(Re)building cooperation: Effects of a cognitive intervention on cooperative behavior in games

Ismail Guennouni[1], Samuel Dupret[1], Quentin JM Huys[2], Maarten Speekenbrink[1]

Abstract

Social trust is an important building block of strong social bonds, and its absence is a risk factor for social dysfunction. As such, interventions to foster and strengthen trust-based cooperation are highly desirable. Using the Repeated Trust Game paradigm, we assess the effectiveness of a cognitive intervention derived from Dialectical Behavior Therapy. The intervention's goal was to repair the potential breakdown of cooperation from a pre-programmed, one-off defection by the opponent. Over two games, participants are given the role of the trustee and face what they believe are two different players. In between games, they either receive a brief cognitive intervention or not. Post-intervention, participants showed more cooperative behavior both before and after defection by the opponent. Analysing participants' actions with a hidden Markov model shows that participants in the intervention group had a higher proportion of cooperative states than participants in the control group. This is consistent with participants inferring from the intervention that pro-social and trustworthy behavior may generally provide more beneficial outcomes to them in the long run.

[1] *Department of Experimental Psychology, Division of Psychology and Language Sciences, UCL.*
[2] *Division of Psychiatry and Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology, UCL.*

---

1 Introduction

Determining the other's goals and intentions is fundamental to successful social interaction. Inference of these latent motives is fraught with uncertainty (Barnby, Dayan, and Bell 2023; FeldmanHall and Shenhav 2019), as they are not directly observable and we may require extensive prior experience with the person to infer them. Absent a history of interaction, we may decide to trust that the other's goals and intentions are aligned with our own. Such trust is risky and if misplaced, it can come at a high cost. But not trusting others is also risky as we may forego mutally beneficial cooperation.

Prior research emphasises the importance of social trust in determining why some people fare better than others physically and mentally (Giordano and Lindström 2016; Meng and Chen 2014). Social trust affects the health status of individuals by reinforcing social support networks, maintaining community norms and facilitating collective action. Research into the determinants of psychopathology has linked trust-based constructs to the emergence of mental health disorders. Fonagy and Allison (2014) identified epistemic trust – the belief in the authenticity and personal relevance of interpersonally transmitted knowledge – as an important function of early attachment relationships. It allows the receiver of social information to let go of their natural self-protective vigilance, which can become pathological hypervigilance after adverse or traumatic experiences and a key factor in the emergence of several mental health disorders (Fonagy and Campbell 2017). Since a lack of epistemic trust is a risk factor for social dysfunction (Fonagy and Campbell 2017), and given the importance of trust for building and maintaining strong social bonds, interventions that foster and strengthen trust-based cooperation would be highly beneficial to society. Such interventions would allow people to more easily repair broken relationships and continue harvesting the benefits of cooperation even in the presence of accidental or intentional social norm violations.

A well-established paradigm in the study of trust is the Repeated Trust Game [RTG; Joyce, Dickhaut, and McCabe (1995)]. In this game, the "investor" decides how much of an endowment to send to the other player (the "trustee").

The amount that is sent is tripled and the trustee decides, in return, how much of the tripled amount to send back to the investor. The Nash equilibrium for a single-round version of the game is for the investor to send nothing. In the repeated version, rewards for both players are maximised if they build trust and share the benefits of higher investments. If the investor is rewarded for taking the risk of sending an investment, they are likely invest again in future rounds. But if the investor obtains a low return on their investment, they will likely reduce future investments, thereby diminishing the potential gains for both players. To encourage the emergence and maintenance of trust in this setting, some studies focused on modification of the game such as introducing a third-party who monitors the actions of the other players (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler and Haruvy 2017). Others chose to intervene directly on the participants. For example, Drążkowski, Kaczmarek, and Kashdan (2017) found that trust (as measured by the amount invested) was increased when participants were asked to think about and write down five things that they were grateful for, and Burnham, McCabe, and Smith (2000) found that trust increased when participants were primed with the concepts of friend and foe.

Whilst these results show that it is possible to improve cooperative outcomes at the start of the game, they do not address how to repair a breakdown of trust after intentional or accidental non-cooperative actions by the players. Cooperation in the RTG can easily break down after a transgressive behavior such as a nil or very low investment by the investor or a return by the trustee below the investment sent (Bendor, Kramer, and Stout 1991). Such breakdown of cooperation are particularly evident when the trustee suffers from social disorders such as Borderline Personality Disorder [BPD; Lieb et al. (2004)]. Trustees with BPD fail to engage in trust-repairing behaviors such as coaxing the investor by signalling trustworthiness via sending high returns. This failure may be due to BPD trustees not realising that low returns in the game violate social norms (King-Casas et al. 2008).

To devise interventions to repair social trust, we can take inspiration from cognitive interventions championed by successful therapies that aim to improve social dysfunction in BPD. Psychotherapies such as Mentalisation Based Therapy (Allen and Fonagy 2006) and Dialectical Behavior Therapy (Linehan 1993) have been shown to improve social skills in BPD (Gunderson et al. 2018). However, response to these treatments is highly variable, and determining which interventions are effective for individual patients is challenging (Rudge, Feigenbaum, and Fonagy 2020; Arch et al. 2012). One promising approach is the study of how specific components of psychotherapeutic treatment affect quantitative markers of behavior such as those inferred through computational models (Huys, Maia, and Frank 2016; Reiter et al. 2021; Dercon et al. 2022). Combining the use of specific cognitive probes inspired by therapeutic interventions and computational models of behavior may allow us to uncover the cognitive mechanisms targeted by common forms of psychotherapy. In turn, this may provide the basis for choosing effective psychotherapeutic interventions for given individuals.

In this study, we assess the effectiveness of a cognitive intervention aimed at repairing the potential breakdown of cooperation from a low investment sent by the investor. The intervention focuses on explaining the potential harm from reciprocating non-cooperative actions and suggesting a non-impulsive course of action to coax the investor back into cooperation. Participants are given the role of the trustee and are randomly assigned to either a control or intervention group. They play two instances of the Repeated Trust Game with two different computerized investors. Between the first and second instance of the game, they either receive the cognitive intervention (intervention condition) or perform an unrelated task solving anagrams (control condition). The computerized investor in both instances of the game was programmed to play according to a hidden Markov model estimated from real players' data in prior research. A key aspect of this model is that the actions of the investor depend on a latent "trust state" which reacts to the trustees returns. To foreshadow our results, we find that the intervention led to more cooperative actions (higher returns) by the participants and countered a tendency to send back lower returns after a transgression from the investor. However, we found no evidence that the effects of the intervention transfer to a different (Repeated Prisoner's Dilemma) game.

2 Method

2.1 Participants

A total of 318 participants were recruited on the Prolific Academic platform (prolific.co). The mean age of participants was 31.3 years, with a 9.9 years standard deviation. Participants were paid a fixed fee of £5 plus a bonus payment dependent on their performance that averaged £0.71.

2.2 Design and Procedure

The experiment had a 2 (Condition: Intervention or Control) by 3 (Game: Trust-Game Pre Intervention, Trust-Game Post Intervention, Prisoner's Dilemma Post Intervention) design, with repeated measures on the second factor. Participants were randomly assigned to one of the two levels of the first factor. The games were designed and implemented online using Empirica (Almaatouq et al. 2021).

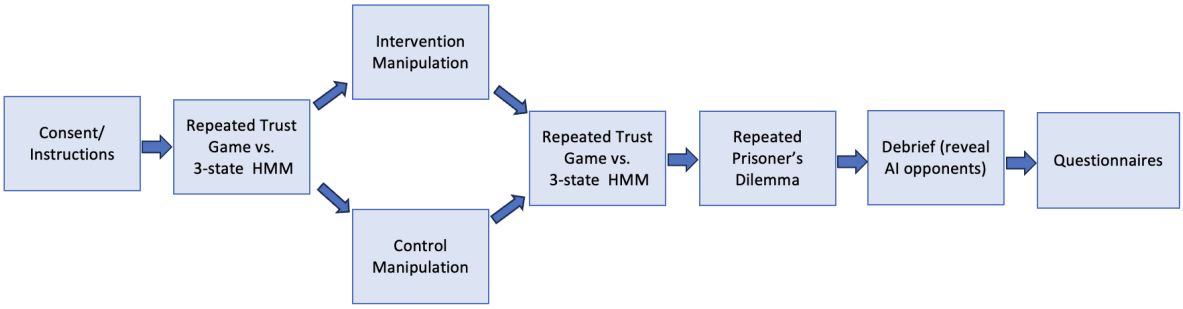

Figure 1: Experiment overview. After playing 15 rounds of the RTG as the trustee, participants were randomized to either be part of the control or intervention condition. They then played the second set of 15 rounds of RTG (again as the trustee) to examine intervention effects, and 7 rounds of a repeated Prisoner's Dilemma to examine generalization of intervention effects. Finally, participants answered questionnaires and were debriefed.

2.3 Tasks and Measures

2.3.1 Repeated Trust Game

Participants played a 15-round Repeated Trust Game (Joyce, Dickhaut, and McCabe 1995) in the trustee role against a computer-programmed investor. On each round the investor is endowed with 20 units and decides how much of that endowment to invest. This investment is tripled and the trustee then decides how to split this tripled amount between them and the investor. If the trustee returns more than one third of the amount, the investor makes a gain.

The strategy of the computerised investor was modelled on behavior of human investors in the Repeated Trust Game (RTG) over 10-rounds with the same (human) opponent. Full detail on the datasets used in the Supplementary Information. Using this data, we estimated a hidden Markov model (HMM) on investors' behavior with three latent states. Each latent state was associated with a state-conditional distribution over the possible investments from 0 to 20 (Figure 2.A). These distributions reflect "low-trust", "medium-trust", or "high-trust". Over rounds, the investor can move between states, and the probability of these transitions was modelled as a function of their net return (i.e return - investment) in the previous round (see Figure 2.B). In order to instigate a potential breakdown of trust, thereby allowing us to probe efforts to repair trust, the computerised agent was programmed to provide a low investment on round 12 (pre-intervention) or round 13 (post-intervention). On all other rounds, the investor's actions were determined by randomly drawing an investment from the state-conditional distribution, with the state over rounds determined by randomly drawing the next state from the state-transition distribution as determined from the net return on the previous round (disregarding the net return immediately after the pre-programmed low investment rounds). The starting state in each instance of the game was the XXX.

On each round, immediately after being informed of the investment sent, participants in the intervention condition were asked to provide an evaluation of their emotion in terms of valence (from negative to positive) and arousal (from low to high). Participants in the control condition were asked to evaluate the investment in terms of speed (from slow to fast) and magnitude (from low to high). These evaluations were made by clicking on a two-dimensional field with labelled axes.

2.3.2 Repeated Prisoner's Dilemma

To ascertain whether any effect of the intervention would transfer to a different game, participants played 7 rounds of a Repeated Prisoner's Dilemma (RPD). In each round, participants could choose between a cooperative action

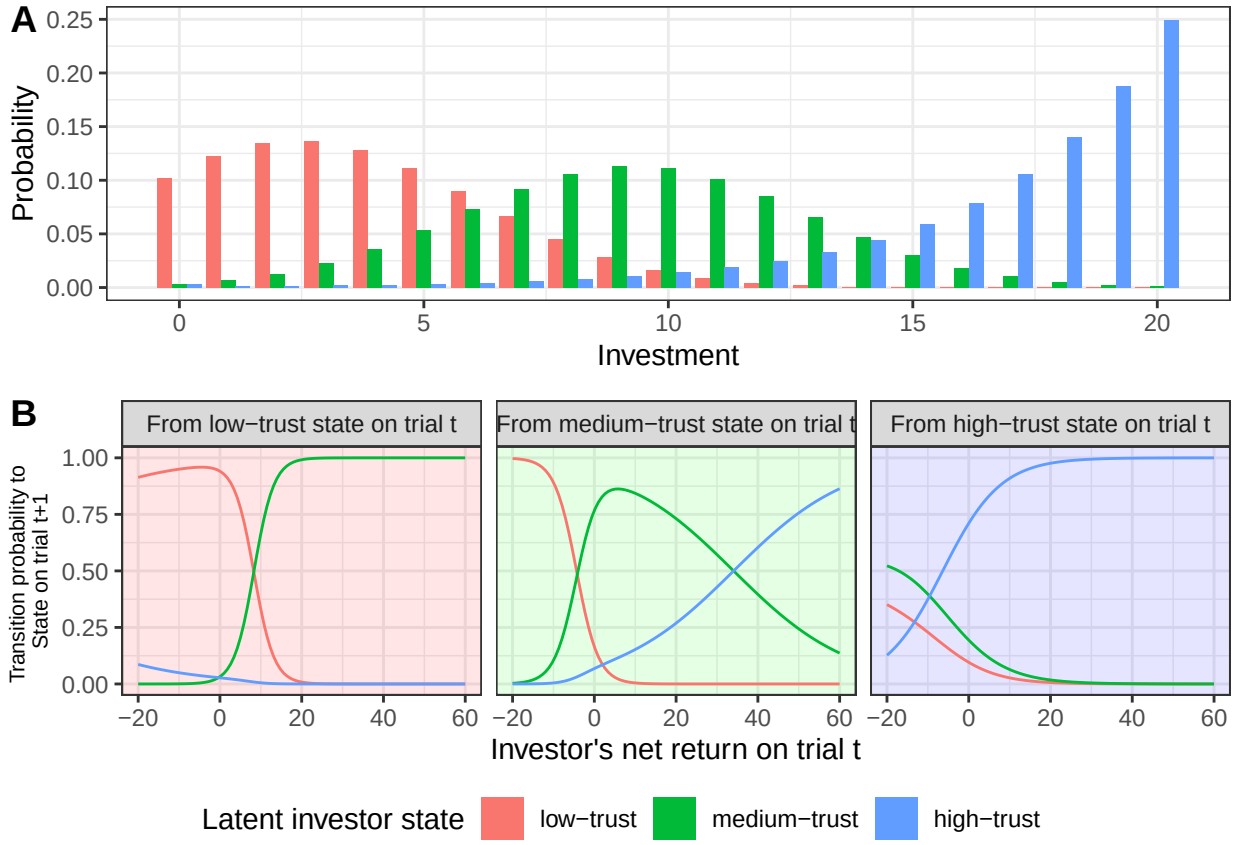

Figure 2: We construct the artificial investor agent by fitting a three-state hidden Markov model to data of human investors engaged in the 10 round Repeated Trust Game against human trustees. From the fitted HMM, we get the distribution of investments by the artificial investor agent conditional on its latent state as shown in Panel A. The fitted HMM also yields the transition probability of the agent to a state on trial t+1 as a function of the net return (difference between the investment sent and the amount received in return) on trial t as shown in Panel B. Each plot in Panel B represents a different starting latent state on trial t, and each line represents the probability of transitioning to a particular state in trial t+1. The policy in each state in Panel A and the way the agent transitions between states as defined in Panel B characterises the artificial agent's behavior in the Repeated Trust Game.

with a reward of 5 (the other player also cooperates) or 1 (the other player defects), or a defect action with a reward of 7 (the other player cooperates) or 2 (the other player defects). The Nash equilibrium for a single-round version is to choose the non-cooperative action. In the repeated version, both players can maximise their reward by choosing to cooperate. In this game, the computerised agent was programmed to act according to a tit-for-tat strategy (Axelrod and Hamilton 1981), starting with a cooperative action and then mirroring what the other player chose in the preceding round. On round 4, the computerised agent was pre-programmed to choose the defect action, regardless of the participant's preceding action.

2.4 Intervention

The intervention was built on Dialectical Behavior Therapy (DBT) skills training, asking patients to reflect on the consequences of actions taken in emotional states (Linehan 2015). Specifically, participants were presented with a hypothetical situation in which they receive a low investment and asked to indicate how they would respond. They were then presented with an educational slide inviting them to consider that the ultimate aim in the game is to maximise their total reward and to reflect on whether punishing the investor for the low investment is beneficial to achieving that aim. Participants were told that punishment can create a negative feedback loop where the other player might trust them even less. An alternative action was suggested, whereby players would respond kindly to such a transgression in the hope of gaining trust from the investor. Participants were then asked whether the information just received would change their behavior in such a hypothetical situation and to justify their answer. Full details on the intervention are provided in the supplementary information.

In the control condition, participants were asked to solve five anagrams ("listen", "triangle", "deductions", "players", "care"). They provided their answers in a free-form text box. The time given to solve the anagrams was the same as that given to respond to questions in the intervention manipulation.

2.5 Procedure

At the start of the experiment (Figure 1), participants provided informed consent and were instructed the study would consist of three phases in which they would face a different other player. Participants were told their goal was to maximise the number of points in all phases. They were not told the number of rounds of each phase. Phase one was a 15 round Repeated Trust Game (RTG) in which participants took the role of trustee, facing the same investor over all 15 rounds. On each round, after being informed about the amount sent by the investor, participants were asked to evaluate their emotion (invervention condition) or the investment (control condition). Participants then decided on how much of the tripled investment to return to the investor, before continuing to the next round. After completing 15 rounds of the RTG, participants rated how cooperative, selfish, trustworthy and friendly they perceived the investor to be (all on a scale from 1 to 10). After phase one, participants in the intervention condition completed the intervention, and participants in the control condition solved anagrams. Subsequent phase two was similar to phase one, with participants being told they would face a new player. Phase three consisted of 7 rounds of the Repeated Prisoner's Dilemma game (RPD), with participants informed they would face a third player. Participants then completed questionnaires related to mentalising abilities, emotion regulation, and BPD traits (see the supplement for details). They were then asked about the strategy in the games, as well as whether they thought the other players were human or computer agents. They were then debriefed and thanked for their participation.

2.6 Statistical analysis

To explore whether participants behaved differently in the RTG after the intervention compared to the control group, we model the percentage return (percentage of tripled investment returned to investor) using a linear mixed-effects model as described below:

$$\begin{aligned} R_{ij} =& \beta_0 + \beta_1 \left(\text{Condition}\right)_i + \beta_2 \left(\text{Game}\right)_i + \beta_3 \left(\text{Investment}\right)_i + \\ & \beta_4 \left(\text{Condition} \times \text{Game}\right)_i + \beta_5 \left(\text{Condition} \times \text{Investment}\right)_i + \beta_6 \left(\text{Game} \times \text{Investment}\right)_i + \\ & \beta_7 \left(\text{Condition} \times \text{Game} \times \text{Investment}\right)_i + b_{0j} + b_{1j} \left(\text{Game}\right)_i + \epsilon_{ij} \end{aligned}$$

where:

- R_{ij}: percentage of tripled investment returned to investor for participant j in observation i
- β_0: intercept
- β_1: effect of Condition (intervention vs. control)
- β_2: effect of Game (RTG game pre vs. post-intervention)
- β_3: effect of Investment
- β_4: interaction effect between Condition and Game
- β_5: interaction effect between Condition and Investment
- β_6: interaction effect between Game and Investment
- β_7: three-way interaction effect between Condition, Game and Investment
- b_{0j}: participant-wise random intercept for participant j
- b_{1j}: participant-wise random slope for Game for participant j
- ϵ_{ij}: error term for participant j in observation i

The model was estimated using the `afex` package (Singmann et al. 2022) in R (R Core Team 2022). More complex models with additional random effects could not be estimated reliably, and as such the estimated model can be considered to include the optimal random effects structure (Matuschek et al. 2017). A similar process was used to establish the random effects structures of other linear mixed-effects models used throughout the statistical analyses. For the F-tests, we used the Kenward-Roger approximation to the degrees of freedom, as implemented in the R package "afex". We Z-transform the Investment variable as centering is beneficial to interpreting the main effects more easily in the presence of interactions.

To model participants' returns in the RTG across games and conditions, we fit various hidden Markov models (Visser and Speekenbrink 2022) to participants' returns using the depmixS4 package (Visser and Speekenbrink 2021) for R. The transition between latent states is assumed to depend on the investment received and a dummy variable to characterise the group that the participant belongs to. Details on how the models are constructed can be found in the supplement. We fit models with different numbers of hidden states, and use the Bayesian Information Criterion (Schwarz 1978) to select the best model.

3 Behavioral results

Average investments and returns prior to the "defection round" (Figure 3) were within the range of reported investments (40-60% of endowment) and returns (35-50% of total yield) in the literature (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011). Mixed-effects analysis on the percentage returns shows a significant main effect of Condition (intervention vs. control), $F(1, 317.20) = 9.53$, $p = .002$, with higher percentage returns in the intervention compared to the control condition. Importantly, we also find an interaction between Condition and Game (RTG pre- vs. post-intervention), $F(1, 318.30) = 26.91$, $p < .001$. Post-hoc tests show an increase in the percentage returned in the intervention condition, pre - post, $\Delta M = -0.03$, 95% CI $[-0.05, -0.02]$, $t(316.54) = -4.85$, $p < .001$, but a decrease in the control condition, $\Delta M = 0.02$, 95% CI $[0.00, 0.03]$, $t(319.69) = 2.34$, $p = .020$ (see Figure 4.A). This indicates the intervention was effective in increasing cooperative behavior. There was also a significant main effect of Investment, $F(1, 9208.68) = 373.23$, $p < .001$, such that higher investments were associated with higher percentage returns. An Investment by Condition interaction, $F(1, 9208.68) = 45.35$, $p < .001$, indicates the positive effect of investment on percentage returns was greater in the control than intervention condition. There was also an Investment by Game interaction, $F(1, 8990.38) = 4.31$, $p = .038$. Finally, we find a three way interaction between Game, Condition and Investment, $F(1, 8990.38) = 24.56$, $p < .001$, showing that the differentiated effect of the investment on the proportion returned by condition is itself moderated by the Game (pre- vs post intervention).

To explore whether the HMM investors behaved differently in the RTG after the intervention compared to the control group, we estimate a linear mixed-effects model of investments sent by the computerised HMM agent with Condition and Game and their interaction as fixed effects, and a similar random effects structure to the returns model. This shows a main effect of Condition, $F(1, 317) = 8.72$, $p = .003$, and Game, $F(1, 317) = 8.32$, $p = .004$. As can be seen in Figure 4.B, investment was higher in the intervention compared to the control condition across games, and higher in the second game compared to first across conditions.

We next analysed returns separately for rounds prior to the pre-programmed defection (rounds 1 to 11 pre-intervention and 1 to 12 post-intervention) and rounds after (rounds 12 to 15 pre-intervention and rounds 13 to 15 post-intervention). Applying the same mixed-effects model as before to returns before the defection largely replicates the results over all trials. Participants in the intervention condition increased their returns in the second game, $\Delta M = -0.04$, 95% CI $[-0.05, -0.02]$, $t(317.20) = -5.19$, $p < .001$. There was no evidence that participants in the

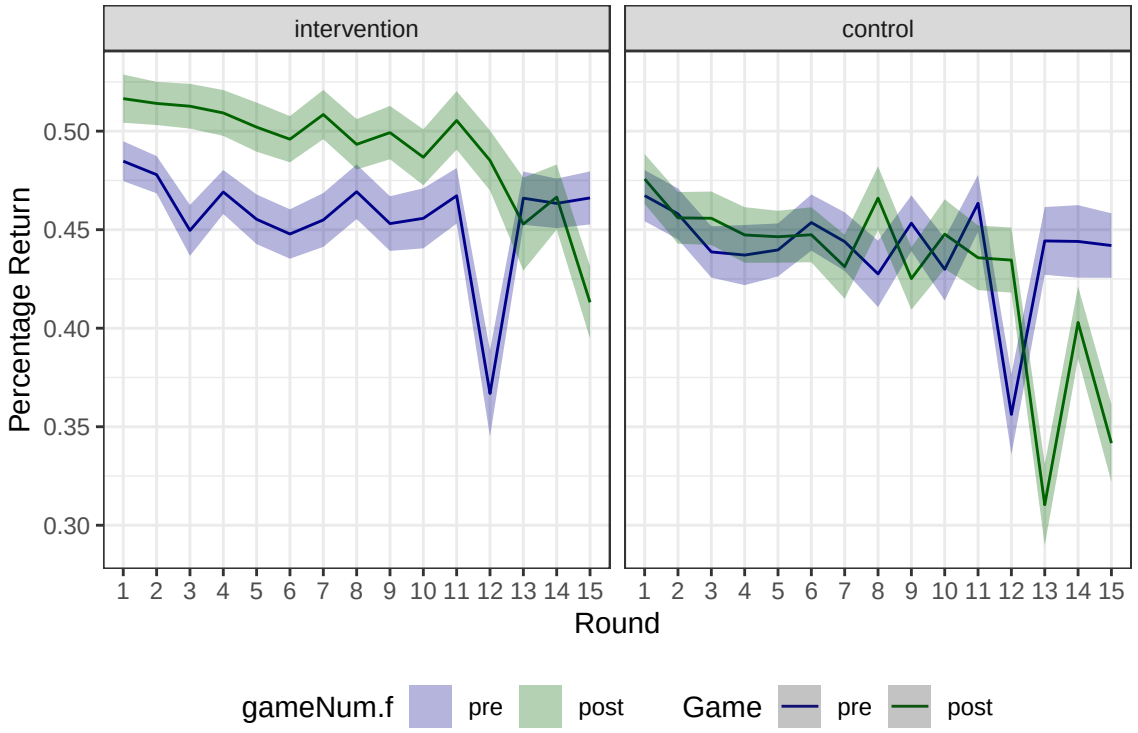

Figure 3: Averages and standard errors of the trustee's return as a percentage of the multiplied investment received by Condition, Phase, and game round. The blue line shows the returns pre-manipulation and the green line post-manipulation. We note a different reaction to the pre-programmed one-off low investment between the two conditions: Whilst there is a dip in returns pre-manipulation for both conditions, post manipulation we see higher returns in the intervention condition compared to the dip in returns seen in the control condition in the right panel

control condition changed their returns in pre-defection trials between the first and second RTG, $\Delta M = 0.00$, 95% CI $[-0.01, 0.02]$, $t(318.64) = 0.21$, $p = .833$. Full results are provided in the supplement.

Applying the same model to the returns after the defection (see the supplement for full results), we again find a significant interaction between Condition and Game. Participants in the control condition decreased their post-defection returns from the first to the second RTG, $\Delta M = 0.07$, 95% CI $[0.04, 0.10]$, $t(325.03) = 5.17$, $p < .001$. There was no significant change for participants in the intervention condition, $\Delta M = -0.01$, 95% CI $[-0.04, 0.02]$, $t(313.90) = -0.83$, $p = .407$.

Taken together, we find that participants in the control condition sent lower percentage returns in the second game, despite the HMM investor sending on average higher investments in that game. Those in the intervention group returned higher percentage returns in the second game, with the investor also sending higher investments. These higher returns in the intervention compared to the control condition were not purely driven by reciprocity towards higher investments, since we found a Condition by Game interaction whilst controlling for investment in the model, and a reduced effect of investment in the intervention condition. The intervention did not increase participants' returns after the defection by the other player. Instead, it countered the tendency shown by participants in the control condition to lower returns after the defection in the second game compared to the first game.

3.1 Emotion self-reports

To assess the impact of the intervention on the participant's emotional reactions, we used linear mixed-effects models (one for valence, and one for arousal) with fixed effects for Game (pre vs. post intervention) and Investment, as well as interaction between Investment and Game, with participant-wide random intercepts and random slopes for Game. This showed that higher investments were associated with more positive emotions, $F(1, 3448.17) = 2108.08$, $p < .001$, and higher arousal, $F(1, 3453.24) = 1505.03$, $p < .001$. In addition, the positiveness of emotion declined between the two games, $F(1, 117.20) = 17.99$, $p < .001$, as did arousal, $F(1, 117.19) = 5.52$, $p = .021$. There was no indication that the effect of the investment on either aspect of emotion was affected by the intervention, as there was no interaction between Investment and Game on valence, $F(1, 3419.70) = 1.49$, $p = .222$, or arousal, $F(1, 3409.69) = 0.12$, $p = .726$. This indicates that participants in the intervention condition returned higher amounts post-intervention, despite their emotional reaction to investments remaining largely the same.

3.2 Evaluation of the investor

For the Investor evaluation, we estimate a mixed-effects model for participants ratings with Game and Condition as fixed effects and participant-wise random intercepts as random effects. Participants rated the HMM investor in the second game as less cooperative ($\Delta M = 0.42$, 95% CI $[0.15, 0.69]$, $t(317) = 3.10$, $p = .002$), less trustworthy ($\Delta M = 0.43$, 95% CI $[0.16, 0.70]$, $t(317) = 3.19$, $p = .002$), less friendly ($\Delta M = 0.40$, 95% CI $[0.17, 0.64]$, $t(317) = 3.36$, $p < .001$) and more selfish ($\Delta M = -0.36$, 95% CI $[-0.61, -0.10]$, $t(317) = -2.76$, $p = .006$), than the HMM investor in the first game. Participants in the intervention condition rated players higher than those in the control condition on cooperativeness ($\Delta M = 0.40$, 95% CI $[0.00, 0.80]$, $t(317) = 1.95$, $p = .052$) and lower on selfishness ($\Delta M = -0.41$, 95% CI $[-0.80, -0.02]$, $t(317) = -2.04$, $p = .042$). There was no evidence for an interaction effect between Condition and Game on any of the attributes.

When asked during debrief whether they thought the investors they faced were Human or not, 40% of participants thought they were either facing a human or were not sure of the nature of the opponent. Many answers reflected participants projecting human traits such as "spitefulness" or "greed" onto the artificial opponent's behavior.

3.3 Transfer to the Repeated Prisoner's Dilemma game

We next asked whether the intervention had any discernible effect on participants' behavior in a different, Repeated Prisoner's Dilemma game. Predicting the probability of a cooperative action with a logistic mixed-effects regression model, with Condition and Phase (before or after defection trial) as fixed effects and a random intercept for participants, showed a decline in cooperation after defection by the other player, $\chi^2(1) = 237.67$, $p < .001$, but no evidence for an overall different cooperation rate in the intervention condition compared to the control condition, $\chi^2(1) = 0.10$, $p = .754$, or a different response to defection between the conditions, $\chi^2(1) = 0.23$, $p = .635$. As such, there is no evidence that the intervention affected behavior in this game.

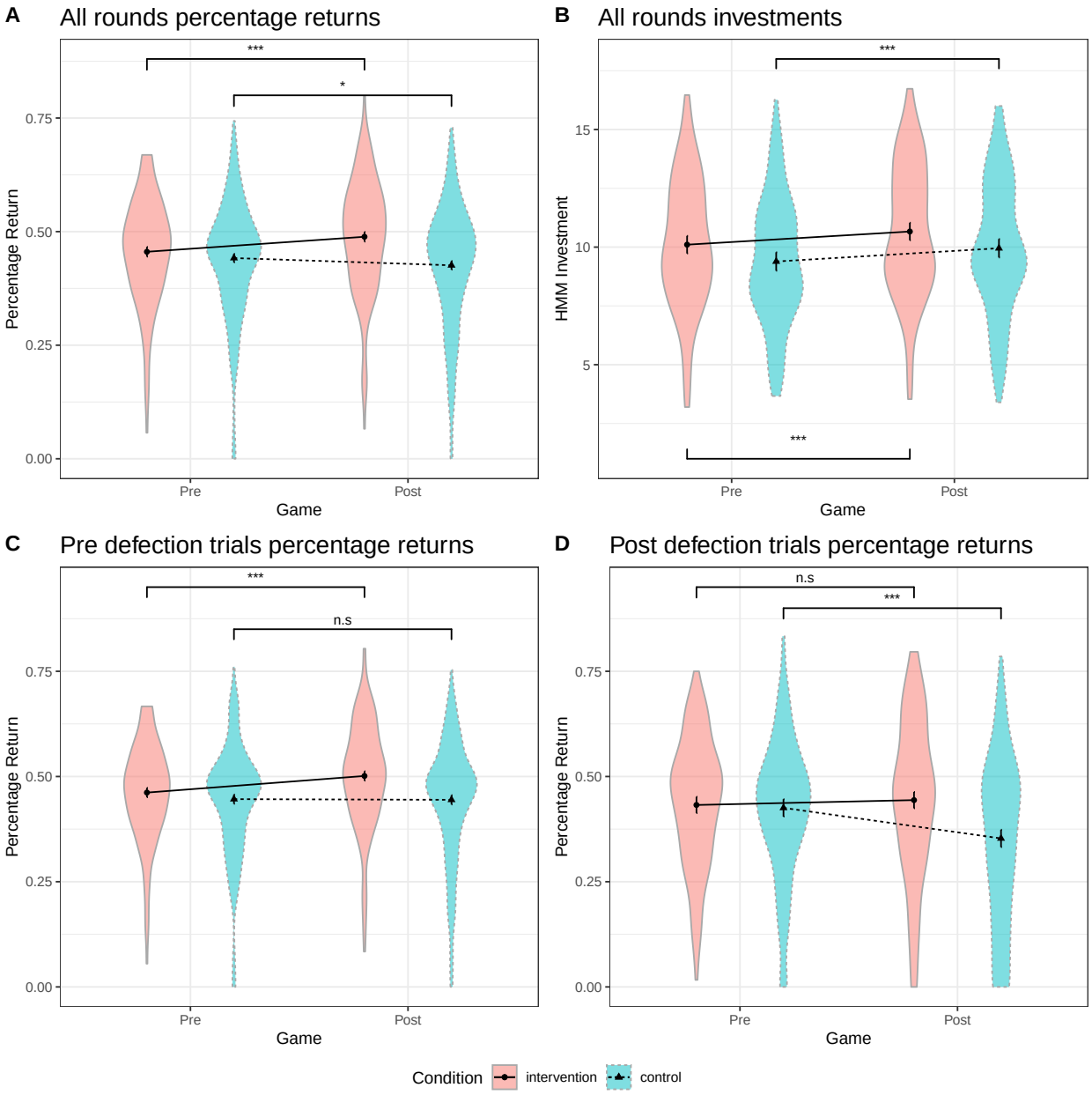

Figure 4: Participants in the Intervention condition returned higher proportions of the multiplied investment received in the second game compared to the first game over all ronds (Panel A), whilst those in the Control condition sent back lower returns. Focusing on pre-defection trials (Panel C), those in the Control condition returned similar amounts between games, while returns were higher in the Intervention condition in the second game. Post-defection (Panel D), returns were similar for those in the Intervention condition while those in the Control condition returned lower amounts in the second game. Investments by the HMM agent (Panel B) were higher in the second game compared to the first game across conditions. All panels show marginal means and distributions of either investments or percentage returns across participants by Game and Condition

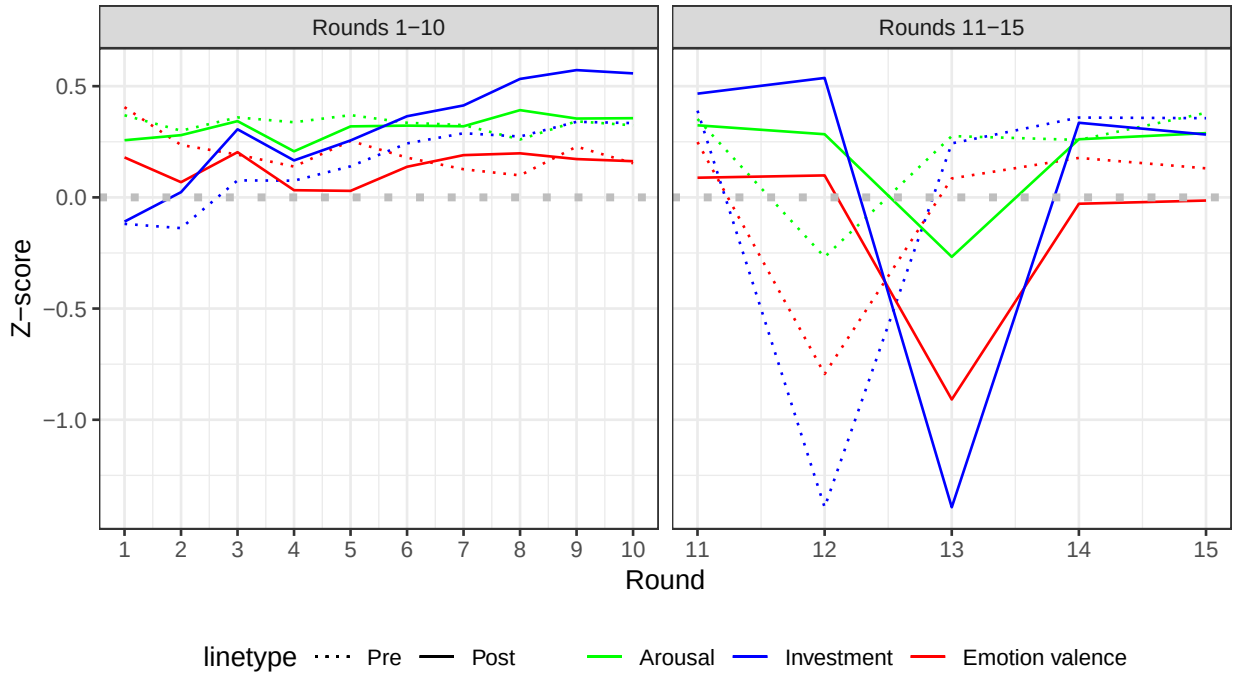

Figure 5: Self-reported emotion valence and arousal as well as investment z-scores for each round of the Repeated Trust Game averaged across participants in the intervention condition only. The participants' emotion reaction measured during the investor's defection round were similar before (round 12) and after (round 13) the intervention.

4 HMM analysis of participant returns

We used hidden Markov models (HMMs) to further assess differences between the intervention and control condition in participants' reactions to the investor in the Repeated Trust Game. As in the models for the investor, these HMMs assume behavior is governed by latent states, with participants' switches between states now dependent on the investments made. We also allowed for differences between games and conditions in how investments govern state transitions: We fitted five main models which all regressed state transition probabilities onto investments, as well as on additional contrast-coded predictors for condition and/or Game. In the most complex model (HMM-full), the transition probabilities were allowed to differ between all four combinations of Game and condition. The HMM-coax model allowed differences between post-intervention and the other three conditions (pre-intervention, pre-control, post-control) treating these latter conditions as the same. Similarly, the HMM-ctrl model allowed differences between post-control and the other three conditions. The HMM-prepost model allowed differences between the first and second RTG. Finally, the HMM-inv model did not allow transition probabilities to differ between conditions or games, modelling them only as a function of investment.

As the number of hidden states was unknown, we estimated models with 2 to 7 latent states for the most complex HMM-full model, and used the BIC to compare them. The best fitting HMM-full model according to the BIC had 6 latent states. Further details on the HMMs and estimation procedure are provided in the Supplementary Information.

Focusing on the HMM-full model with 6 latent states, likelihood ratio tests showed that the HMM-full model fits significantly better than HMM-ctrl ($\chi^2(60) = 108.44$, $p < .001$), HMM-coax ($\chi^2(60) = 129.85$, $p < .001$) and HMM-prepost ($\chi^2(60) = 110.01$, $p < .001$).

Taking the best-fitting 6-state HMM-full model, we used a local decoding procedure to assign observations (participants' returns on trials) to latent states. The states are ordered by expected return, with state 1 having the lowest mean return and state 5 the highest. Figure 6 shows that participants were more likely to be in a lower return state in the control condition compared to the intervention condition both pre and post defection. For instance, in round 5, state 1 was the most likely posterior state for only 5% of participants in the intervention condition compared to 12% in the control condition ($\chi^2(1) = 4.73, p = 0.03$). For the post-defection trial after the intervention (round 14), state 1 was the most likely state for only 15% of participants in the intervention condition compared to 31% in the control condition ($\chi^2(1) = 14.70, p < 0.001$). Whilst the posterior states indicate that the intervention was effective, a non-negligible proportion of participants in the intervention condition did not exhibit the coaxing behavior promoted by the intervention. Directly following the low investment in round 13, 17% of participants in the intervention condition were assigned to state 1 with the lowest average returns highlighting individual differences in the effectiveness of the intervention.

5 Discussion

In this experiment, human participants took the role of the trustee in a Repeated trust game (RTG) where they faced artificial computer agents who's behavior was partly determined by participants' returns. The behavior of the artificial investors was determined by a 3-state hidden Markov model (HMM), which was estimated from the behavior of humans in the RTG in prior research. Overall, investments and returns by the artificial and human agents replicated that of human dyads in prior research using the RTG (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011). This, together with participants' reported uncertainty about whether they faced a human or artificial investor, shows the potential of using HMM-based artificial agents to mimic human behavior in economic games, whilst offering a higher degree of experimental control than human dyadic interaction.

The aim of the cognitive intervention was to articulate the potential unwanted effects of acting on impulse after a transgressive action from the investor in the form of a one-off low investment. After the intervention, participants sent back higher returns compared to before the intervention, and did not decrease their returns after a transgressive action like participants in the control condition did. The overall higher returns after the intervention occurred despite participants' emotional reactions to the investments remaining largely the same as before the intervention. This indicates the intervention achieved its goal of encouraging participants to respond in a non-impulsive and considered manner, possibly overriding the urge to retaliate.

That participants generally send higher returns to the investor after the intervention is unlikely due to a general learning effect unrelated to the intervention, as participants in the control condition did not increase their returns. Also, as participants in the intervention and control condition faced the same HMM investor, the higher post-intervention returns are not solely due to a difference in investor behavior. As the investor reacts to participants'

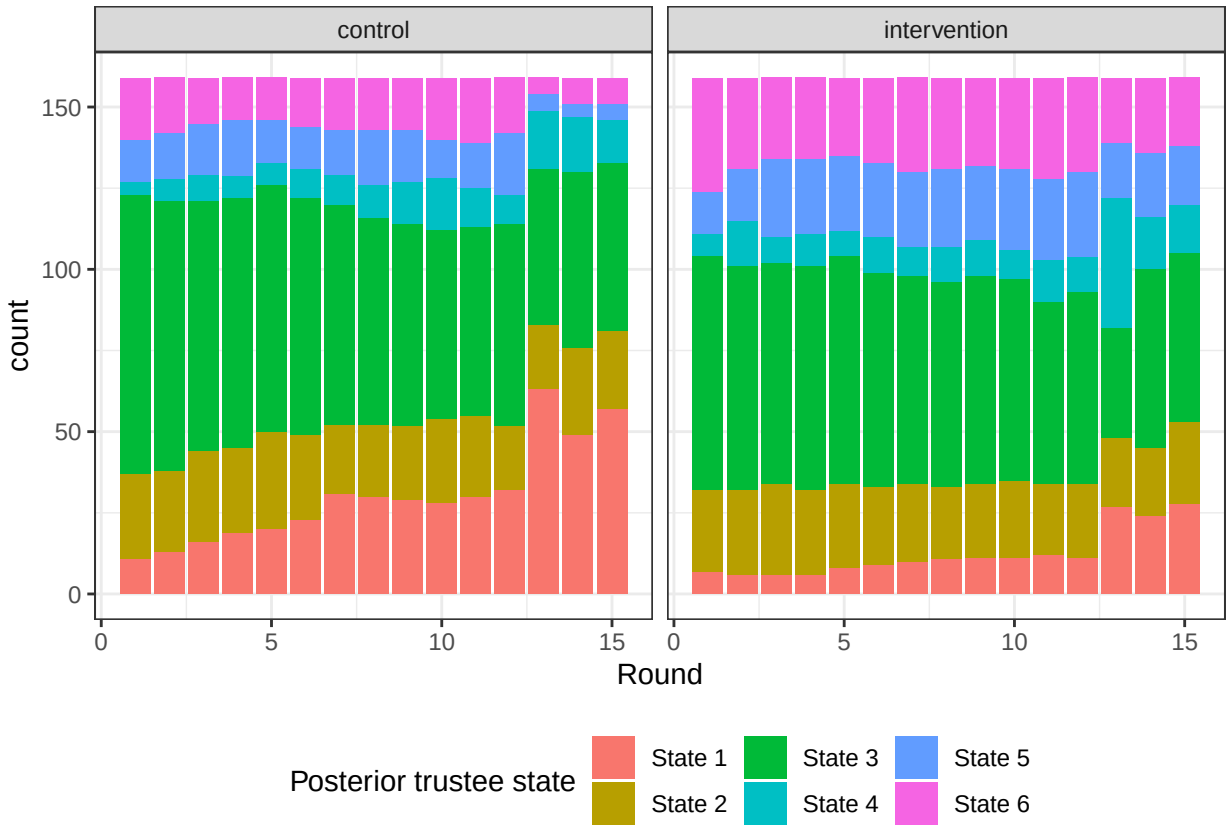

Figure 6: Distribution of posterior trustee states post-manipulation by condition for all rounds, as estimated by the most likely posterior state in the best fitting HMM model (HMM-full) using a local decoding procedure. States are represented in increasing degrees of average percentage returns from the lowest (state 1) to the highest (state 6) return state. This figure characterises behavioral differences between conditions as the result of participants being in more pro-social (higher return) states in the intervention condition compared to the control condition both pre- and post-defection.

returns, those who return more will generally see higher investments. But this is driven by the magnitude of their returns, not by a change in the strategy of the investor. Finally, as there were no differences between conditions in how participants rated the first and second HMM agent on attributes such as cooperativeness and trust, the increased returns are unlikely to be the result of a more favourable evaluation of the investor. Rather, we find it most likely that the increased returns are due to participants inferring from the intervention that pro-social and trustworthy behavior may generally motivate the investor to send high investments, which can provide more beneficial outcomes to them in the long run.

The effect of the intervention was not transferred to the Repeated Prisoner's Dilemma (RPD) game. There was no difference between the intervention and control condition in the rate of cooperation, whether before or after a preprogrammed defection by the artificial agent. As the prisoner's dilemma is a popular economic game, it is possible that participants had a strong prior commitment towards the strategy they would adopt, which was not overridden by the intervention. The RPD also involves much coarser actions (cooperate vs defect) than the finer-grained returns in the RTG. This makes it more difficult to observe more subtle effects of the intervention. As such, the RPD might not be the best choice of task to measure transfer. In any case, we can not rule out that the effects of the brief cognitive intervention, which explicitly focused on the RTG, are confined to the RTG. Future research will need to assess whether transfer to other games is possible.

Analysing participants' behavior with hidden Markov models, we found clear individual differences in how returns changed between the pre- and post-intervention RTG, which can be seen as a proxy for the effectiveness of the intervention. Some participants may not have been convinced that coaxing via high returns was a good way to establish cooperation and decided to reduce their returns in the second trust game. Their impulse to "punish" the other player for a defection may have been too strong to be overridden by the intervention. This was also evident from participants' replies to a question about whether they would change their behavior, just after receiving the intervention. Other participants responded to the intervention and increased their returns. Heterogeneity in response to treatment is common in psychiatry and related fields. Such heterogeneity may reflect the complex nature of mental health problems, which may be best viewed as complex systems involving interactions between neuro-computational processes and socio-environmental contexts evolving over time (Fried and Cramer 2017). This view was used to justify computational psychiatry's difficulty in establishing differential and reliable predictors of likely treatment response (Hitchcock, Fried, and Frank 2022). Here, we found heterogeneity in reaction to a relatively explicit intervention by a sample of participants from the general population. This suggests that the issue of variable treatment responses may result from the interaction of two sources of variability: the phenotyping of the disorder as well as the phenomenological aspects of the intervention itself. As such, a rigorous exploration of the determinants of inter-individual differences to an intervention in the general patient population is required.

Overall, we are encouraged that our brief cognitive intervention, consisting of reading a short text detailing a non-impulsive reaction to low investments, can lead to clearly differentiated behavior. In future studies, we aim to explore the effects of improved cognitive interventions to enhance cooperative behavior. We may enhance engagement by making the intervention more interactive and visually appealing, rather than the "dry" textual format used in the experiment. It would also be of interest to test such interventions with participants that suffer from an inability or unwillingness to repair relationships after an accidental breakdown of trust, such as people with Borderline Personality Disorder. The relative ease by which online interventions can be assigned, and the opportunity for people to test the effect of their behavior with artificial but human-like agents, may pave the way for efficient, low-cost, effective treatment programmes which may help a wide-variety of people overcome detrimental actions in social situations.

References

Allen, Jon G., and Peter Fonagy, eds. 2006. *The Handbook of Mentalization-Based Treatment.* The Handbook of Mentalization-Based Treatment. Hoboken, NJ, US: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470712986.

Almaatouq, Abdullah, Joshua Becker, James P Houghton, Nicolas Paton, Duncan J Watts, and Mark E Whiting. 2021. "Empirica: A Virtual Lab for High-Throughput Macro-Level Experiments." *Behavior Research Methods* 53 (5): 2158–71.

Arch, Joanna J., Kate B. Wolitzky-Taylor, Georg H. Eifert, and Michelle G. Craske. 2012. "Longitudinal Treatment Mediation of Traditional Cognitive Behavioral Therapy and Acceptance and Commitment Therapy for Anxiety Disorders." *Behaviour Research and Therapy* 50 (7-8): 469–78. https://doi.org/10.1016/j.brat.2012.04.007.

Axelrod, Robert, and William D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211 (4489): 1390–96. https://doi.org/10.1126/science.7466396.

Barnby, Joseph M, Peter Dayan, and Vaughan Bell. 2023. "Formalising Social Representation to Explain Psychiatric Symptoms." *Trends in Cognitive Sciences*.

Bendor, Jonathan, Roderick M. Kramer, and Suzanne Stout. 1991. "When in Doubt...: Cooperation in a Noisy Prisoner's Dilemma." *Journal of Conflict Resolution* 35 (4): 691–719. https://doi.org/10.1177/0022002791035004007.

Burnham, Terence, Kevin McCabe, and Vernon L Smith. 2000. "Friend-or-Foe Intentionality Priming in an Extensive Form Trust Game." *Journal of Economic Behavior & Organization* 43 (1): 57–73. https://doi.org/10.1016/S0167-2681(00)00108-6.

Charness, Gary, Ramón Cobo-Reyes, and Natalia Jiménez. 2008. "An Investment Game with Third-Party Intervention." *Journal of Economic Behavior & Organization* 68 (1): 18–28. https://doi.org/10.1016/j.jebo.2008.02.006.

Dercon, Quentin, Sara Z Mehrhof, Timothy R Sandhu, Caitlin Hitchcock, Rebecca P Lawson, Diego A Pizzagalli, Tim Dalgleish, and Camilla L Nord. 2022. "A Core Component of Psychological Therapy Causes Adaptive Changes in Computational Learning Mechanisms." *Psychological Medicine*, 1–11.

Drążkowski, Dariusz, Lukasz D. Kaczmarek, and Todd B. Kashdan. 2017. "Gratitude Pays: A Weekly Gratitude Intervention Influences Monetary Decisions, Physiological Responses, and Emotional Experiences During a Trust-Related Social Interaction." *Personality and Individual Differences* 110 (May): 148–53. https://doi.org/10.1016/j.paid.2017.01.043.

FeldmanHall, Oriel, and Amitai Shenhav. 2019. "Resolving Uncertainty in a Social World." *Nature Human Behaviour* 3 (5): 426–35.

Fiedler, Marina, and Ernan Haruvy. 2017. "The Effect of Third Party Intervention in the Trust Game." *Journal of Behavioral and Experimental Economics* 67 (April): 65–74. https://doi.org/10.1016/j.socec.2016.10.003.

Fiedler, Marina, Ernan Haruvy, and Sherry Xin Li. 2011. "Social Distance in a Virtual World Experiment." *Games and Economic Behavior* 72 (2): 400–426. https://doi.org/10.1016/j.geb.2010.09.004.

Fonagy, Peter, and Elizabeth Allison. 2014. "The Role of Mentalizing and Epistemic Trust in the Therapeutic Relationship." *Psychotherapy* 51: 372–80. https://doi.org/10.1037/a0036505.

Fonagy, Peter, and Chloe Campbell. 2017. "Mentalizing, Attachment and Epistemic Trust: How Psychotherapy Can Promote Resilience." *Psychiatria Hungarica: A Magyar Pszichiatriai Tarsasag Tudomanyos Folyoirata* 32 (3): 283–87.

Fried, Eiko I., and Angélique O. J. Cramer. 2017. "Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology." *Perspectives on Psychological Science* 12 (6): 999–1020. https://doi.org/10.1177/1745691617705892.

Giordano, Giuseppe Nicola, and Martin Lindström. 2016. "Trust and Health: Testing the Reverse Causality Hypothesis." *Journal of Epidemiology and Community Health* 70 (1): 10–16. https://doi.org/10.1136/jech-2015-205822.

Gunderson, John G., Sabine C. Herpertz, Andrew E. Skodol, Svenn Torgersen, and Mary C. Zanarini. 2018. "Borderline Personality Disorder." *Nature Reviews Disease Primers* 4 (1): 18029. https://doi.org/10.1038/nrdp.2018.29.

Hitchcock, Peter F., Eiko I. Fried, and Michael J. Frank. 2022. "Computational Psychiatry Needs Time and Context." *Annual Review of Psychology* 73 (1): 243–70. https://doi.org/10.1146/annurev-psych-021621-124910.

Huys, Quentin J M, Tiago V Maia, and Michael J Frank. 2016. "Computational Psychiatry as a Bridge from Neuroscience to Clinical Applications." *Nature Neuroscience* 19 (3): 404–13. https://doi.org/10.1038/nn.4238.

Joyce, Berg, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1): 122–42.

King-Casas, Brooks, Carla Sharp, Laura Lomax-Bream, Terry Lohrenz, Peter Fonagy, and P. Read Montague. 2008. "The Rupture and Repair of Cooperation in Borderline Personality Disorder." *Science* 321 (5890): 806–10. https://doi.org/10.1126/science.1156902.

Lieb, Klaus, Mary C Zanarini, Christian Schmahl, Marsha M Linehan, and Martin Bohus. 2004. "Borderline Personality Disorder." *The Lancet* 364 (9432): 453–61. https://doi.org/10.1016/S0140-6736(04)16770-6.

Linehan, Marsha M. 1993. *Cognitive-Behavioral Treatment of Borderline Personality Disorder.* Cognitive-Behavioral

Treatment of Borderline Personality Disorder. New York, NY, US: Guilford Press.
———. 2015. *DBT® Skills Training Manual, 2nd Ed.* DBT® Skills Training Manual, 2nd Ed. New York, NY, US: Guilford Press.
Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. "Balancing Type i Error and Power in Linear Mixed Models." *Journal of Memory and Language* 94: 305–15.
Meng, Tianguang, and He Chen. 2014. "A Multilevel Analysis of Social Capital and Self-Rated Health: Evidence from China." *Health & Place* 27 (May): 38–44. https://doi.org/10.1016/j.healthplace.2014.01.009.
R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Reiter, Andrea MF, Nadim AA Atiya, Isabel M Berwian, and Quentin JM Huys. 2021. "Neuro-Cognitive Processes as Mediators of Psychological Treatment Effects." *Current Opinion in Behavioral Sciences*, Computational cognitive neuroscience, 38 (April): 103–9. https://doi.org/10.1016/j.cobeha.2021.02.007.
Rudge, Susie, Janet Denise Feigenbaum, and Peter Fonagy. 2020. "Mechanisms of Change in Dialectical Behaviour Therapy and Cognitive Behaviour Therapy for Borderline Personality Disorder: A Critical Review of the Literature." *Journal of Mental Health* 29 (1): 92–102. https://doi.org/10.1080/09638237.2017.1322185.
Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2). https://doi.org/10.1214/aos/1176344136.
Singmann, Henrik, Ben Bolker, Jake Westfall, Frederik Aust, Mattan S. Ben-Shachar, Søren Højsgaard, John Fox, et al. 2022. "Afex: Analysis of Factorial Experiments."
Visser, Ingmar, and Maarten Speekenbrink. 2021. "depmixS4: Dependent Mixture Models - Hidden Markov Models of GLMs and Other Distributions in S4."
———. 2022. "Hidden Markov Models." In *Mixture and Hidden Markov Models with r*, 125–72. Springer.