(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

Cooperation, underpinned by trust, is crucial for social interactions yet fragile once broken. This study investigates the effectiveness of a brief cognitive intervention in promoting resilient cooperative behavior in economic games. Using a randomized controlled online experiment, we examined how 318 participants, acting as trustees in the Repeated Trust Game (RTG), responded to an intervention inspired by Dialectical Behavior Therapy. Participants played two 15-round RTGs against a novel computer-simulated investor using a hidden Markov model (HMM) trained on human data, enabling adaptive, human-like behavior while maintaining experimental control. The intervention significantly increased cooperative behavior, with participants showing more resilient cooperation even after experiencing low investments. HMM analysis revealed that the intervention increased the probability of participants occupying high-cooperation states. However, these effects did not transfer to a subsequent Repeated Prisoner's Dilemma game. Our findings demonstrate the potential of targeted cognitive interventions to enhance cooperative behavior, while showcasing the value of HMM-based agents in studying dynamic social interactions.

[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.*

[*] *Email: ismail.guennouni@zi-mannheim.de*

Introduction

Cooperation, defined as individuals or entities working together towards a shared goal, plays a fundamental role in promoting collective success and social harmony (1). 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 (2). Trust enables individuals to engage in risky interactions where immediate self-interest could easily override long-term benefits (3). 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 (4). 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. Cooperation emerges when both parties act in ways that promote mutual gains over multiple rounds. However, trust is fragile, and a single instance of defection—where one player fails to reciprocate appropriately—can lead to the breakdown of cooperation (5). Once trust is violated, it is often difficult to re-establish, even if doing so would be mutually beneficial (6).

Previous research has explored ways to encourage initial cooperation in trust games, such as using third-party enforcement (7,8), gratitude priming (9), or with the concepts of friend and foe (10). While these approaches improve early cooperative behavior, they often fail to address the more challenging task of repairing cooperation after trust has been broken. Once defection occurs, individuals may react impulsively by reducing their own cooperative

efforts, even if re-establishing trust would be mutually beneficial. Thus, interventions aimed at restoring cooperation in these contexts are crucial but understudied.

In this study, we focus on the role of the trustee in the RTG, as a driver of cooperation dynamics. While the trustee does not exercise trust in the same way as the investor, their decisions to reciprocate (or not) play a critical role in maintaining or disrupting the cooperative relationship. A lack of reciprocation from trustees erodes trust over time (11) and is particularly evident when the trustee suffers from some personality disorders (12). These trustees fail to engage in trust-repairing behaviors such as coaxing the investor by signalling trustworthiness via sending high returns (13). Additionally, unpredictable behavior from trustees fosters mistrust and impedes future cooperation (14). In contrast, consistent cooperation from trustees promotes trust and encourages further collaboration, as evidenced by neural data (15). 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 (16) and Mentalisation Based Therapy (17) have shown promise in enhancing social skills among individuals with interpersonal difficulties. These approaches often focus on helping individuals recognize the impact of their actions on others and consider alternative behavioral strategies. However, response to these treatments is highly variable, and determining which interventions are effective for particular sub-groups of patients is challenging (18,19). 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 (20–22). 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, potentially extending to improving trustee behavior and fostering cooperation in the RTG.

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 Dialectical Behavior Therapy (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 behavioral change (23). Specifically, we hypothesize that encouraging participants to reflect on the consequences of their actions and 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

A total of 320 participants were recruited on the Prolific Academic platform (prolific.co). Two players had incomplete trust game entries and their data was disregarded, leaving data for 318 participants for analysis, equally split between the two conditions. The required sample size was determined using an *a priori* power analysis to have an 80% probability to detect a small effect size (Cohen's f = 0.10) for a within-between interaction with a 5% type I error rate in a repeated measures ANOVA. The sample size calculation assumed 2 groups, 2 measurement per group and was performed using the G*Power software (24). 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.

Design and Procedure

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. Participants were randomly assigned to either the intervention or control condition. Following the post-intervention Trust Game, all participants completed a Repeated Prisoner's Dilemma (RPD) game to assess potential transfer effects (See Figure 1.A for experiment overview). The games were designed and implemented online using Empirica (25). This research received approval from a UCL ethics board (ID:21029/001) and the experiment was performed in accordance with the ethics board guidelines and regulations.

Tasks and Measures

Repeated Trust Game

Participants played two separate 15-round Repeated Trust Games (4) in the role of trustee. They were told (correctly) that each 15-round game was played against the same co-player, and that they will face a new player when they started a new 15-round game. Participants were led to believe they were interacting with human co-players when in fact they faced the same computerised investor in all trust games. This design allows us to examine the effects of our intervention while controlling for individual differences in play style, as well as minimizing the influence of reputational concerns or carry-over effects that might occur if participants believed they were facing the same partner twice. In each round of the RTG, the investor (computer-simulated) was endowed with 20 units and decided how much of that endowment to invest. This investment was tripled, and the participant, as the trustee, then decided how to split this tripled amount between themselves and the investor. If the trustee returns more than one third of the amount, 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 as shown in Figure 1.C for the intervention condition.

Adaptive Computerised Investor

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) co-player from existing datasets. Full detail on the data sources used are 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 1.B). 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 1.D). 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 initial state for the HMM investor in each instance of the game was the "mid-trust" state.

Our use of HMMs to model investor behavior represents a significant advancement in experimental design for economic games. Unlike traditional fixed-strategy computer opponents, our HMM-based agents adapt their behavior based on the participant's actions, mirroring the complex decision-making processes observed in human players. This approach allows us to maintain experimental control while significantly enhancing the ecological validity of our study.

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 maximise their reward by choosing to cooperate.

In this game, the computerised agent was programmed to act according to a tit-for-tat strategy (26), 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.

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 (23). 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 order to separate any general practice effects from the effect of the intervention, we added 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.

Procedure

At the start of the experiment, participants provided informed consent and were instructed the study would consist of three phases. Participants across conditions were told their goal was to maximise 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 details 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 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. 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:

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

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

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 (27) 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 (28). 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 R package "afex". 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.

To model participants' returns in the RTG across games and conditions, we fit various hidden Markov models (29) to participants' returns using the depmixS4 package (30) 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 (31) to select the best fitting model.

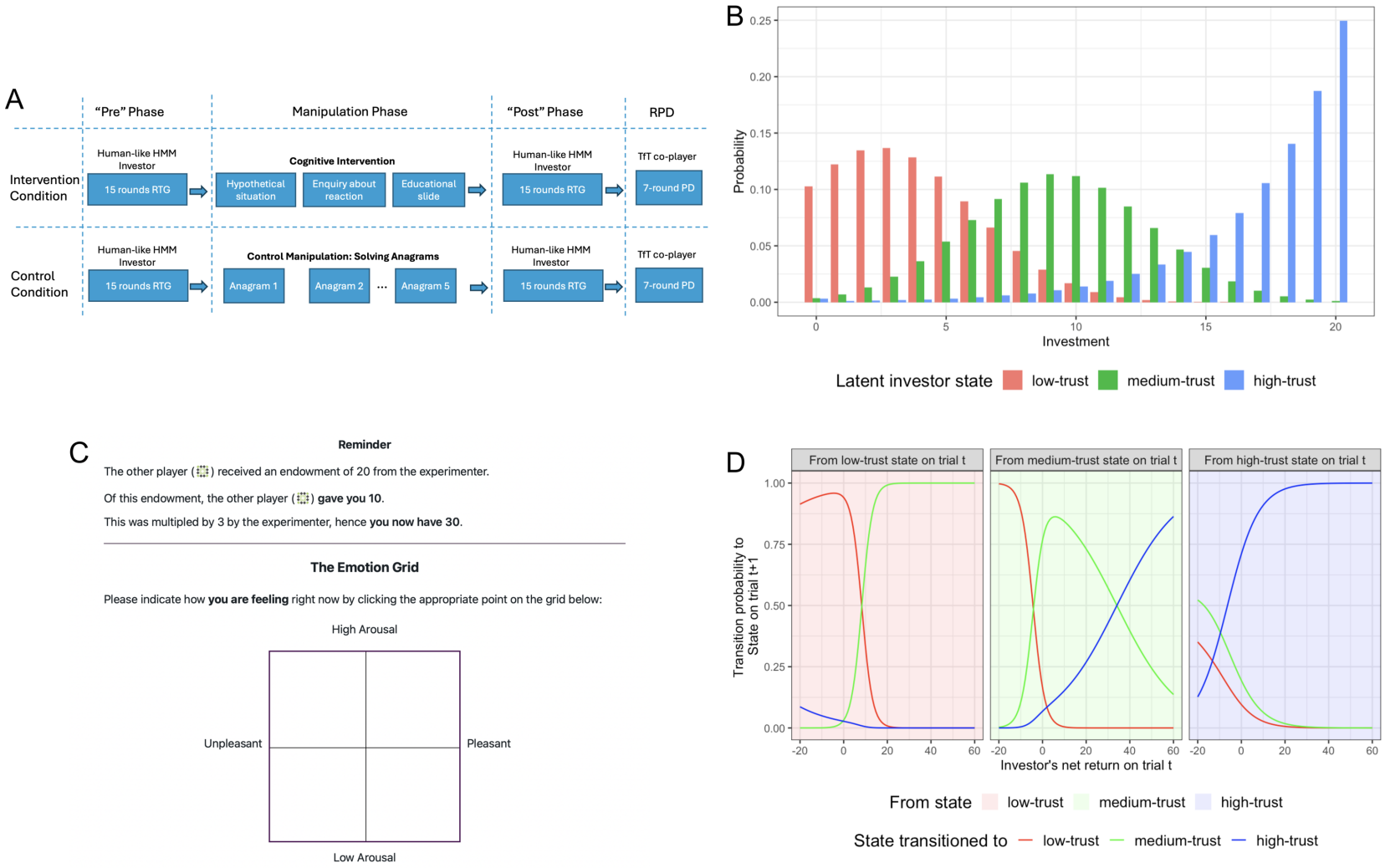

Figure 1: Panel A: Experimental design overview. Participants in both conditions played two 15-round Repeated Trust Games (RTGs) with human-like Hidden Markov Model (HMM) investors, separated by a manipulation phase. The intervention condition received a cognitive intervention, while the control condition 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. Immediately before making return decisions in the RTG, participants are shown a two dimensional grid (Panel C) where they need to report their reaction to the investment. Those in the intervention condition were asked questions about their emotional state and those in the control condition were asked to report features of the investment unrelated to their emotion. Panels B - 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 B. 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 "defection round" (Figure 2.A) were within the range of reported investments (40-60% of endowment) and returns (35-50% of total yield) in the literature (7,32). Mixed-effects analysis on the percentage returns shows a significant main effect of Condition (intervention vs. control), $F(1, 317.20) = 9.53$, $p = .002$, due to overall 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 2.C). 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 that returns were more affected by investments in the control 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 the HMM investors' behavior across games and conditions, 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 2.D, 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 low investment by the HMM (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 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, pre-post, $\Delta M = 0.07$, 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.

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 in our questionnaires.

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 (Figure 2.B).

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 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, 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.

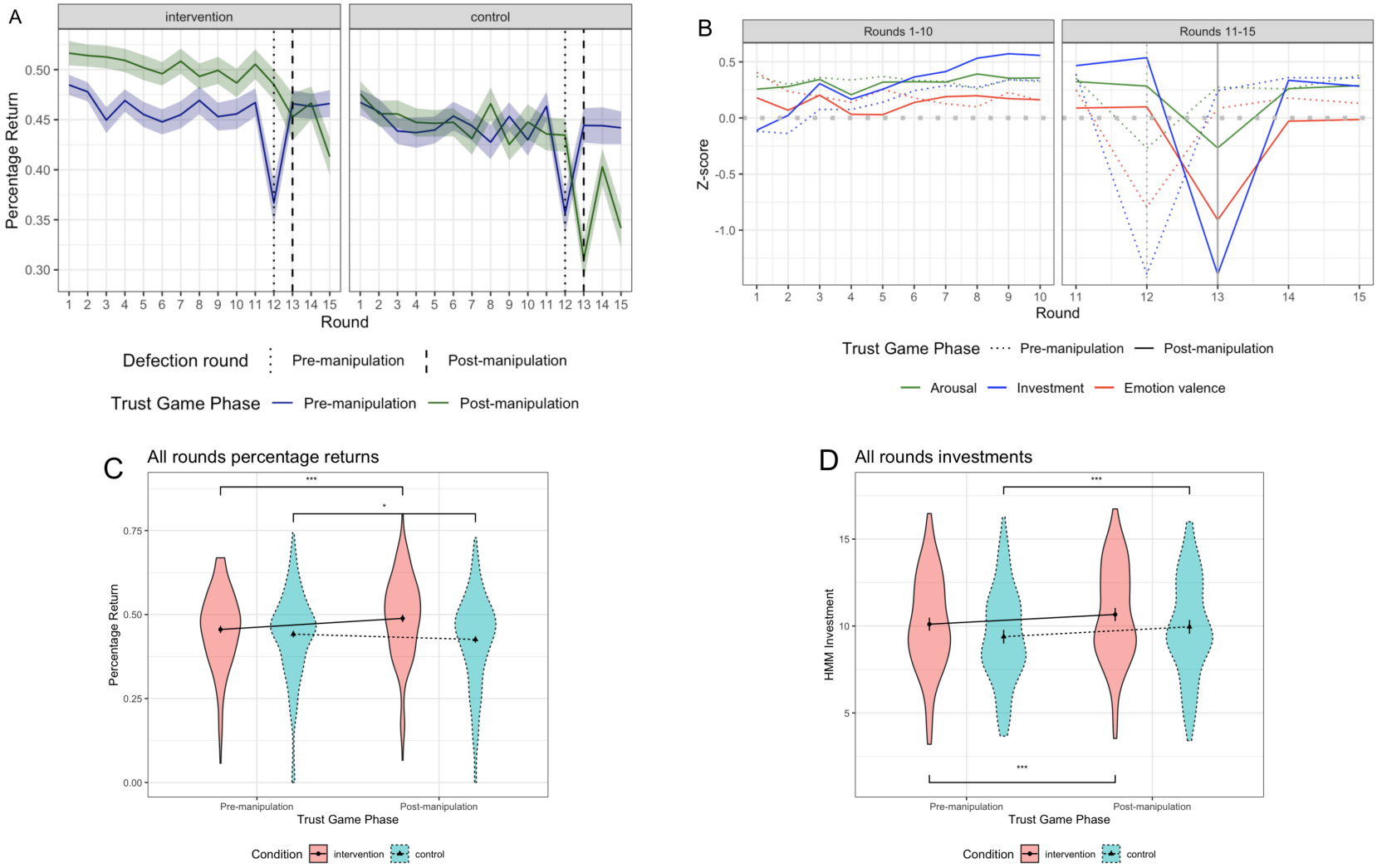

Figure 2: Panel A: Averages and standard errors of the trustee's return as a percentage of the multiplied investment received by Condition, Phase, and game round. We note a different reaction to the pre-programmed one-off low investment (vertical black lines) between the two conditions: Whilst there 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. 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' emotion reaction measured during the investor's pre-programmed one-off low investment round (vertical grey lines) were similar before (round 12) and after (round 13) the intervention. The bottom panels show marginal means and distributions of either investments or percentage returns across participants by Game and Condition. Panel C shows that 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, 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.

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. The estimated state-dependent policy of trustee actions, according to the HMM-full model, is depicted in Figure 3.A.

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 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 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.

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 and reduced their returns post-defection. An HMM analysis showed that those in the intervention condition were less prone to low-return states after defection. These findings suggest that cognitive interventions can promote resilience in cooperative behaviors following trust violations.

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

The increased cooperation observed post-intervention could potentially be attributed to demand expectancy effects (33,34), given that the educational slide explicitly suggests a preferred reaction to co-player defection. However, several factors challenge this interpretation. First, participants in the intervention group exhibited higher returns on average even before encountering the pre-programmed defection that the intervention specifically addressed. This behavior diverges from what would be expected if participants were merely responding to perceived experimental

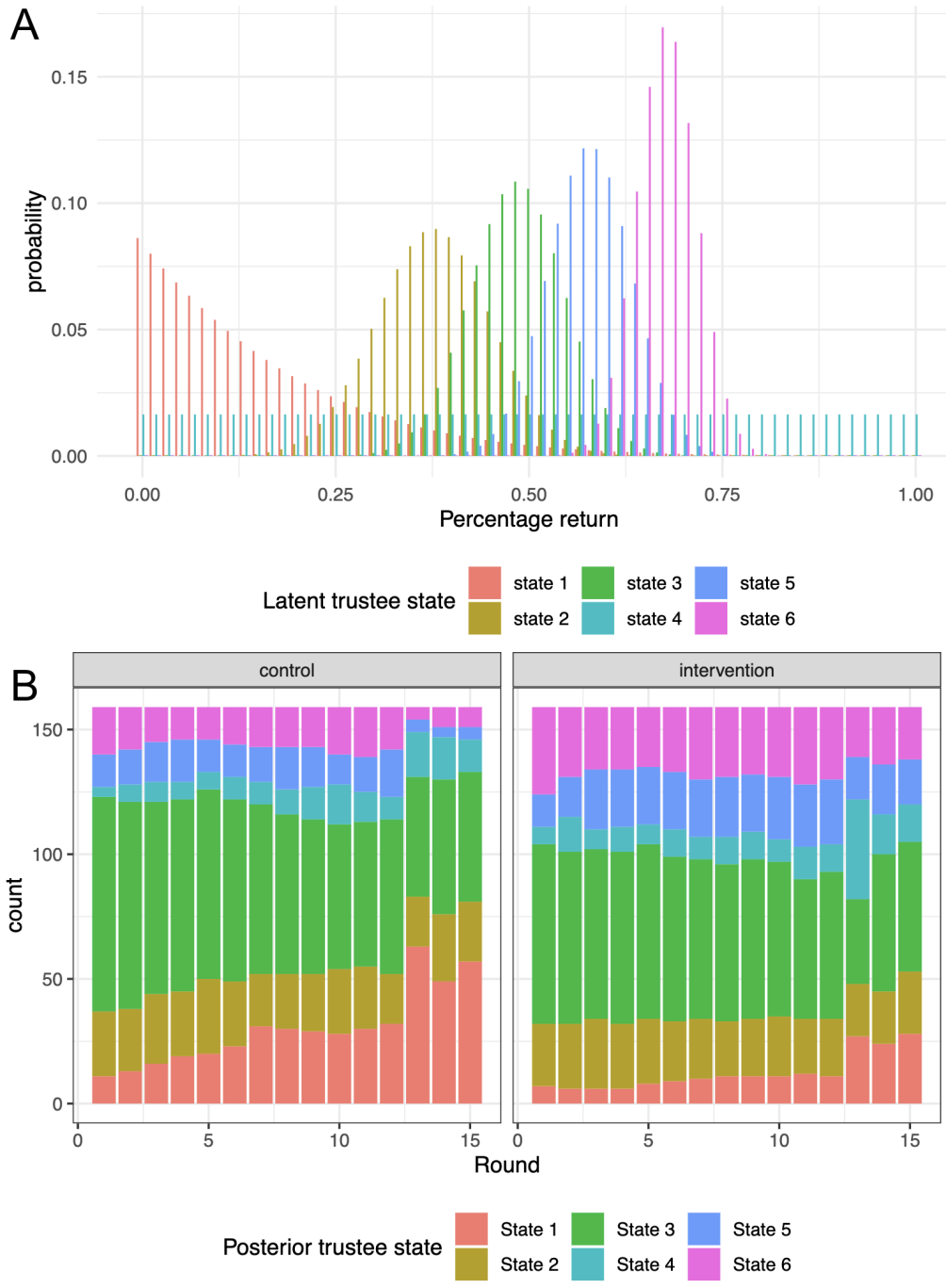

Figure 3: Panel A: the distribution of participants' percentage return for each of the latent states in the 6 state HMM-full model shows distinct policies centered around different return levels. The latent states are ordered by the mean of the discretized Gaussian representing the policy in that state, 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.

demands. It is still possible that participants inferred from the intervention a "demand" to behave in cooperative ways in general. However, the effect of the intervention did not transfer to the Repeated Prisoner's Dilemma (RPD) game, with no difference between conditions in cooperation rates. This specificity argues against the intervention merely prompting participants to align with perceived experimenter expectations. If participants were simply trying to please the experimenter by acting cooperatively post-intervention, we would expect to see elevated cooperation rates in the RPD game for the intervention group. The absence of such an effect suggests that the intervention's impact was specific to the Trust Game context. Rather, we posit that the increased returns stem from participants inferring that pro-social and trustworthy behavior may motivate the investor to send high investments, potentially leading to more beneficial long-term outcomes.

While we cannot entirely rule out the influence of demand characteristics on the observed behavior, it's important to consider whether this effect, if present, diminishes the intervention's value, particularly in clinical contexts. For individuals with high interpersonal dysfunction, such as those with Borderline Personality Disorder, the ability to recognize and respond to social cues - even if initially driven by perceived expectations - can be a crucial step towards more adaptive behavior (16). Research has shown that explicitly teaching social skills and appropriate responses can lead to improved outcomes in various clinical populations (35). Moreover, the concept of "therapeutic demand characteristics" has been proposed, suggesting that aligning with perceived therapeutic expectations can be an active ingredient in treatment effectiveness (36). In the context of our study, even if participants' initial cooperation was partly motivated by perceived experimenter expectations, this could potentially translate into more adaptive real-world behaviors over time. This perspective aligns with the principles of cognitive-behavioral therapies, which often explicitly teach and reinforce desired behaviors (37). Therefore, rather than viewing potential demand characteristics as a limitation, we might consider them as a possible mechanism for initiating behavioral change, particularly in populations that may not inherently be motivated to please therapists or adhere strictly to treatment protocols.

Interestingly, we observed a decrease in both emotional positivity and arousal in the second RTG compared to the first, across both conditions. 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 about whether they would change their behavior, just after receiving the intervention. 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 (38). This view was used to justify computational psychiatry's difficulty in establishing differential and reliable predictors of likely treatment response (39). 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.

Several limitations of this study warrant consideration. First, the multi-component nature of our intervention makes it challenging to disentangle the specific effects of each element. While this design reflects real-world therapeutic approaches, future research could employ factorial designs to investigate the relative contributions of individual components. Second, our focus on the trustee role in the RTG and the use of HMMs as co-players, while allowing us to examine responses to potential trust violations, limits our ability to directly assess changes in trust as traditionally conceptualized in the first-mover role. Finally, the brief nature of our intervention and the use of a general population sample may limit generalizability to clinical settings or longer-term behavioral changes.

Despite these limitations, 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 visually appealing. 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 ease of assigning online interventions and the use of artificial but human-like agents open up possibilities for efficient, low-cost treatment programs to help a wide variety of people overcome detrimental actions in social situations.

Author contributions statement

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.

Competing interests statement

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Additional information

Correspondence

All correspondence and requests for materials should be addressed to I. Guennouni.

Transparency and data availability

Preregistration: The hypotheses and methods were not preregistered.The analysis plan was not preregistered. Materials: All study materials are publicly available (https://github.com/ismailg/CoaxIntervention). Data: All primary data are publicly available (https://github.com/ismailg/CoaxIntervention). Analysis scripts: All analysis scripts are publicly available (https://github.com/ismailg/CoaxIntervention).

References

1. Tomasello, M., Melis, A. P., Tennie, C., Wyman, E. & Herrmann, E. Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis. *Current Anthropology* **53,** 673–692 (2012).
2. Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review* **23,** 393–404 (1998).
3. Balliet, D. & Van Lange, P. A. M. Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin* **139,** 1090–1112 (2013).
4. Joyce, B., Dickhaut, J. & McCabe, K. Trust, Reciprocity, and Social History. *Games and Economic Behavior* **10,** 122–142 (1995).
5. Bendor, J., Kramer, R. M. & Stout, S. When in Doubt...: Cooperation in a Noisy Prisoner's Dilemma. *Journal of Conflict Resolution* **35,** 691–719 (1991).
6. Harth, N. S. & Regner, T. 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,** 18–25 (2017).
7. Charness, G., Cobo-Reyes, R. & Jiménez, N. An investment game with third-party intervention. *Journal of Economic Behavior & Organization* **68,** 18–28 (2008).

8. Fiedler, M. & Haruvy, E. The effect of third party intervention in the trust game. *Journal of Behavioral and Experimental Economics* **67,** 65–74 (2017).
9. Drążkowski, D., Kaczmarek, L. D. & Kashdan, T. B. 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 (2017).
10. Burnham, T., McCabe, K. & Smith, V. L. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization* **43,** 57–73 (2000).
11. Servátka, M., Tucker, S. & Vadovič, R. Words speak louder than money. *Journal of Economic Psychology* **32,** 700–709 (2011).
12. Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M. & Bohus, M. Borderline personality disorder. *The Lancet* **364,** 453–461 (2004).
13. King-Casas, B. *et al.* The Rupture and Repair of Cooperation in Borderline Personality Disorder. *Science* **321,** 806–810 (2008).
14. Rigdon, M. L., McCabe, K. A. & Smith, V. L. Sustaining Cooperation in Trust Games. *The Economic Journal* **117,** 991–1007 (2007).
15. King-Casas, B. *et al.* Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science* **308,** 78–83 (2005).
16. Linehan, M. M. *Cognitive-behavioral treatment of borderline personality disorder.* xvii, 558 (Guilford Press, 1993).
17. *The handbook of mentalization-based treatment.* xxi, 340 (John Wiley & Sons, Inc., 2006). doi:10.1002/9780470712986
18. Rudge, S., Feigenbaum, J. D. & Fonagy, P. 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,** 92–102 (2020).
19. Arch, J. J., Wolitzky-Taylor, K. B., Eifert, G. H. & Craske, M. G. Longitudinal treatment mediation of traditional cognitive behavioral therapy and acceptance and commitment therapy for anxiety disorders. *Behaviour Research and Therapy* **50,** 469–478 (2012).
20. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience* **19,** 404–413 (2016).
21. Reiter, A. M., Atiya, N. A., Berwian, I. M. & Huys, Q. J. Neuro-cognitive processes as mediators of psychological treatment effects. *Current Opinion in Behavioral Sciences* **38,** 103–109 (2021).
22. Dercon, Q. *et al.* A core component of psychological therapy causes adaptive changes in computational learning mechanisms. *Psychological Medicine* **54,** 327–337 (2024).
23. Linehan, M. M. *DBT® skills training manual, 2nd ed.* xxiv, 504 (Guilford Press, 2015).
24. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* **41,** 1149–1160 (2009).
25. Almaatouq, A. *et al.* Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods* **53,** 2158–2171 (2021).
26. Axelrod, R. & Hamilton, W. D. The Evolution of Cooperation. *Science* **211,** 1390–1396 (1981).
27. Singmann, H. *et al.* Afex: Analysis of Factorial Experiments. (2022).
28. Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H. & Bates, D. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* **94,** 305–315 (2017).
29. Visser, I. & Speekenbrink, M. in *Mixture and Hidden Markov Models with R* (eds. Visser, I. & Speekenbrink, M.) 125–172 (Springer International Publishing, 2022). doi:10.1007/978-3-031-01440-6_4
30. Visser, I. & Speekenbrink, M. depmixS4: Dependent Mixture Models - Hidden Markov Models of GLMs and Other Distributions in S4. (2021).
31. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6,** (1978).
32. Fiedler, M., Haruvy, E. & Li, S. X. Social distance in a virtual world experiment. *Games and Economic Behavior* **72,** 400–426 (2011).
33. Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* **17,** 776–783 (1962).
34. Nichols, A. L. & Maner, J. K. The Good-Subject Effect: Investigating Participant Demand Characteristics. *The Journal of General Psychology* **135,** 151–166 (2008).
35. Bornovalova, M. A. & Daughters, S. B. How does Dialectical Behavior Therapy facilitate treatment retention among individuals with comorbid borderline personality disorder and substance use disorders? *Clinical Psychology Review* **27,** 923–943 (2007).

36. Trachsel, M. & Grosse Holtforth, M. How to Strengthen Patients' Meaning Response by an Ethical Informed Consent in Psychotherapy. *Frontiers in Psychology* **10,** 1747 (2019).
37. Beck, J. S. *Cognitive behavior therapy : Basics and beyond.* (The Guilford Press, 2011).
38. Fried, E. I. & Cramer, A. O. J. Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology. *Perspectives on Psychological Science* **12,** 999–1020 (2017).
39. Hitchcock, P. F., Fried, E. I. & Frank, M. J. Computational Psychiatry Needs Time and Context. *Annual Review of Psychology* **73,** 243–270 (2022).