(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

Restoring cooperation after trust breaches remains a major hurdle in social, clinical, and organizational contexts. Yet, most interventions focus on initial trust formation rather than repair. We introduce a novel, online cognitive intervention—rooted in Dialectical Behavior Therapy—to reduce psychological reactivity and promote resilient cooperation. In a randomized controlled online trial (N = 318), participants played a Repeated Trust Game against an HMM-based adaptive agent trained on human data, ensuring dynamic, human-like responses under full experimental control. Compared to an active control, the cognitive intervention prevented retaliation following programmed trust violations and yielded significantly higher cooperative returns. Mixed-effects analysis revealed that while control participants became increasingly reciprocal over time (approaching tit-for-tat strategies), intervention participants maintained more stable cooperative behavior that was less contingent on partner investment levels. This change did not generalize to Prisoner's Dilemma, underscoring the intervention's specificity. These results validate HMM agents as a powerful experimental tool and suggest that targeted cognitive strategies may help support cooperative behavior; clinical efficacy and real-world deployment remain to be established.

[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.*

[*] *Current email: ismail.guennouni@iwr.uni-heidelberg.de*

Author Note

Author Contributions: I. Guennouni, QJM. Huys and M. Speekenbrink designed and developed the study concept. Experiment design was done by S. Dupret and I. Guennouni. Testing and data collection were performed by I. Guennouni. I. Guennouni analysed and interpreted the data under the supervision of QJM. Huys and M. Speekenbrink. M. Speekenbrink and I. Guennouni jointly performed the HMM modelling. All authors jointly wrote and approved the final version of the manuscript for submission.

Funding: I. Guennouni was supported by the UK Engineering and Physical Sciences Research Council under grant EP/S515255/1. QJM. Huys acknowledges support by the UCLH NIHR BRC. He has received fees and options for consultancies for Aya Technologies and Alto Neuroscience.

Data Availability: Data, analysis code, and study materials are openly available at https://github.com/ismailg/CoaxIntervention/tree/submission-decision.

Correspondence: Correspondence concerning this article should be addressed to Ismail Guennouni, Current address: Interdisciplinary Center for Scientific Computing, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany. Email: ismail.guennouni@iwr.uni-heidelberg.de

Introduction

Cooperation, defined as individuals or entities working together towards a shared goal, is fundamental to collective success and social harmony (Tomasello et al., 2012). At the heart of cooperation is trust—the belief that others will act in ways that are mutually beneficial, even when they have the opportunity to exploit the situation (Rousseau et al., 1998). Trust enables individuals to engage in risky interactions where immediate self-interest could easily override long-term benefits (Balliet & Van Lange, 2013). Without trust, cooperation tends to break down, leading to suboptimal outcomes for all parties involved. Understanding how to maintain and repair cooperation following such breakdowns is therefore of significant interest to researchers and practitioners alike.

The Repeated Trust Game (RTG) has emerged as a well-established paradigm for studying trust and cooperation in controlled settings (Joyce et al., 1995). In this game, an "investor" decides how much of an endowment to send to a "trustee." The amount sent is typically multiplied by 3, and the trustee then decides how much of this multiplied amount to return to the investor and how much to keep for themselves. Cooperation emerges when both parties act in ways that promote mutual gains. However, trust is fragile, and a single instance of defection—where one player fails to reciprocate appropriately—can lead to a breakdown of cooperation (Bendor et al., 1991). Once trust is violated, it is often difficult to re-establish, even if doing so would be mutually beneficial (Harth & Regner, 2017).

Previous research has explored ways to encourage initial cooperation in trust games, such as using third-party enforcement (Charness et al., 2008; Fiedler & Haruvy, 2017) and priming with concepts of gratitude (Drążkowski et al., 2017) or friend and foe (Burnham et al., 2000). While these approaches increase early cooperation, they often fail to address the more challenging task of repairing cooperation after trust has been broken. After an act of defection from another party, individuals may react impulsively by reducing their own cooperative efforts, even though re-establishing trust could be more beneficial to them as well as the other party. Interventions aimed at restoring cooperation in these situations are therefore crucial, yet understudied.

Here, we focus on the role of the trustee in the RTG in establishing and maintaining cooperation. While the trustee does not exercise trust in the same way as the investor, their decisions whether to reciprocate or not play a critical role in maintaining or disrupting a cooperative relationship. A lack of reciprocation from trustees erodes trust over time (Servátka et al., 2011), which is particularly evident when the trustee suffers from personality disorders such as borderline personality disorder (Lieb et al., 2004). These trustees fail to engage in trust-repairing behaviors such as coaxing the investor by signaling trustworthiness via sending high returns (King-Casas et al., 2008). Additionally, unpredictable behavior from trustees fosters mistrust and impedes future cooperation (Rigdon et al., 2007). In contrast, consistent cooperation from trustees promotes trust and encourages further collaboration, as evidenced by neural data (King-Casas et al., 2005). Therefore, emphasizing the role of trustees in rebuilding cooperation is essential; when trustees demonstrate reliability and reciprocity, even after breaches of trust, cooperation can be restored and sustained.

Given the pivotal role of trustees' behavior in shaping cooperative dynamics, it is worth exploring whether interventions aimed at improving interpersonal skills could positively influence their decision-making in the RTG. In the broader field of psychological therapies, cognitive interventions inspired by Dialectical Behavior Therapy [DBT; Linehan (1993)] and Mentalisation Based Therapy [MBT; Allen & Fonagy (2006)] have shown promise in enhancing social skills among individuals with interpersonal difficulties. These approaches focus on helping individuals recognize the impact of their actions on others and considering alternative strategies. Drawing inspiration from such therapeutic approaches, we employ a randomized control trial to evaluate a cognitive intervention aimed at repairing cooperation after low investments from a computerized investor. The intervention in this study is a brief, multi-component cognitive intervention inspired by DBT principles. It combines elements aimed at understanding long-term consequences of actions and promoting prosocial behavior. This approach mirrors real-world cognitive interventions that often employ multiple strategies to effect behavior change (Linehan, 2015). Specifically, we hypothesize that encouraging participants to reflect on the consequences of their actions and to consider a non-impulsive course of action might lead to more resilient cooperative behavior, even in the face of perceived non-cooperation from their partner.

We conducted an online experiment with 318 participants acting as trustees in two RTG rounds, with the intervention administered between rounds to a subset of participants. While previous studies have often relied on predetermined or simplistic computer strategies in economic games, our study introduces a novel approach using Hidden Markov Models (HMMs) to create more realistic, adaptive computer agents. A key aspect of these agents is that their actions depend on a latent "trust state" which reacts dynamically to the trustees' returns, simulating real-life trust-building scenarios. 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.

Methods

Participants and design

The experiment employed a 2 (Condition: Intervention or Control) by 2 (Game: Trust-Game Pre-Intervention, Trust-Game Post-Intervention) design, with repeated measures on the second factor (see Figure 1.A for a graphical overview of the experiment structure). A total of 320 participants were recruited on the Prolific Academic platform (prolific.co). The required sample size was determined via a priori power analysis with G*Power (Faul et al., 2009) to ensure power of .8 to detect a small within-between interaction effect (Cohen's f = 0.10) in a 2 by 2 mixed ANOVA with a .05 significance level. Participants were randomly assigned to either the intervention or control condition. Two players had incomplete trust game entries and their data was disregarded, leaving data from 318 participants for analysis, equally split between the two conditions. The mean age of participants was 31.3 years (SD = 9.9). Participants were paid a fixed fee of £5 plus a performance-contingent bonus payment of £0.71 on average. All participants provided informed consent and the study received ethical approval from the local UCL ethics board (ID:21029/001).

Tasks and Measures

Repeated Trust Game

Participants played two separate 15-round Repeated Trust Games (Joyce et al., 1995) in the role of trustee. In each round of the RTG, the (computer-simulated) investor is endowed with 20 units and decides how much of that endowment to invest. This investment is tripled, and the trustee (participant) then decides how to split this tripled amount between themselves and the investor. If the trustee returns more than one third of the amount they receive, the investor makes a gain.

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 indicating what they were asked to report (see supplement for a screen shot of the grid).

The strategy of the computerized investor was modelled on behavior of human investors in a 10-round Repeated Trust Game (RTG) with the same co-player. Details of the data sources used for this are provided in the Supplementary Information. Using this data, we estimated a hidden Markov model (HMM) for 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 1.C). These distributions reflect "low-trust", "medium-trust", or "high-trust". Over rounds, the investor can switch state, and the probability of such state transitions are a function of the net return (i.e. return - investment) in the previous round (see Figure 1.D). In order to instigate a potential breakdown of trust allowing us to probe efforts to repair trust, the computerized 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 initial state for the HMM investor in each instance of the game was the "mid-trust" state.

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 maximize 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. The full text of the education slide of the intervention is provided in Figure 1.B

In order to distinguish general practice effects from the effect of the intervention, we included a control condition in which 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.

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 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 maximize their reward by choosing to cooperate.

In this game, the computerized agent was programmed to act according to a tit-for-tat strategy (Axelrod & Hamilton, 1981), starting with a cooperative action and then mirroring what the other player chose in the preceding round. On round 4, the computerized agent was pre-programmed to choose the defect action, regardless of the participant's preceding action.

Post game questionnaires

Failure to repair a breakdown in trust in the repeated trust game has been associated with trustees with BPD traits (King-Casas et al., 2008). Theories of social dysfunction in BPD have focused on dysfunction in the patients' mentalizing ability (Allen & Fonagy, 2006) as well as difficulties in emotional regulation (Rudge et al., 2020). The questionnaires we included in the experiment tried to assess borderline traits (PAI-BOR; Morey (1991)), emotional regulation capabilities (DERS; Gratz & Roemer (2004)) and mentalizing ability (RFQ8; Fonagy et al. (2016)).

Procedure

At the start of the experiment, participants provided informed consent and were instructed the study would consist of three phases. Participants were led to believe they were interacting with human co-players. They were told they would face the same co-player within all rounds of a game, but that the co-player would change between games (i.e. from the first to the second Repeated Trust Game, and then to the Repeated Prisoner's Dilemma). Participants in both conditions were told that their goal was to maximize the number of points in all phases. Participants had to pass comprehension checks about the number of phases, the fact the co-player was the same within each phase, and that they would face a new player at the start of each new phase. They were not told the number of rounds of each phase.

Phase one was a 15-round RTG in which participants took the role of trustee, facing the same investor over all 15 rounds. Participants were given detailed instructions about the game and had to pass comprehension checks to test their understanding of their and their co-player's payoff in a hypothetical situation. On each round, after being informed about the amount sent by the investor, participants were asked to evaluate their emotion (intervention 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 co-player.

Phase three consisted of 7 rounds of the Repeated Prisoner's Dilemma game (RPD), with participants informed they would face a third co-player. Participants then completed questionnaires related to mentalizing 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. Finally, participants were debriefed and thanked for their participation.

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:

$$
\begin{aligned}
R_{ij} = &\beta_0 + \beta_1 \text{ (Condition)}_i + \beta_2 \text{ (Game)}_i + \beta_3 \text{ (Investment)}_i \\
&\beta_4 \text{ (Condition} \times \text{Game)}_i + \beta_5 \text{ (Condition} \times \text{Investment)}_i + \beta_6 \text{ (Game} \times \text{Investment)}_i + \\
&\beta_7 \text{ (Condition} \times \text{Game} \times \text{Investment)}_i + \beta_8 \text{ (RoundNumber)}_i + \beta_9 \text{ (IsDefectionRound)}_i + \\
&b_{0j} + b_{1j} \text{ (Game)}_i + b_{2j} \text{ (Investment)}_i + \epsilon_{ij}
\end{aligned}
$$

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
- β_8: effect of round number
- β_9: effect of defection round
- 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

Our choice of linear mixed-effects models (LMMs) over mixed ANOVA was based on their greater flexibility in handling our complex data structure, including continuous predictors and nested repeated measures. LMMs offer increased statistical power and more flexible assumptions, particularly regarding sphericity, which is often violated in repeated measures designs. The model was estimated using the `afex` package (Singmann et al., 2022) in R. 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. As there is no agreed upon way to calculate effect sizes for mixed effects models, we will report instead on testing differences in marginal means. For the F-tests, we used the Kenward-Roger approximation to the degrees of freedom, as implemented in the `afex` package. For all post-hoc pairwise comparisons following significant effects in the mixed-effects models, we used Tukey's Honestly Significant Difference (HSD) test to adjust for multiple comparisons, unless otherwise stated. We Z-transform the Investment variable as centering is beneficial to interpreting the main effects more easily in the presence of interactions.

For emotion self-reports, we analyzed valence and arousal using linear mixed-effects models with fixed effects for Game (pre vs. post), Investment (z-scored), and their interaction, and participant-wise random intercepts and random slopes for Game.

To examine the temporal dynamics of responses to trust violations with finer resolution, we conducted an event-study analysis around the pre-programmed low investment rounds (round 12 in the pre-phase, round 13 in the post-phase). We modeled percentage returns using relative time as a factor (t-2, t-1, t=defection, t+1, t+2), fully interacted with Condition and Game, while controlling for investment level. This approach separates immediate responses at the defection trial from anticipatory effects before defection and sustained behavioral changes after defection.

To model participants' returns in the RTG across games and conditions, we fit various hidden Markov models (Visser & Speekenbrink, 2022) to participants' returns using the `depmixS4` package (Visser & 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 fitting model.

Data and materials availability

All data, analysis code, and study materials are openly available at https://github.com/ismailg/CoaxIntervention/tree/submission-decision.

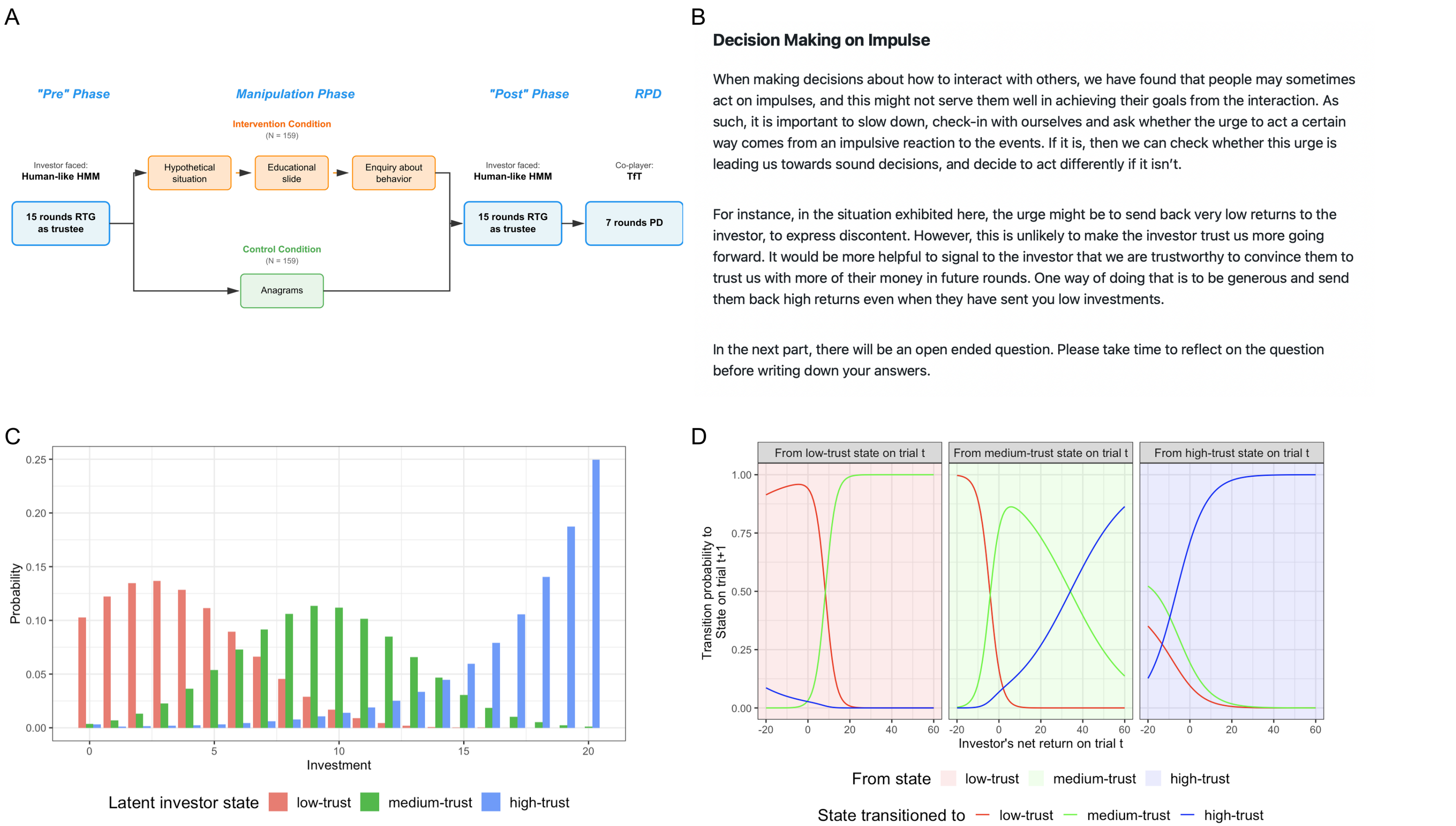

Figure 1: Panel A: Experimental design overview. Participants in both conditions played two 15-round Repeated Trust Games (RTGs) with human-like Hidden Markov Model (HMM) investors, separated by a manipulation phase. The intervention condition received a cognitive intervention, while the control condition solved anagrams. Both groups then completed a 7-round Repeated Prisoner's Dilemma (RPD) with a tit-for-tat (TFT) co-player to assess transfer effects. This design allows for comparison of cooperative behavior before and after the intervention, as well as between conditions. Panel B: Full text of the educational slide shown to the intervention group. Panels C - D: 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 C. 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 D. Each plot in Panel D 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.

Behavioral results

Average investments and returns prior to the pre-programmed low investment round were within the range reported in previous studies (investments: 40-60% of endowment; returns: 35-50% of total yield; Cochard et al. (2004); Charness et al. (2008); Figure 2.A).

The intervention increased cooperative behavior. Percentage returns were higher overall in the intervention condition compared to control (main effect of Condition: $F(1, 316.08) = 8.41$, $p = .004$). Critically, this effect varied across game phases (Condition × Game interaction: $F(1, 316.08) = 23.40$, $p < .001$).

Post-hoc contrasts revealed divergent trajectories between conditions. The intervention group increased cooperation from pre to post ($\Delta M = 0.03$, 95% CI $[0.02, 0.05]$, $t(313.85) = 4.64$, $p < .001$), while the control group decreased cooperation ($\Delta M = -0.01$, 95% CI $[-0.03, 0.00]$, $t(319.32) = -2.07$, $p = .039$). Groups did not differ at baseline ($t(316.29) = 1.02$, $p = .307$), but the intervention group showed significantly higher returns post-manipulation ($t(316.05) = 4.26$, $p < .001$; Figure 2.C).

Several additional factors influenced cooperative behavior. First, participants showed conditional cooperation: higher investments elicited higher percentage returns (main effect of Investment: $F(1, 337.35) = 41.05$, $p < .001$). Second, cooperation eroded over repeated rounds (main effect of Round: $F(1, 8632.78) = 106.01$, $p < .001$). Third, during pre-programmed rounds where the computer investor sent a low investment (round 12 in the pre phase and round 13 in the post phase), participants returned less compared to other rounds ($F(1, 8628.87) = 21.82$, $p < .001$; $\Delta M = 0.03$, 95% CI $[0.02, 0.04]$, $z = 4.67$, $p < .001$).

The intervention and control groups differed in how strongly they reciprocated partner investments. The control group showed steeper reciprocity slopes—larger increases in returns for higher investments—compared to the intervention group (Investment × Condition interaction: $F(1, 315.26) = 8.27$, $p = .004$; $\Delta M = -0.02$, 95% CI $[-0.04, -0.01]$, $z = -2.88$, $p = .004$).

The intervention fundamentally altered how participants responded to partner behavior, as revealed by a three-way Game × Condition × Investment interaction ($F(1, 8855.73) = 24.89$, $p < .001$). At baseline, both groups showed similar reciprocity patterns ($\Delta M = -0.01$, 95% CI $[-0.02, 0.01]$, $z = -0.90$, $p = .366$). Post-intervention, the groups diverged: the intervention group became less contingent on partner investments (i.e., maintained higher returns regardless of investment level), whereas the control group became more contingent (i.e., returned more only when the partner invested more; $\Delta M = -0.04$, 95% CI $[-0.06, -0.02]$, $z = -4.47$, $p < .001$). This shift from conditional to more unconditional cooperation in the intervention group was statistically significant ($\Delta M = 0.03$, 95% CI $[0.02, 0.04]$, $z = 4.99$, $p < .001$).

HMM investments were higher in the intervention than control, and higher in the second game than the first (main effects: Condition $F(1, 316) = 8.88$, $p = .003$, Game $F(1, 316) = 7.80$, $p = .006$). See Figure 2.D.

We analyzed participants' responses to trust violations using aggregate periods before versus after the pre-programmed low investment. Before defection (rounds 1-11 pre-phase, 1-12 post-phase), returns increased in the intervention group ($\Delta M = 0.03$, 95% CI $[0.02, 0.05]$, $z = 4.64$, $p < .001$) while decreasing in the control group ($\Delta M = -0.01$, 95% CI $[-0.03, 0.00]$, $z = -2.07$, $p = .038$). After defection (rounds 12-15 pre-phase, 13-15 post-phase), control participants decreased returns from first to second RTG ($\Delta M = -0.06$, 95% CI $[-0.09, -0.04]$, $t(337.93) = -4.49$, $p < .001$), whereas the intervention group showed no significant change.

An event-study analysis pinpointed the exact timing of these effects. At baseline (pre-phase), groups showed no differences at any time point relative to defection (all p > .05). Post-intervention, the intervention group maintained significantly higher returns than controls at the exact moment of defection (t=0: $\Delta M = 0.13$, $z = 5.21$, $p < .001$), with this advantage persisting at t+1 ($\Delta M = 0.05$, $p = 0.036$) and t+2 ($\Delta M = 0.06$, $p = 0.008$). Within-group comparisons revealed that intervention participants increased cooperation from pre to post specifically during defection ($\Delta M = 0.09$, $p < .001$), while controls decreased cooperation ($\Delta M =$ -0.04, $p = 0.015$) and continued retaliating thereafter. This demonstrates the intervention specifically enhanced resilience at the critical moment of trust violation.

We also examined whether participants' questionnaire scores were associated with their behavior or interacted with the experimental conditions. Linear mixed-effects models including these scores as covariates revealed no significant associations or interactions with other variables such as condition and game, suggesting that the observed effects were not moderated by the individual differences measured by our questionnaires.

Emotion self-reports

To assess the impact of the intervention on participants' emotional reactions, we used linear mixed-effects models for valence and for arousal, with fixed effects for Game (pre vs. post intervention) and Investment, as well as interaction between Investment and Game, with participant-wise 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, $F(1, 3409.69) = 0.12$, $p = .726$. Despite similar investment-related emotional responses, participants in the intervention condition returned higher amounts post-intervention (Figure 2.B).

Evaluation of the investor

To analyse participants' evaluations of the investor, 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 computerized investor in the second game as less cooperative ($\Delta M = -0.42$, 95% CI $[-0.69, -0.15]$, $t(317) = -3.10$, $p = .002$), less trustworthy ($\Delta M = -0.43$, 95% CI $[-0.70, -0.16]$, $t(317) = -3.19$, $p = .002$), less friendly ($\Delta M = -0.40$, 95% CI $[-0.64, -0.17]$, $t(317) = -3.36$, $p < .001$) and more selfish ($\Delta M = 0.36$, 95% CI $[0.10, 0.61]$, $t(317) = 2.76$, $p = .006$), than the computerized 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 co-player. Many answers reflected participants projecting human traits such as "spitefulness" or "greed" onto the artificial co-player's behavior.

Transfer to the Repeated Prisoner's Dilemma game

We next asked whether the intervention had any discernible effect on participants' behavior in a different game (the Repeated Prisoner's Dilemma). Predicting the probability of a cooperative action with a mixed-effects logistic 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.

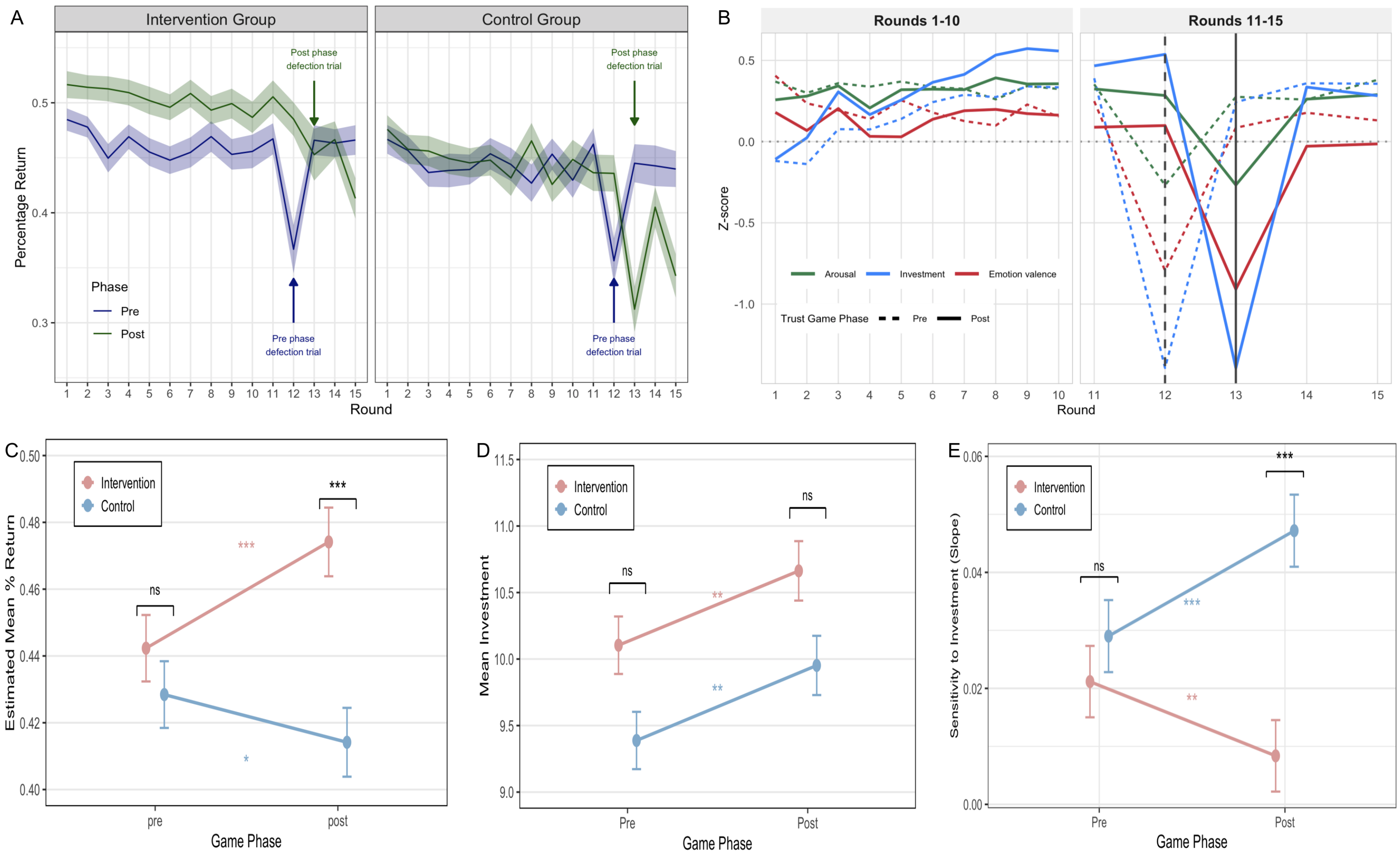

Figure 2: Panel A: Averages and standard errors of the trustee's return as a percentage of the multiplied investment received by Condition, Phase, and game round. We note a different reaction to the pre-programmed one-off low investment between the two conditions: While there is a dip in returns in the Pre phase for both conditions, we see higher returns in the intervention condition compared to the dip in returns seen in the control condition during the Post phase. Panel B: 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' emotional reaction measured during the investor's pre-programmed one-off low investment round (vertical grey lines) was similar before (round 12) and after (round 13) the intervention. The bottom panels show estimated marginal means of percentage returns (Panel C), investments (Panel D), and emotion trend interactions (Panel E) across the Pre and Post phases for both the Intervention and Control conditions. Error bars represent the standard error of the means. Panel C shows that participants in the Intervention condition returned higher proportions in the second game compared to the first game over all rounds, whilst those in the Control condition sent back lower returns. Investments by the HMM agent (Panel D) were higher in the second game compared to the first game across conditions. Those in the Intervention group became less sensitive to investments received in the post phase, whilst those in the control group became more sensitive.

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 models 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 ($\chi^2(60) = 110.01$, $p < .001$). As such, there is evidence that participants reacted differently to the investments in the four combinations of Condition and Game. Consistent with the mixed-effects trends, the intervention was associated with reduced sensitivity to investment levels and a shift in posterior state occupancy toward higher-return states (see Figure 2.E and Figure 3). The estimated state-dependent policy of trustee actions, according to the HMM-full model, is depicted in Figure 3.A. In HMM-full, pre (Game 1) is shared across conditions while post (Game 2) is allowed to differ by condition via a three-level contrast.

Taking the best-fitting 6-state HMM-full model, we used a local decoding procedure to determine participants' state on each round of the RTG. The states are ordered by expected return, with state 1 having the lowest mean return and state 6 the highest. Figure 3.B 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 state for only 5% of participants in the intervention condition compared to 12% in the control condition ($\chi^2(1) = 4.73, p = .03$). 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 ($\chi^2(1) = 14.70, p < .001$). While 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.

Discussion

Following a cognitive intervention, participants increased their returns without a corresponding change in emotional response, indicating the intervention's effectiveness in preserving cooperation and reducing retaliation to a breach of cooperation. Those in the control condition returned similar proportions pre-defection but reduced their returns post-defection. The intervention produced an increase in trustee reciprocity, with the proportion of the endowment returned rising by approximately 6 percentage points post-intervention relative to the control group (from 41.4% to 47.4%, or a 14.5% relative increase). For context, baseline reciprocity in Trust Games typically falls between 35–50 percent of trustee receipts (Johnson & Mislin, 2011). This effect is larger than the 4 percentage point gains observed from interventions that allowed limited communication between players in the trust game (Ben-Ner et al., 2011), but more modest than the 11 points reported by Charness & Dufwenberg (2006) when full communication between players is allowed. Beyond the magnitude effects, an HMM analysis showed that those in the intervention condition were less prone to low-return states after defection. This aligns with the intervention's target: participants learned to avoid getting stuck in the low-cooperation states that typically follow trust violations, demonstrating successful acquisition of the intervention's core behavioral strategy.

The increased returns in the intervention condition are unlikely due to a general learning effect, as participants in the control condition did not increase their returns. Additionally, since both conditions faced the same computerized investor, the higher post-intervention returns are not solely due to differences in investor behavior. While the investor reacts to participants' returns, with higher returns generally leading to higher investments, this is driven by the

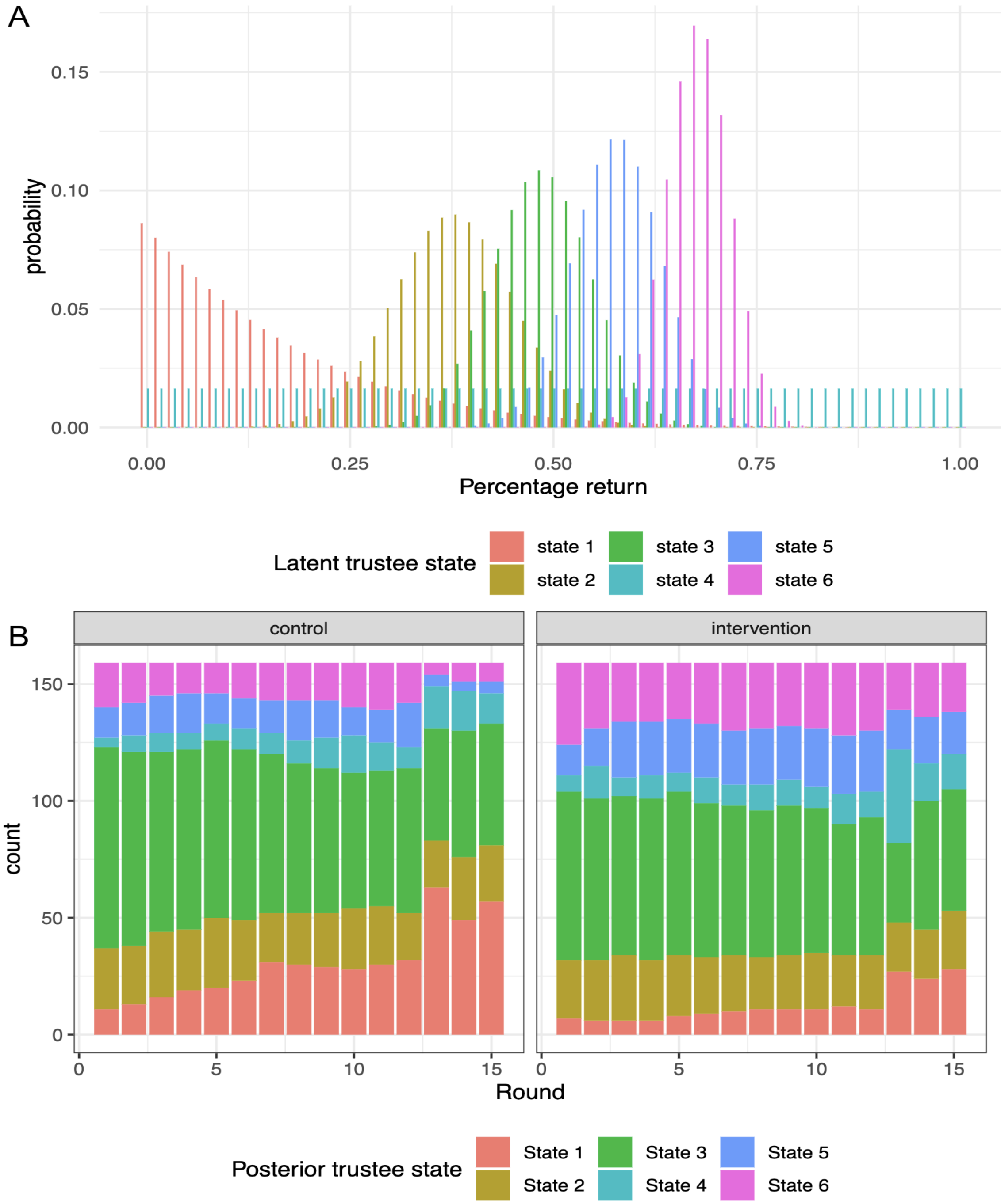

Figure 3: Panel A: the distribution of participants' percentage return for each of the latent states in the 6 state HMM-full model shows distinct policies centered around different return levels. The latent states are ordered by the mean of the discretized Gaussian representing the policy in that state, so higher numbered states can be considered more pro-social. Panel B: 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. The 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 is visibly more weighted towards higher return states. As in Panel A, states here are represented in increasing degrees of average percentage returns from the lowest (state 1) to the highest (state 6) return state.

magnitude of participants' returns, not by a change in the strategy of the investor. Furthermore, the absence of differences between conditions in participants' ratings of the first and second HMM agent suggests that the increased returns are also not due to a more favorable evaluation of the investor.

Because the intervention explicitly advocated a cooperative response, demand effects are a concern. However, the effect did not generalize to the Prisoner's Dilemma, arguing against a global "please-the-experimenter" bias. While we cannot exclude demand characteristics entirely, the pattern suggests a context-specific change in trust-repair behavior rather than generalized compliance.

Interestingly, across both conditions, we observed a decrease in both emotional positivity and arousal in the second RTG compared to the first. This general decline in emotional intensity is likely attributable to factors such as decreased engagement or increased fatigue as the experiment progressed, rather than being a specific effect of the intervention. Importantly, despite this reduction in emotional intensity, participants in the intervention condition maintained higher levels of cooperative behavior, suggesting that the intervention may have promoted more deliberate, strategic decision-making rather than emotionally-driven responses.

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 following the intervention about whether they would change their behavior. 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 & Cramer, 2017). This view was used to justify computational psychiatry's difficulty in establishing differential and reliable predictors of treatment response (Hitchcock et al., 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 response to treatment 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.

These individual differences notwithstanding, our findings have broader implications for understanding therapeutic mechanisms. They show that a core component of Dialectical Behavior Therapy affects reactions to social defections in a highly controlled laboratory setting. This is particularly remarkable given that these effects emerged in individuals without clinical diagnoses, suggesting that therapeutic mechanisms may be more fundamental and broadly applicable than previously recognized. The clinical literature indicates DBT effectively treats borderline personality disorder, a condition characterized by profound deficits in repairing social trust after interpersonal violations (King-Casas et al., 2008). Our results provide experimental validation for the theoretical mechanisms underlying this therapeutic approach. This study contributes to an emerging paradigm investigating psychotherapeutic mechanisms through computational and experimental methods. Recent work has shown that cognitive interventions influence effort sensitivity (Norbury et al., 2024), distancing techniques impact emotional dynamics (Malamud et al., 2024), and behavioral activation relates to Pavlovian biases (Huys et al., 2022). These studies are laying crucial foundations for detailed computational and neurobiological investigations of how psychotherapeutic interventions create change. More broadly, our findings suggest that explicit psychoeducation about defection responses could prove valuable beyond therapeutic settings—in organizational dynamics, educational environments, and conflict resolution. The intervention's specificity (trust games but not Prisoner's Dilemma) suggests that targeted approaches could be developed for specific social contexts, offering precision methods for improving interpersonal functioning.

Despite these complexities, we are encouraged that our brief cognitive intervention led to differentiated behavior. Future studies could explore improved cognitive interventions to enhance cooperative behavior, possibly making them more interactive and engaging. Testing such interventions with participants who struggle to repair relationships after trust breakdowns, such as those with Borderline Personality Disorder, could be particularly valuable. The relative ease of delivering online interventions and repeated interactions and training with artificial but human-like agents, open up possibilities for efficient, low-cost treatment programs to help a wide variety of people overcome tendencies for detrimental actions in social situations.

Constraints on generality

Following recommendations by Simons et al. (2017), we detail the limits on generalizability of our findings across participants, materials, procedures, and contexts.

Participants: Our sample comprised adult participants recruited from Prolific Academic, predominantly from Western countries with internet access. The generalizability to non-WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations remains uncertain, as cultural differences in trust norms, cooperation patterns, and responses to cognitive interventions may moderate effects. Additionally, our sample consisted of individuals without diagnosed mental health conditions. Whether similar effects would emerge in clinical populations (e.g., those with borderline personality disorder, who show particular difficulties with trust repair) requires direct empirical testing.

Materials: The intervention consisted of a brief text-based psychoeducational module inspired by DBT principles, presented in a single session. Its multi-component nature (combining psychoeducation, consequence reflection, and behavioral guidance) makes it challenging to disentangle specific active ingredients; factorial designs would be needed to isolate individual component effects. Effects might differ with alternative intervention formats (e.g., video-based, interactive, or multi-session interventions) or with interventions drawing from other therapeutic traditions (e.g., cognitive-behavioral therapy, acceptance and commitment therapy). The computerized investor was based on a Hidden Markov Model trained on human data, providing consistent but probabilistic responses. Results may not generalize to interactions with human partners, who bring additional complexities such as theory of mind, strategic sophistication, and emotional reactivity.

Procedures: The study employed a between-subjects design with participants randomly assigned to intervention or control conditions, with the control condition involving an active task (anagram solving). Effects might differ under within-subjects designs or with different control conditions (e.g., no-task control, attention placebo). Participants played as trustees (second movers), which allowed us to examine responses to trust violations but limits direct assessment of changes in trust as traditionally conceptualized in the investor (first-mover) role. The Repeated Trust Game structure (15 rounds, tripling multiplier, programmed defection on round 12 pre-intervention and round 13 post-intervention) represents a specific instantiation of cooperative exchanges. Variations in game length, payoff structures, timing of trust violations, or frequency of violations could moderate effects. The lack of transfer to Prisoner's Dilemma already demonstrates some boundary conditions—unlike the RTG where signaling trustworthiness can increase future investments, the short-horizon Prisoner's Dilemma with a predictable tit-for-tat opponent reduces the value of unilateral "coaxing" after a one-off defection.

Context: Data collection occurred entirely online, with participants completing the study independently in their own environments. Laboratory settings with in-person interactions might yield different results. The economic incentive structure (£5 base payment plus performance-contingent bonus averaging £0.71) may have influenced engagement and decision-making; different payment schemes could moderate effects.

Invariances: Based on our design and results, we expect the core finding—that brief cognitive interventions can reduce retaliatory responses to trust violations—would remain robust across variations in (1) the specific wording of the intervention, provided core DBT principles are preserved; (2) minor variations in Trust Game parameters (e.g., doubling vs. tripling multiplier); and (3) the specific platform for online data collection. However, these expectations require empirical verification.

References

Allen, J. G., & Fonagy, P. (Eds.). (2006). *The handbook of mentalization-based treatment* (pp. xxi, 340). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470712986

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*(5), 1090–1112. https://doi.org/10.1037/a0030939

Bendor, J., Kramer, R. M., & Stout, S. (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

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

Burnham, T., McCabe, K., & Smith, V. L. (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, G., Cobo-Reyes, R., & Jiménez, N. (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

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Cochard, F., Van, P. N., & Willinger, M. (2004). Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization*, *55*(1), 31–44.

Drążkowski, D., Kaczmarek, L. D., & Kashdan, T. B. (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*, 148–153. https://doi.org/10.1016/j.paid.2017.01.043

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fiedler, M., & Haruvy, E. (2017). The effect of third party intervention in the trust game. *Journal of Behavioral and Experimental Economics*, *67*, 65–74. https://doi.org/10.1016/j.socec.2016.10.003

Fonagy, P., Luyten, P., Moulton-Perkins, A., Lee, Y.-W., Warren, F., Howard, S., Ghinai, R., Fearon, P., & Lowyck, B. (2016). Development and Validation of a Self-Report Measure of Mentalizing: The Reflective Functioning Questionnaire. *PLOS ONE*, *11*(7), e0158678. https://doi.org/10.1371/journal.pone.0158678

Fried, E. I., & Cramer, A. O. J. (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

Gratz, K. L., & Roemer, L. (2004). Multidimensional Assessment of Emotion Regulation and Dysregulation: Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment*, *26*(1), 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94

Harth, N. S., & Regner, T. (2017). The spiral of distrust: (Non-)cooperation in a repeated trust game is predicted by anger and individual differences in negative reciprocity orientation. *International Journal of Psychology*, *52*(S1), 18–25. https://doi.org/10.1002/ijop.12257

Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational Psychiatry Needs Time and Context. *Annual Review of Psychology*, *73*(1), 243–270. https://doi.org/10.1146/annurev-psych-021621-124910

Huys, Q. J. M., Russek, E. M., Abitante, G., Kahnt, T., & Gollan, J. K. (2022). Components of behavioral activation therapy for depression engage specific reinforcement learning mechanisms in a pilot study. *Computational Psychiatry*, *6*(1), 238–255. https://doi.org/10.5334/cpsy.81

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

Joyce, B., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The Rupture and Repair of Cooperation in Borderline Personality Disorder. *Science*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, *364*(9432), 453–461. https://doi.org/10.1016/S0140-6736(04)16770-6

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder* (pp. xvii, 558). Guilford Press.

Linehan, M. M. (2015). *DBT® skills training manual, 2nd ed* (pp. xxiv, 504). Guilford Press.

Malamud, J., Guloksuz, S., Winkel, R. van, Delespaul, P., De Hert, M. A. F., Derom, C., Thiery, E., Jacobs, N., Os, J. van, & Huys, Q. J. M. (2024). Characterizing the dynamics, reactivity and controllability of moods in depression with a kalman filter. *PLOS Computational Biology*, *20*(10), e1012457. https://doi.org/10.1371/journal.pcbi.1012457

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Morey, L. C. (1991). *The Personality Assessment Inventory TM: Professional Manual*. PAR, Psychological Assessment Resources, Incorporated.

Norbury, A., Robbins, T. W., Seymour, B., et al. (2024). Different components of cognitive-behavioral therapy affect specific cognitive mechanisms. *Science Advances*, *10*(1), adk3222. https://doi.org/10.1126/sciadv.adk3222

Rigdon, M. L., McCabe, K. A., & Smith, V. L. (2007). Sustaining Cooperation in Trust Games. *The Economic Journal*, *117*(522), 991–1007. https://doi.org/10.1111/j.1468-0297.2007.02075.x

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review*, *23*(3), 393–404. https://www.jstor.org/stable/259285

Rudge, S., Feigenbaum, J. D., & Fonagy, P. (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, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

Servátka, M., Tucker, S., & Vadovič, R. (2011). Words speak louder than money. *Journal of Economic Psychology*, *32*(5), 700–709. https://doi.org/10.1016/j.joep.2011.04.003

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2022). *Afex: Analysis of Factorial Experiments*.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. *Current Anthropology*, *53*(6), 673–692. https://doi.org/10.1086/668207

Visser, I., & Speekenbrink, M. (2021). *depmixS4: Dependent Mixture Models - Hidden Markov Models of GLMs and Other Distributions in S4*.

Visser, I., & Speekenbrink, M. (2022). Hidden Markov Models. In I. Visser & M. Speekenbrink (Eds.), *Mixture and Hidden Markov Models with R* (pp. 125–172). Springer International Publishing. https://doi.org/10.1007/978-3-031-01440-6_4