimental model for how stereotypes, "beliefs about the characteristics, attributes, and behaviors of members of certain groups" (Hilton & Von Hippel, 1996), affect trust. If state trust information dominates participants should only gradually learn to trust individuals from groups previously perceived as untrustworthy. However, recent work on stereotypes has suggested that they are rapidly discounted if participants are given individuating information (Rubinstein, Jussim, & Stevens, 2018). If so, participants should only be affected in their initial behavior towards people drawn from groups stereotyped as untrustworthy. We explored this using a mix of behavioral analysis and RL modeling.

Method

Participants

The sample consisted of 74 participants (36 female, 38 male) with an average age of 27.1 (*SD* = 4.7). All participants were paid for their participation by receiving two movie tickets (approximate value, 260 SEK). All participants read and signed an informed consent form. The experiment was approved by the Regional Ethical Review Board in Stockholm (2017/1116-31/4).

Materials

All parts of the experiment took place in a near-soundproof room. The participants were seated 60-70 cm (*M* = 64.5, *SD* = 2.14) in front of a 24 " LED screen with a refresh rate of 144 Hz and resolution of 1920 x 1080. The experiment was presented using Psychopy 1.83.0 (Peirce, 2007) in Python 2.7.10. Additionally, participants eye-movements were monitored, but those data are not reported here.

The pictures of the partners were selected from The Chicago Face Database (CFD; Ma, Correll, & Wittenbrink, 2015) and consisted of 64 Caucasian faces (32 female) with neutral facial expressions. All pictures were then modified in Adobe Photoshop to change the color of their shirt, so that each image has a blue and yellow shirted version.

Procedure

The experiment was divided into two parts, a group phase and a character phase. In both parts of the experiment, the participants took the roll of the trustor in a modified version of a repeated trust game (cf. Delgado et al., 2005). In both parts of the experiment the participants interacted with partners drawn from two different groups (blue and yellow-shirted) and were initially instructed that groups would behave differently, but not how or which.

Groups Participants were told that while the partner was computer controlled, the behaviors of the partners were based on results from a previous study and that the partner's behaviors was unrelated to the faces of people depicted in the experiment (to minimize the effect of visual trustworthiness on participants behavior).

In the group phase of the experiment, one of the groups was randomly assigned to be the untrustworthy group and one to be the trustworthy group. When participants invested in the trustworthy group their partner invested back 75% of the time, for the untrustworthy group this was only 25%. The group phase consisted of 60 trials, 30 trials with each group.

In the character phase, participants met four novel partners, two from each group. Of the four partners, one from each

group behaved as in the group phase. The remaining two behaved as a member of the opposite group. This ensured that participants met one partner from a trustworthy group that seemingly had changed to be untrustworthy and one partner from the untrustworthy group that seemingly had changed to be trustworthy. Participants interacted with each partner 12 times in a pseudorandom order, giving 48 trials.

Trial structure Each trial had the same structure (Fig. 1). The participants were assigned 10 points and the picture of their partner was shown as they were asked what they wanted to do. The participants had two options; (1) keep the points, thereby choosing not to engage with the partner and gaining the 10 points that were assigned or (2) invest the 10 points in the partner. If they chose to invest their points, the outcome was dependent on the partner's choice. If their partner invested back, participants gained 20 points and were told that their partner also gained 20 points. If their partner instead chose to keep the invested points, the participants gained no point and were told that their partner gained 40 points. If participants choose to keep the points they were not informed of the partner's actions but if they invested they were.

Participants were informed that the points they gained in the character phase were used in a lottery that occurred at the end of the experiment. The more points they gained, the higher the chance of winning. If participants won in the lottery they were awarded with and extra movie ticket (approximate value, 130 SEK).

At end of the group phase participants were asked to select which group they would rather play an additional round with; how trustworthy they thought each group was and how many times they estimated that partners had invested back to them. At the end of the character phase participants were again asked which group they would rather play with, and to rate each individual partner on trustworthiness and typicality (-5 to +5).

At the end of the study, several psychometric scales were administered, but those data are not reported here.

Reinforcement learning

We fit several reinforcement learning models to participants data in the group phase to provide a computational account of their learning. Models were fit to each individual participant's data using maximum likelihood estimation as implemented in the **mle2** function from the **bbmle** package in R using the L-BFGS-S optimizer. Models were compared by calculating Akaike Information Criteria weights (AICw) for each model (Wagenmakers & Farrell, 2004), where a higher score indicates a better fit. Model selection was done by counting the number of times a model had a higher score compared to the other models for each participant.

We tested four simple models based on the existing literature on trust learning (Fareri et al., 2015), a standard RW model, a bias model, a loss gain model and a combined model. The RW model functions as a baseline model. The bias model includes a term which can be interpreted as capturing participants' propensity to trust. The loss-gain model adds flexibility to how state trust learning is modelled, providing a stronger competitor model to the bias model.

We assumed that participants made a choice to invest or keep their points on each trial (t) depending on the expected value (EV) of the outcomes, which for invest decisions was the likelihood of the partner reciprocating (P_r) multiplied by the payoff (20).

$$EV_i(t) = P_{r,group}(t) * 20$$

Initial probabilities were set to 0.5. For keep decisions the EV was simply 10. Choices were calculated using the softmax rule, for example, the probability of investing was given by:

$$p(invest) = \frac{e^{EV_i * \beta}}{e^{EV_i * \beta} + e^{10 * \beta}}$$

Where β is a temperature parameter which controls the stochasticity of choice.

RW Model For the simplest model, we used a simple Rescorla-Wagner update rule, updating participants' expectations of the probability that a partner would reciprocate:

$$P_{r,color}(t+1) = P_{r,color}(t) + \alpha * (R - P_{r,group}(t))$$

Where R was the observed reciprocation, taking values 1 for reciprocation and 0 otherwise, and α is a free parameter indicating a participant's learning rate.

Bias Model This model included a bias term, θ, which allowed for participants to place an additional value on reciprocation (Fareri et al., 2015). This would in turn bias their choices towards investing over and above any learning. In this model the EV equation was thus:

$$EV_i(t) = P_{r,color}(t) * (20 * \theta)$$

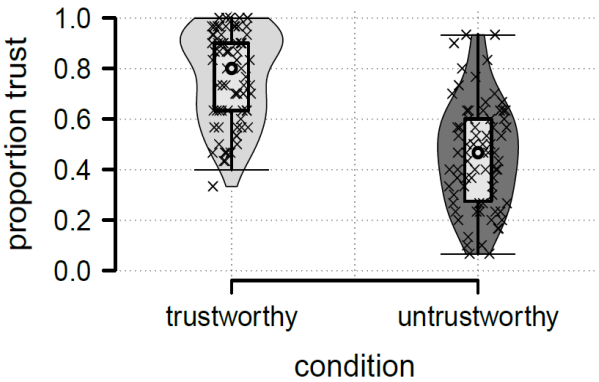

Figure 2. Proportion of trust decisions by participants when facing partners in the trustworthy and untrustworthy groups during the learning phase.

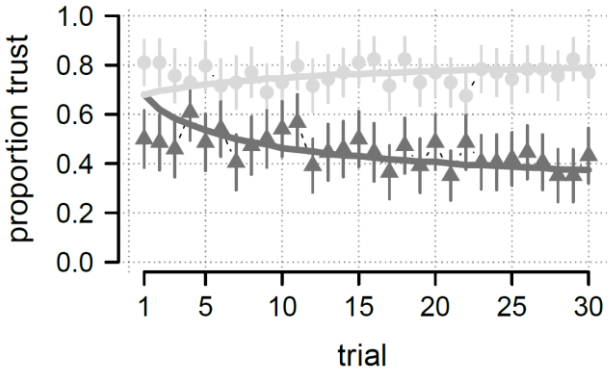

Figure 3. Trial-by-trial choices (points) and 95% confidence intervals (segments) from the learning phase. Thick lines represent predictions from the Bias model. Light, trustworthy group, dark, untrustworthy group.

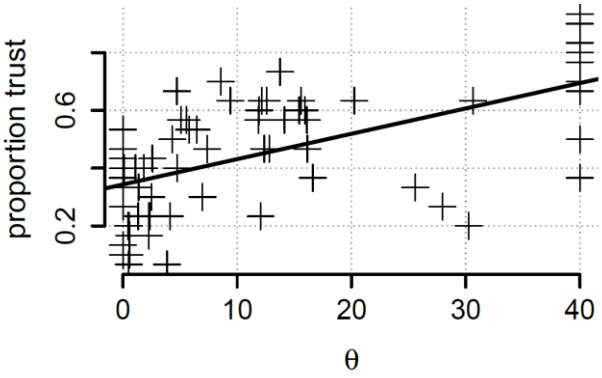

Figure 4. θ values reflecting reciprocation bias fitted to group phase data and proportion trusting choices to the untrustworthy group.

Loss-gain Model This model allowed participants to learn at different rates if the partner reciprocated or not, as has been suggested previously (Chang et al., 2010). The update equation was altered to depend on the value of *R*, for *R*=1 the learning rate was set to α_{gain} and otherwise α_{loss}.

Combined Loss-gain and Bias Model The final model combined the modifications of both the bias and loss-gain models.

Analysis

Participants' choice and response time data were analyzed using multi-level regression models as implemented by the **brms** package (Bürkner, 2016). Models were fit with the maximal random-effects structure using zero-centered weakly informative priors. Coefficients were assessed using 95% credible intervals and Bayes Factors calculated using the Savage-Dickey ratio.

Results

Group Phase

We first analyzed participants' behavior in the initial group phase of the experiment.

Trust Behavior Trust was defined as the number of invest decisions participants made. Overall participants trusted in 76% of trials when interacting with the trustworthy group and in 45% of trials when interacting with the untrustworthy group, see Fig. 2. A mixed-effects logistic regression confirmed the robust differentiation between conditions, b = 1.62, *95% CrI* = [1.22, 2.05], $BF_{10} = 2.79*10^{19}$. Participants were more varied in their responses to the untrustworthy group compared to the trustworthy group (Fig. 2).

Response Times Participants took longer time to make choices when facing untrustworthy group (M = 2.1s, SD = 2.5) compared to when facing the trustworthy group (M = 1.9s, SD = 2.2). Modelling the response times using an ex-Gaussian distribution (Heathcote, Popiel & Mewhort, 1991), revealed differences between conditions in the *tau* parameter of the exponential component of the distribution, reflecting the size of the tail of the RT distribution, b = -0.14, *95% CrI* = [-0.22, -0.057], BF_{10} = 66.53 (longer tail for untrustworthy condition). There was no effect on the *mu* parameter of the distribution (BF_{10} = 0.03), indicating no difference in mode.

Reinforcement Learning Model We found that a model which included a static bias term ('Bias model') boosting the expected value of reciprocation performed best, although closely followed by a model which included different learning rates for reciprocation compared to keep outcomes (see Table 1).

The selected Bias model captured the overall trial-by-trial dynamics of participants' choices (see Fig. 3), indicating a good fit. On an individual level, the fitted size of the reciprocation bias parameter, theta, strongly correlated with participant's tendency to make trusting choices towards the untrustworthy group, r = 0.60, $p = 1.72*10^{-8}$, see Fig. 4), but not for the trustworthy group, r = 0.22, p = .062).

Table 1. Model comparison (number of participants a model was best for) and average parameter values.

Model	AICw wins	α	α_{loss}	α_{gain}	β	θ
RW	20	0.14	-	-	0.52	-
Bias	26	0.20	-	-	0.50	12.5
LG	22	-	0.17	0.54	0.44	-
Bias+LG	6	-	0.26	0.37	0.44	14.9

Character Phase

Trust Behavior We first analyzed participants' trust behavior with respect to each of the four individual partners they met in the character phase using a mixed-effects logistic regression model. Participants differentiated between the individual partners based on actual trustworthiness very well, b = 1.91, *95% CrI* = [1.53, 2.30], $BF_{10} = 1.1*10^{16}$. However,

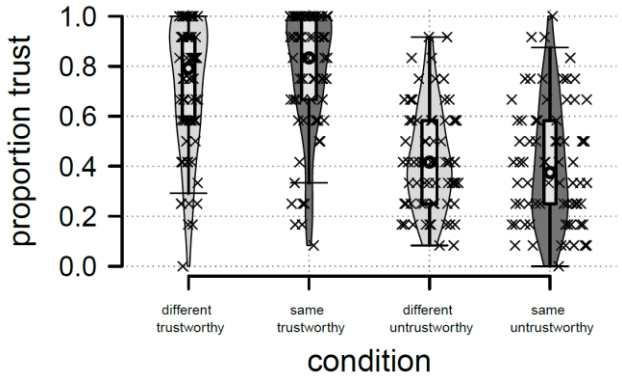

Figure 5. Proportion of trust decisions by participants when facing the four individual partners during the character phase. Same and different indicates how the partner was behaving compared to the expectation based on their group membership from the group phase.

we did not observe an interaction between actual trustworthiness and change (b = 0.54, *95% CrI* = [-0.052, 1.15], BF_{10} = 1.31), which would have reflected different adaptation to the partners' behavior depending on their group membership. Overall participants trusted the trustworthy partners belonging to the trustworthy ('same') group 78% of trials compared to 72% when the individual belonged to a group who had been untrustworthy previously ('different', see Fig. 5). For the untrustworthy partners, participants trusted partners behaving same as previously 41% of trials while the partners behaving differently 43% of trials. This indicates that participants were overall effective in ignoring the previously learnt information about the groups and, generally interacted with the individual partners as appropriate.

However, examining behavior on the very first trial revealed a different pattern. For the partners who belonged to the trustworthy group, participants initially trusted to a high degree, 89%, as would be expected – slightly higher than final average trust rate during the group phase (see Fig. 3). For the partners who belonged to the untrustworthy group, participants also exhibited a high degree of initial trust – 66%, far removed from their final behavior in the group phase.

To better understand this behavior, we regressed participants' first trial behavior with the partners from the untrustworthy group on their fitted θ values and their final estimates of the reciprocation probability (P_r) of the untrustworthy group using a mixed-effects logistic regression. Surprisingly, we found that P_r did not reliably predict initial trust (b = 0.14, *95% CrI* = [-0.40, 0.69], BF_{10} = 0.62). Instead, θ weakly predicted initial trust, b = 0.49, *95% CrI* = [-0.05, 1.0], BF_{10} = 2.9 (see Fig. 6), suggesting that participants' propensity to trust dominated when facing the prospect to interact with a novel partner even when they came from a previously untrustworthy group.

Response Times We again analyzed participants response times using an ex-Gaussian distribution. We found clear

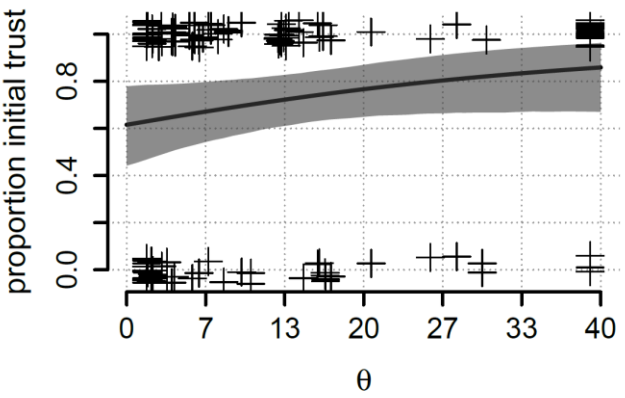

Figure 6. Posterior predictive plot of relationship between θ and initial trust decisions on the first trial to the learned untrustworthy partners.

effects of trustworthiness on the *tau* parameter, b = -0.23, *95% CrI* = [-0.31, -0.14], BF_{10} = $7.43*10^{11}$, reflecting the long response times when participants face the untrustworthy partners (same: M = 1.75s, different: M = 1.74s) compared to when facing the trustworthy partners (same: M = 1.48s, different: M = 1.62s). In line with this, we found a weak trustworthiness X change interaction effect, b = -0.15, *95% CrI* = [-0.33, 0.01], BF_{10} = 2.1. There were no effects on the *mu* parameter of the distribution (all BF_{10} < 0.49).

Partner Trust Ratings Finally, we examined the individual trustworthiness ratings of each partner. The two trustworthy partners were rated the highest, with the partner behaving same as his group membership would predict being rated higher, M = 2.2, SD = 2.5 than the partner behaving differently, M = 1.5, SD = 2.6. The two untrustworthy partners were rated as such, with the one behaving differently being rated M = -2.6, SD = 2.2 and the one behaving as expected rated M = -2.4, SD = 2.4. Analysis revealed a main effect of trustworthiness (b = 4.32, *95% CrI* = [3.8, 4.9], BF_{10} = $8.82*10^{48}$). There was weak evidence for partners behaving differently than expected to be rated lower (b = 0.40, *95% CrI* = [-0.08, 0.84], BF_{10} = 1.9), nevertheless suggesting that while participants didn't differentiate in their behavior between partners that behaved as their group and those who didn't, some distrust was manifest due to this change.

Discussion

We investigated trust dynamics in the context of learning about the trustworthiness of two arbitrary group followed by repeated interactions with select individuals from those groups. We replicated previous findings indicating that RL models can adequately capture the learning process, and that trust consists of considerations beyond just calculating expected values of payoffs. Instead, we found that a model containing a fixed bias term, upweighting the expected value of partner reciprocation best explained participants behavior. The bias parameter (θ) might capture participants propensity

to trust beyond the immediate situational factors (cf. Mayer, Davis & Schoorman, 1995).

We find that response time were generally faster in response to individuals and groups known to be trustworthy. Response times in decision making capture the amount if evidence available to the decision maker, where faster decisions indicate easier decisions (cf. Krajbich, Bartling, Hare & Fehr, 2015). Response times have not typically been reported in relation to trust games, and the findings here indicate that modeling them fully might yield further insights into the dynamics of trust decisions.

In the subsequent character phase where participants interacted with novel partners drawn from the previously encountered groups, we found that participants rapidly adapted to the partners' actual behavior irrespective of group membership. Even if participants' behavior differed on the first trial, participants were generally more biased towards initially trusting new individuals than would be predicted by their learning, as captured by the reciprocation bias term. Similar findings have been observed in relation to public goods and prisoner dilemma games where players cooperative behavior will "restart" with new partners following its deterioration as typically seen during repeated play (Andreoni, 1988).

Further, as discussed in the introduction, people are less prone to act on stereotypical information given highly individuating information (Rubinstein et al., 2018). It might seem that the novel partners do not represent such a case. However, with the prospect of repeated interaction with the same person, an initial trusting choice might be considered a small price for highly diagnostic information about the partner. Hence, the task structure with binary choices and outcomes might provide additional impetus to participants to display high initial trust. While many real-life situations might have this dichotomous character, others won't, where instead trust will be partly reciprocated. To further understand how learned group information influences character learning graded trust decisions and reciprocation will need to be investigated.

Acknowledgments

This research was supported by the European Research Council (Independent Starting Grant 284366; Emotional Learning in Social Interaction) and the Knut and Alice Wallenberg Foundation (KAW 2014.0237) to A. Olsson and the Swedish Research Council (2016-06793) to P. Pärnamets.

References

Andreoni, J. (1988): "Why free ride? Strategies and learning in public goods experiments," *Journal of Public Economics*, *37*(3), 291-304.

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

Bürkner, P. C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

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

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science*, *9*(3), 245-274.

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

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2).

Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Reviews of Psychology*, *47*, 237–271.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Read Montague, P. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78–83.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, 6, 7455.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122–1135.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162.* 8–13.

Peysakhovich, A., & Rand, D. G. (2015). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Scienc*e, *62*(3), 631-647.

Rubinstein, R. S., Jussim, L., & Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology*, *75*, 54–70.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *PNAS*, *108*(19), 7710–7715.

van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*, 796–803.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196.

Zarolia, P., Weisbuch, M., & McRae, K. (2017). Influence of indirect information on interpersonal trust despite direct information. *Journal of personality and social psychology*, *112*(1), 39.