Repairing cooperation through a cognitive intervention in the repeated Trust Game

Ismail Guennouni Maarten Speekenbrink Quentin Huys Samuel Dupret

Abstract

Social trust is an important building block of strong social bonds, and its absence is a risk factor for social dysfunction. As such, interventions to foster and strengthen trust-based cooperation are highly desirable. Using the repeated Trust Game paradigm, we assess the effectiveness of a cognitive intervention aimed at repairing the potential breakdown of cooperation from a pre-programmed, one-off defection by the opponent. Over two games, participants are given the role of the trustee and face what they believe are two different players. In between games, they either receive a brief cognitive intervention or not. The intervention led to more cooperative behavior both pre and post defection by the opponent. HMM modelling of participants actions shows participants in the intervention group had a lower probability of transitioning to non cooperative states. Posterior latent state analysis also showed a higher proportion of players best described by more cooperative latent states in the intervention condition compared to the control condition.

1 Introduction

At the core of social interaction is figuring out the goals, intentions, and decision-making process of the interaction partner. This inference is however fraught with uncertainty, as we cannot reasonably observe these features without prior knowledge of the person. Absent a history of interaction, one needs to decide whether to take the risk of trusting others. The challenge of trust is that it is by construction a risky endeavour. For example, if we deem a person trustworthy, we might decide to take the risk of investing in the relationship hoping for a collaborative outcome. If we misplace our trust, this can come at a high cost to us. Not trusting others is also risky since opportunities for cooperation may be foregone.

Evidence from the literature emphasises the importance of social trust in determining why some people fare better than others physically and mentally (Giordano and Lindström 2016 ; Meng and Chen 2014). It affects the health status of individuals through reinforcing social support networks, maintaining community norms and facilitating collective action. Research into the determinants of psychopathology has linked trust-based constructs to the emergence of mental health disorders. Fonagy and Allison (2014) identified epistemic trust, or the belief in the authenticity and personal relevance of interpersonally transmitted knowledge, as an important function of early attachment relationships. It allows the individual receiving social information to let go of their natural self-protective vigilance, which can become a pathological hypervigilance frequently observed after traumatic experiences and a key factor for the emergence of multiple mental health disorders (Fonagy and Campbell 2017).

Since a lack of epistemic trust is a risk factor for social dysfunction (Fonagy and Campbell 2017), and given the importance of trust for building and maintaining strong social bonds, interventions to foster and strengthen trust-based cooperation would be highly beneficial to society. Such interventions would allow people to more easily repair broken relationships, and continue harvesting the benefits of cooperation even in the presence of accidental or intentional social norm violations.

A well-established paradigm in the study of trust is the repeated trust game (Joyce, Dickhaut, and McCabe 1995). To encourage the emergence and maintenance of trust in this setting, some studies focused on modifying the game mechanism (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler and Haruvy 2017).

Others chose to intervene directly on the participants. For example, Drążkowski, Kaczmarek, and Kashdan (2017) found that trust was increased when participants were asked to think about and write down five things that they were grateful for. Burnham, McCabe, and Smith (2000) found that trust was also increased when participants were primed with the concepts of friend and foe.

Whilst these interventions show that it is possible to improve cooperative outcomes at the start of the game, they did not address how to repair a breakdown of trust that might occur due to intentional or accidental non-cooperative actions by the players. Indeed, cooperative play in the repeated trust game can easily break down when there is a transgressive behavior, such as a nil or very low investment by the investor after cooperation has been established, or a return of the trustee below the investment sent. Such ruptures of cooperation appear frequently when the trustee suffers from mental health disorders affecting the social domain such as Borderline Personality Disorder (Lieb et al. 2004). In theses situations, BPD trustees fail to engage in trust repairing behavior such as coaxing the investor through sending high return to signal trustworthiness, and this failure may be linked to a failure to perceive their low returns in the game as a violation of social norms (King-Casas et al. 2008).

In devising potential interventions to repair trust, we can derive inspiration from the cognitive interventions championed by successful psychological therapies that aim to improve aspects of interpersonal dysfunction in BPD patients. Whilst there is no proven pharmacological therapy for BPD, some forms of psychotherapy such as Mentalisation Based Therapy (Allen and Fonagy 2006) and Dialectical Behavior Therapy (Linehan 1993), have been clinically validated as efficacious approaches to improve various dysfunctional behaviors in BPD patients, including those related to social interaction (Gunderson et al. 2018). However, these therapies suffer from a high variability in treatment responses. Determining which interventions are effective for particular patients has been very challenging (Rudge, Feigenbaum, and Fonagy 2020; Arch et al. 2012). One promising approach is the study of how specific psychotherapeutic treatment components affect quantitative markers of behavior such as those inferred through computational modelling techniques (Huys, Maia, and Frank 2016; Reiter et al. 2021). As such, using specific cognitive probes inspired by therapeutic interventions complemented by computational modelling of behavior in tasks, may allow us to uncover cognitive mechanisms targeted by common forms of psychotherapy. This may in turn form the basis for targeting of existing psychotherapies to different individuals.

In this study, we assess the effectiveness of a cognitive intervention aimed at repairing the potential breakdown of cooperation from a pre-programmed, one-off low investment sent by an computer investor. The intervention focuses on explaining the potential harm from reciprocating non-cooperative actions and suggesting a non-impulsive course of action to coax the investor back into cooperation. Participants are given the role of the trustee and are randomly assigned to either a control group or an intervention group. They play two instances of the repeated Trust Game facing what they believe are two different players. In between games, they either receive a cognitive intervention (intervention condition) or not (control condition, where participants are asked to solve anagrams). In reality, participants face the same computerised agent in both instances, which is programmed to play according to an HMM fitted to real players data. We explore whether the intervention has an effect on the behaviour of the trustee and whether the learning is transferred to a new repeated Prisoner's Dilemma game when facing a seemingly new player.

2 Method

2.1 Participants and Design

A total of 318 participants were recruited on the Prolific Academic platform (prolific.co). The mean age of participants was 31.3 years. Participants were paid a fixed fee of £5 plus a bonus dependent on their performance. The experiment had a 2 (Condition: Intervention or Control) by 2 (Game : Pre or Post Intervention) design, with repeated measures on the first factor. Participants were randomly assigned to one of the two levels of the second factor.

2.2 Procedure

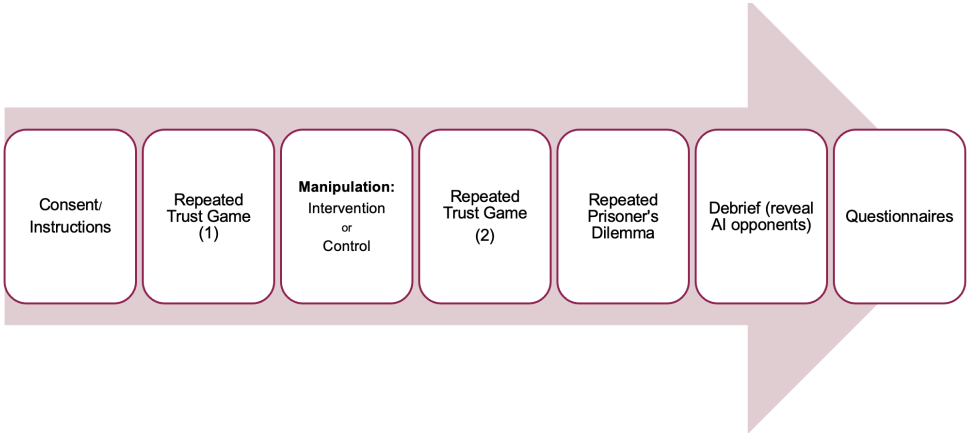

Figure 1: A) Experiment overview. After playing 15 rounds of a repeated trust game (RTG) as trustee, participants were randomized to either receive the control or active intervention. They then played the second set of 15 rounds of RTG (again as trustee) to examine intervention effects, and 7 rounds of a repeated prisoner's dilemma to examine generalization of intervention effects. Finally, participants answered questionnaires and were debriefed.

The experiment as shown in Figure 1 began with participants briefed there will be three phases, and that they would face a new opponent in each phase. Phase one was a 15 round Trust Game (RTG) in which participants had the role of the trustee, facing the same investor opponent. Participants were then assigned to either the control or intervention conditions. Phase two was similar to Phase one, with participants being told they faced a new player. In both phases, participants were asked for feedback in each round after seeing the investment. The nature of the feedback solicited varied by condition, either emotion-related (intervention) or unrelated (control). Participants rated their interaction partner on various attributes after each phase. More details on these attributes, nature of feedback and exact instructions are found in the supplement. Phase three involved 7 rounds of the repeated prisoner's dilemma game (RPD) with a third player. Participants were explicitly told to aim to maximise the number of points in all phases. The total number of rounds was not communicated to the participants in any of the games played.

Post-experiment, participants completed questionnaires related to mentalising abilities, emotion regulation and BPD traits (see details in supplement). They were then quizzed on their game-play, including whether they detected that all opponents were computer agents. The reveal explained that they were facing identical AI agents and that any perceived opponent changes were due to participants' own behavior.

2.3 Interventions

The intervention was built on interventions from DBT skills training, asking patients to reflect on the consequences of actions taking in emotional states (Linehan 2015). Specifically, participants were presented with a low investment and asked to indicate their response. They were then invited to consider what their ultimate aim in the game was and whether this response was most likely to achieve their aim. The intervention is detailed in the supplementary information.

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

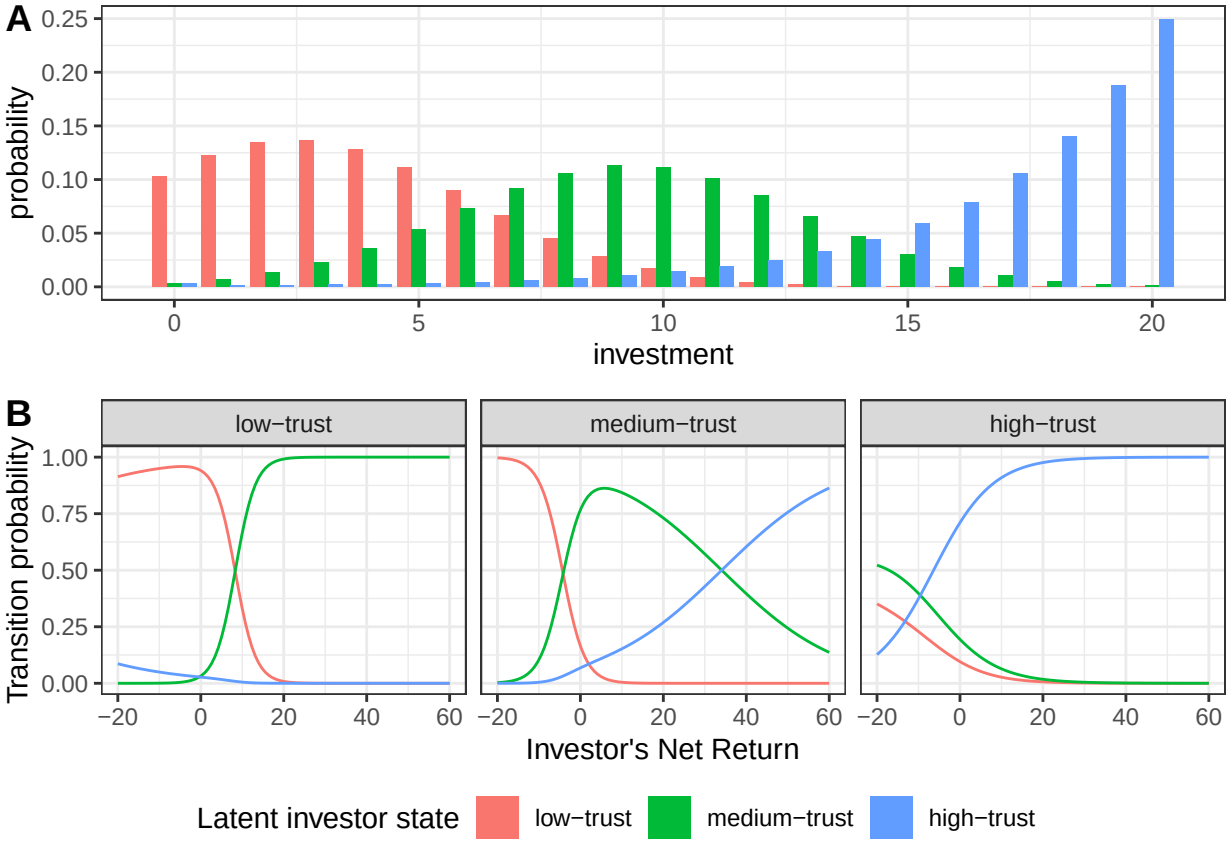

Figure 2: A: Distribution of investments by the artificial investor agent conditional on its latent state as estimated by a three state hidden Markov model fitted to human dyadic play dataset. B: Transition function for the HMM investor conditional on its current state and the net return from the previous round of play.

2.4 Design of the AI agents in the RTG and RPD

The role of the investor was in reality played by an adaptive artificial agent whose strategy was modelled on the behaviour of human participants taking the role of the investor in the RTG. Using data from human dyadic interaction in the 10 round trust game, we estimate a Hidden Markov Model (HMM) to characterise the investor's behavior as emanating from a small number of latent states (Rabiner et al. 1989). In this instance, for both iterations of the repeated trust game, the HMM used was identical: it had three states that can be described as "low-trust", "medium-trust" and "high-trust". Each latent state was associated with a distribution over all possible investor actions (from 0 to 20 investment) that reflected the amount of trust, as presented in Figure 2.A. Transition probabilities between these latent states were modelled as a function of the net return in the previous round conditional on the investor's state as shown in Figure 2.B.

In order to probe efforts to repair trust, we simulated trust breaches in both Trust Games and the Prisoner's Dilemma. The AI was programmed to break trust on specific rounds (12 pre-intervention, 13 post-intervention in the Trust Game; round 4 in Prisoner's Dilemma). After these "defection" rounds, the AI returned to its previous strategy, disregarding the participant's response to the breach. This simulated an accidental defection and a readiness to repair the interaction, providing insights into trust repair dynamics.

2.5 Analysis of participants returns in the RTG

To explore whether participants behaved differently after the intervention compared to the control group over all rounds, we estimate a linear-mixed effects model as implemented in the R package "afex" (Singmann et al. 2022), with fixed effects for Condition (intervention or control), Game-number (pre- or post-intervention) and Investment, as well as interactions between Condition and both Investment and Game-number, and participant-wise random intercepts and random slopes for Game-number. More complex models with additional random effects provided could not be estimated reliably. For the F-tests, we used the Kenward-Roger approximation to the degrees of freedom, as implemented in the R package "afex". We Z-transform the Investment variable as centering is beneficial to interpreting the main effects more easily in the presence of interactions.

To model participants returns in the RTG across games and conditions, we fit various hidden Markov models to participants returns using the depmixS4 package (Visser and Speekenbrink 2021). 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 as well as the various contrast codes used for the dummy variables are found in the supplement. We fit models with different numbers of hidden states, and use the Bayesian Information Criterion (Schwarz 1978) to select the best model.

3 Behavioral results

Average investments and returns prior to the defection round (Figure 3) were within the range of reported investments (40-60% of endowment) and returns (35-50% of total yield) in the literature (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011). Estimated marginal means of percentage returns from fitting the mixed effects model specified in the methods section, as well as the distribution of the returns for participants across all trials are presented in Figure 4.A.

We find an interaction effect between Condition (intervention vs. control) and Game-number (pre- vs. post-manipulation) ($F(1, 314.2) = 26.9, p < 0.001$). Post-hoc tests show that there was an increase in the percentage returned in the intervention group ($t(316.54) = -4.85$, $p < .001$) but not in the control group ($t(319.69) = 2.34$, $p = .020$). We also found a significant main effect for Condition ($F(1, 315.31) = 9.52, p = 0.002$)), such that the intervention group had higher percentage returns than the control group. There was also a significant main effect of Investment($F(1, 9208) = 373.6, p < 0.001$) such that higher investments were associated with higher percentage returns.

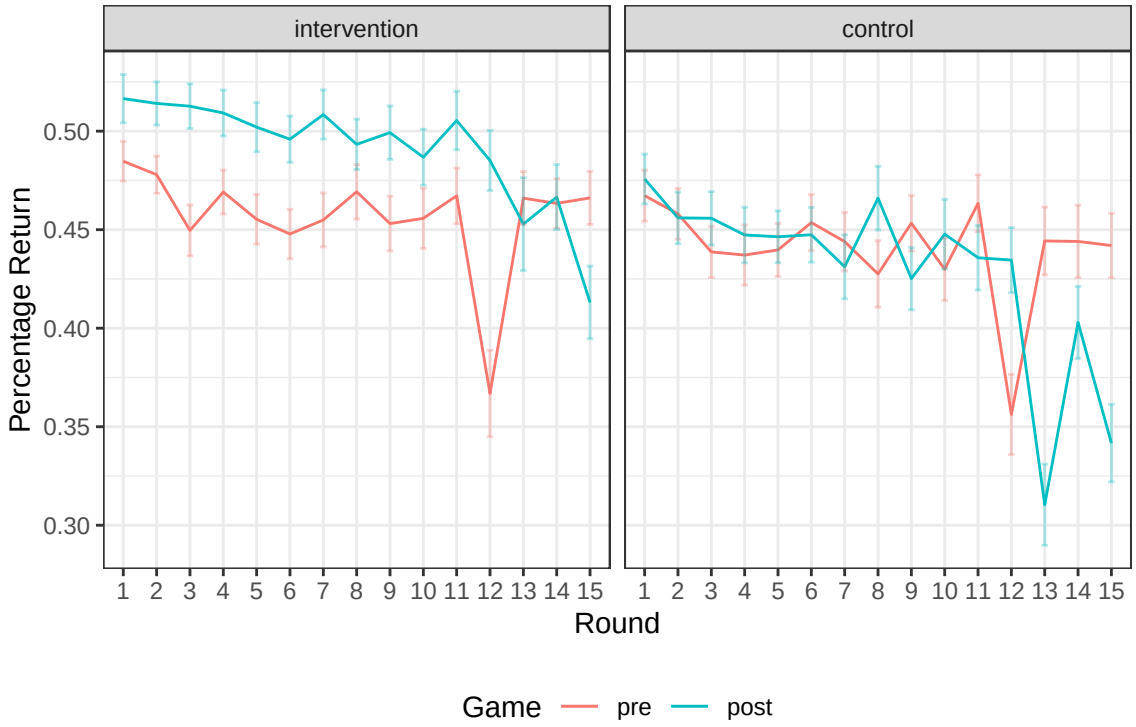

Figure 3: Average and standard errors of the trustee's return as a percentage of the multiplied investment received for each round and for both conditions. The red line shows the returns pre-manipulation and the blue line post-manipulation.We note a different reaction to the pre-programmed one-off low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

There was an interaction effect of Investment by Condition ($F(1, 9207) = 45.38, p < 0.001$), such that the positive effect of investment on percentage returns was greater in the control group than in the intervention group. This suggests that the higher returns we see in the intervention group are not simply due to higher investments. Finally, we find a three way interaction between Game-number, Condition and Investment ($F(1, 8988) = 24.6, p < 0.001$), showing that the differentiated effect of the investment on the proportion returned by condition is itself moderated by the Game-number.

The HMM agent exhibited a differentiated behavior between games and conditions. When examining the HMM agent investments, we fit a linear-mixed effects model, with fixed effects for Condition (intervention or control), Game-number (pre or post intervention), as well as interaction between Condition and Game-number, and participant-wise random intercepts and random slopes for Game-number. We find a main effect for both Condition ($F(1, 317) = 8.7, p = 0.003$) and Game-number($F(1, 317) = 8.3, p = 0.004$). As can be seen in Figure 4.D, investment was higher in the intervention compared to the control condition across games ($p = 0.003$). Across conditions, investment was also higher in the second game compared to the first ($p = 0.003$).

In summary, participants in the control group sent back **lower** percentage returns in the second game ($\Delta M = 0.02$, 95% CI $[0.00, 0.03]$, $t(319.69) = 2.34$, $p = .020$) despite the HMM investor sending, on average higher investments. Those in the intervention group returned **higher** percentage returns in the second game ($\Delta M = -0.03$, 95% CI $[-0.05, -0.02]$, $t(316.54) = -4.85$, $p < .001$), with the investor also sending higher investments. These higher returns in the intervention group compared to the control group were not purely driven by reciprocity towards higher investments since we find this interaction effect whilst controlling for investments in our model of participants returns.

3.1 Looking at pre- and post-defection trials only

We restrict our analysis to the rounds prior to the pre-programmed defection (rounds 1 to 11 pre-manipulation and 1 to 12 post-manipulation) and fit the same linear mixed effects model to percentage returns as the one for the full data. Estimated marginal means of percentage returns as well as the distribution of the returns for participants across all trials are presented in Figure 4.B. We do find an interaction effect between Condition and Game number ($F(1, 317.99) = 17.1, p < 0.001$). This was due to an increase in the percentage returned pre low-investment trial in the intervention condition ($\Delta M = -0.04$, 95% CI $[-0.05, -0.02]$, $t(317.20) = -5.19$, $p < .001$) but not in the control group ($\Delta M = 0.00$, 95% CI $[-0.01, 0.02]$, $t(318.64) = 0.21$, $p = .833$). This indicates that participants were sending back higher returns in the intervention condition but not in the control condition even before the defection trial.

The interventions focused on repairing cooperation in the wake of a transgressive action by the investor. To assess the effect of the intervention, we can therefore consider the percentage return after the "transgression" trials. We fit the same linear mixed effects model to percentage returns as the one for the full data while restricting the data to rounds 12 to 15 in the first game, and 13 to 15 in the second game. These trials represent all trials after the pre-programmed defection of the investor (low investment of 2) until the end of the game (round 15).

Figure 4.C shows the marginal means of trustee returns and their distribution by Game-number and Condition. There was a significant interaction between Game-number and Condition. Participants sent back lower returns between games in the control condition ($\Delta M = 0.07$, 95% CI $[0.04, 0.10]$, $t(325.03) = 5.17$, $p < .001$), but no such difference in the intervention condition ($\Delta M = -0.01$, 95% CI $[-0.04, 0.02]$, $t(313.90) = -0.83$, $p = .407$).

3.2 Emotion self-reports

A key aspect of effective interventions is the emotional change. We examined emotion self-reports through fitting a linear mixed effects model to both emotion dimensions (valence and arousal) for the intervention group to explore whether the intervention changed the emotional appraisal of the investment, with fixed

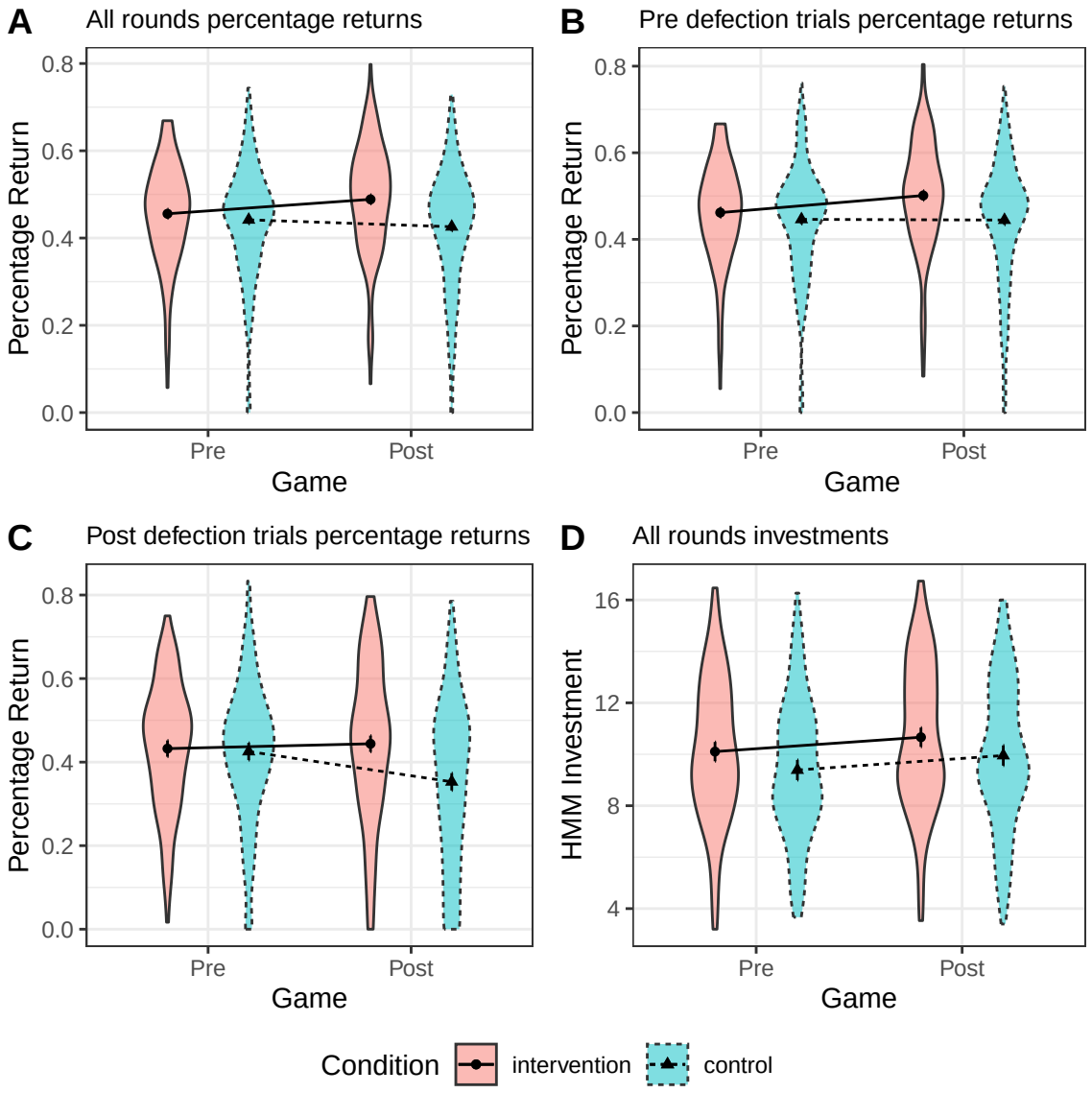

Figure 4: A: Marginal means and distributions of percentage trustee returns over all rounds, shown across participants by Game number and Condition. B: Marginal means and distributions of percentage trustee returns across all participants for pre-defection trials only, by Game number and Condition. C: Marginal means and distributions of percentage trustee returns across all participants for post-defection trials only, by Game number and Condition. D: Marginal means and distributions of investments over all rounds for HMM agents playing the role of the investor, by Game number and Condition

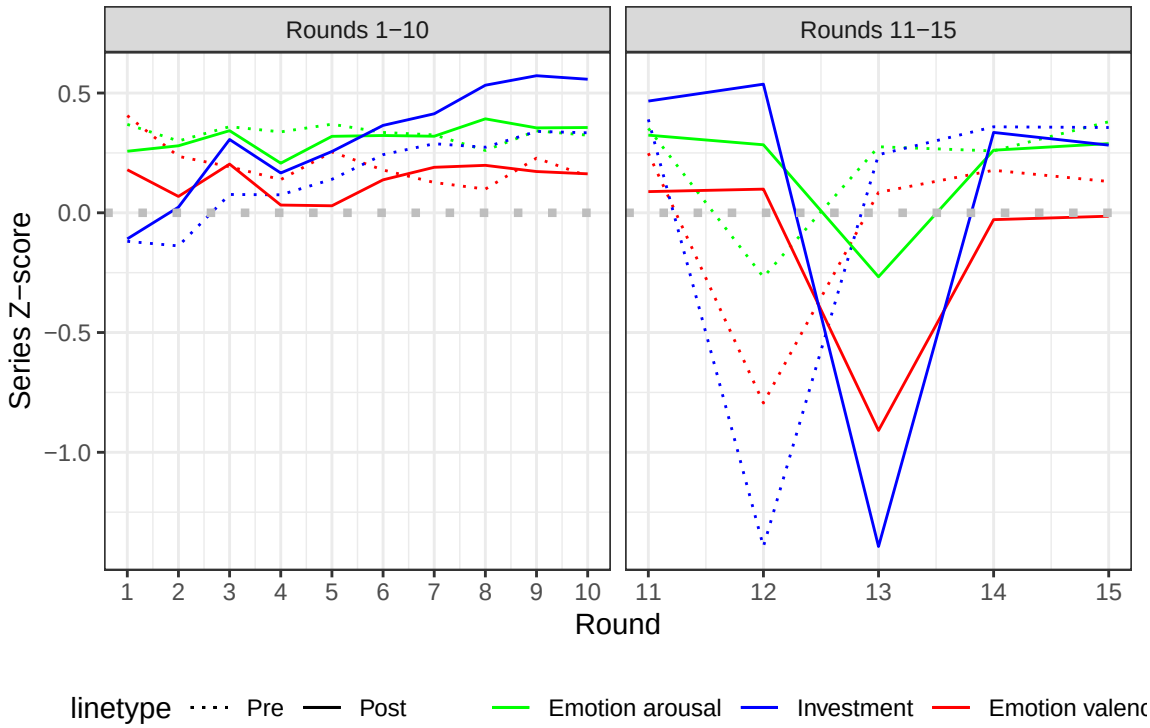

Figure 5: Self-reported emotion valence and arousal as well as investment z-scores for each round of the repeated Trust Game averaged across participants in the intervention condition only.

effects for Game-number (pre or post intervention) and Investment, as well as interaction between Investment and Game-number, with participant-wide random intercepts and random slopes for Game-number.

Higher investments were associated with more positive emotions ($F(1, 4664) = 2902, p < 0.001$), and higher arousal ($F(1, 4668.4) = 1919, p < 0.001$), while positive emotions declined between the two games ($\Delta M = 0.11, 95\%\ CI = [0.05, 0.18], t(\infty) = 3.72, p < .001$). However, there was no evidence that emotional reaction to the investment was lessened after the intervention. This means that participants in the intervention condition returned higher amounts post-intervention despite the emotional reaction to the lower investment being largely unchanged.

3.3 Player rating

Participants rated their artificial opponents differently between conditions (intervention vs control) and games (pre- vs post-intervention). The players ratings submitted at the end of each game showed a main effect of Game-number on all attributes, with participants rating the second player 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 $[-0.61, -0.10]$, $t(317) = -2.76$, $p = .006$). We also find a main effect of Condition as 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 Game-number and Condition on any of the attributes.

3.4 Transfer to the Prisoner's Dilemma Game

We next asked whether the coaxing behavior induced by the intervention generalized to the RPD. Across participants, we look at the rate at which the cooperative action was chosen in each round. Using a logistic mixed-effects model with Condition and Phase (before or after defection trial) as fixed effects and a random intercept for participants, we found no evidence for a different cooperation rate in the intervention condition compared to the control condition. We also found no difference in cooperation rates post defection trial between conditions.

3.5 Self report and debrief questionnaires

We find no interaction effect of the questionnaire scores with the Condition variable, nor a main effect of the questionnaire scores on participants returns in the RTG.

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

4 HMM analysis of participant returns

We analyzed participants' behavior differences in the intervention versus control conditions using hidden Markov models (HMM). Five models were used: "HMM-inv" assumed transition states depended solely on investment, ignoring game number and condition. "HMM-prepost" contrasted Pre and Post Intervention. "HMM-coax" contrasted post-intervention with pre-intervention and both control conditions. "HMM-ctrl" contrasted post-control with pre-control and both intervention conditions. "HMM-full" grouped pre-control and pre-intervention as one, with separate groups for post-control and post-intervention. Models were fitted using 2 to 7 states, selecting the lowest BIC. Generally, 5-7 state models best explained the data. A likelihood ratio test compared the models' goodness of fit, contrasting the complex HMM-full model with nested models equating behavior in certain stages and conditions.

In order to compare the goodness of fit of the various models, we test the relative likelihood of models using a likelihood ratio test. This procedure is useful to compare the most complex model (HMM-full, which allows for differences between pre-intervention and the two conditions post-intervention) to nested models which equate behaviour in some if the stages and conditions.

Using likelihood ratio tests, we find that the HMM-full model fits significantly better than HMM-ctrl ($\chi^2(40) = 138.82$, $p < .001$), HMM-coax ($\chi^2(40) = 265.73$, $p < .001$) and HMM-prepost ($\chi^2(40) = 125.67$, $p < .001$). This is consistent with a differentiated behavior of the trustees between all three groups: the post-intervention group, the post-control group and the pre-manipulation group.

Using the HMM-full model, we can retrieve participants' return distributions based on their latent states (Figures 7.A) and transition probabilities between these states (Figure 6). The states are ranked by mean return, with State 1 having the lowest mean return and State 5 the highest. A higher state number indicates a more pro-social policy. We focused on states related to cooperation's breakdown and repair. We compared the transition probabilities between states when the investment is low for post-control and post-intervention groups. Figure 6 suggests that the intervention group is more forgiving of low investments, as they are less likely to shift to an anti-social state when faced with defection compared to the control group.

To quantitatively explore the differences in transition probabilities between the control and intervention conditions, we can estimate from the model, using local decoding methods from the depmixS4 package (Visser and Speekenbrink 2021), the most likely posterior state of the trustee participants by round given the actions they have taken. Figure 7.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 7% of participants in the intervention condition

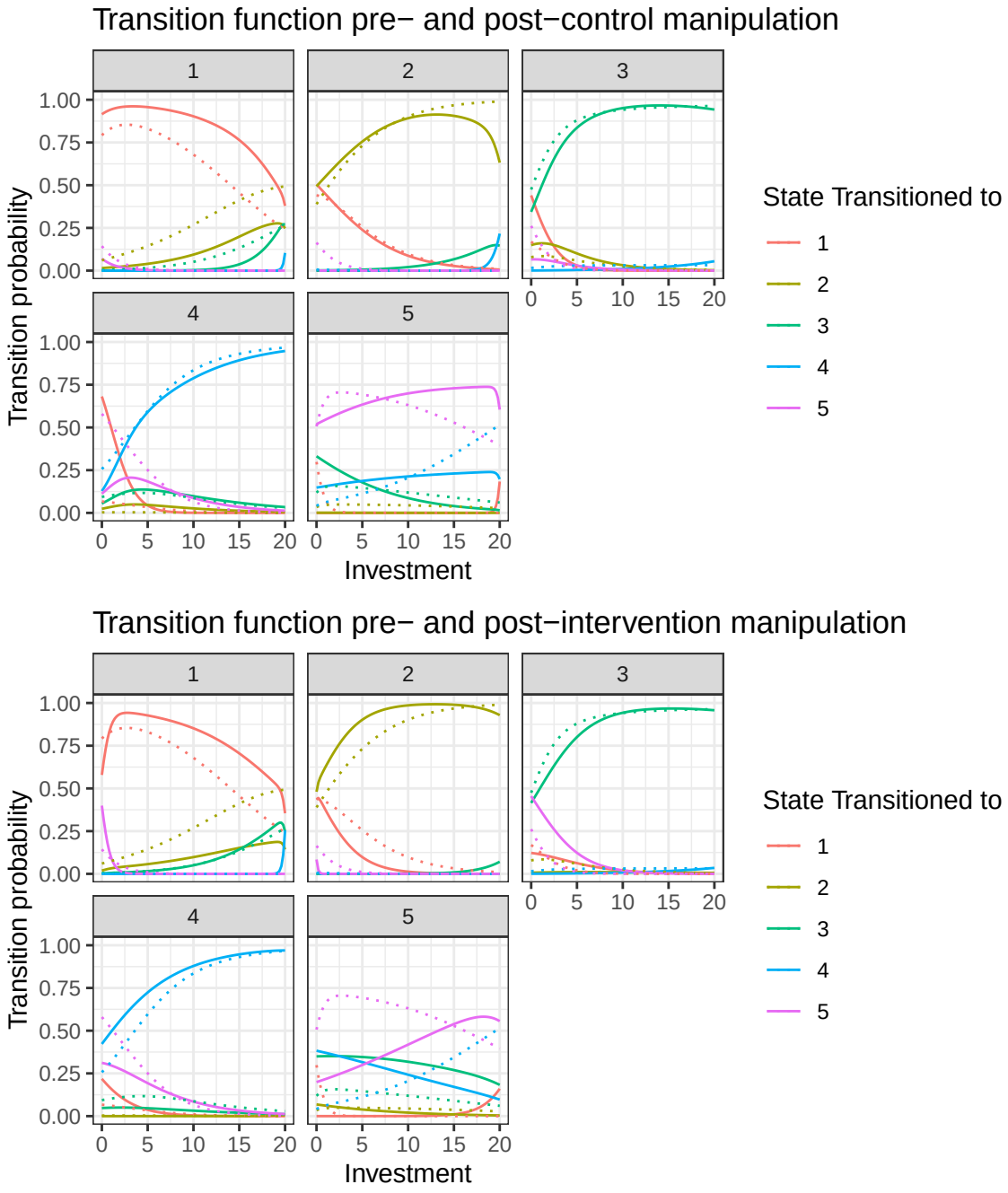

Figure 6: Transition function for the HMM-full trustee model. Each panel represents the state transitioned from, and each color the state transitioned to. Solid lines show estimated transition probabillities post-manipulation. Dotted lines show the same probabilities prior to the manipulation

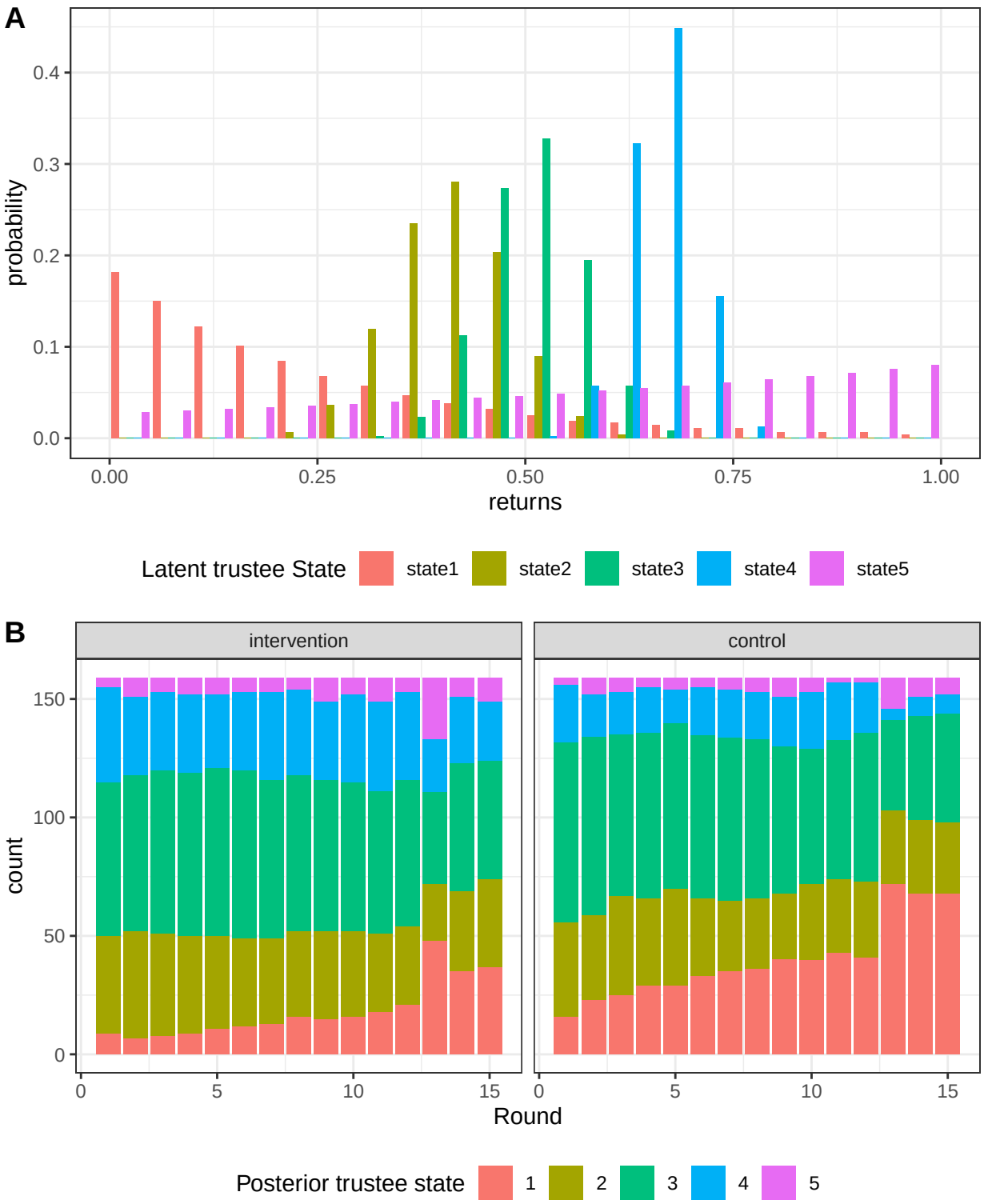

Figure 7: A: Distribution of participants' percentage return for each of the latent states in the 5 state HMM-full model. The latent states are ordered by the mean of the Gaussian that best fits the policy in that state, so higher numbered states are more pro-social. B: Distribution of posterior trustee 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.

compared to 24% in the control condition ($\chi^2(1) = 8.26, p < 0.01$). For the post-defection trial after the intervention (round 14), state 1 was the most likely state for only 22% of participants in the intervention condition compared to 43% in the control condition ($\chi^2(1) = 14.70, p < 0.001$).

The posteriors also suggest that a non-negligible proportion of participants in the intervention condition did not exhibit a behaviour consistent with the goal of the intervention as they were still best fit by low-return states post intervention. For instance, focusing on round 13 post defection 30.2% of those in the intervention condition were most likely to be in the least pro-social state 1. These differences can be seen as an indication of important heterogeneity in the effectiveness of the intervention.

5 Discussion

In this experiment, we made human participants face artificial computer agents endowed with the ability to transition between latent states and react to the participants' returns. The number of states, the policy in each state as well as the way these agents transitioned between states was based on estimating a hidden Markov model to behaviour from real human participants. On average, we saw the emergence of cooperative behavior with investment and returns in line with what is reported in human dyadic interaction in the repeated trust game (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011). The emergence of cooperative behavior and participant's uncertainty about whether they were facing human or artificial opponents, point to the potential of these agents to mimic human behavior in economic games whilst offering a higher degree of experimental control.

The intervention's aim was to articulate the effect of acting on impulse in case of a transgressive action from the investor in the form of a one-off low investment. When the agent defected as programmed, yielding a situation similar to the one presented in the intervention, the intervention led to higher percentage returns post defection compared to the control group, as intended. Participants emotional reactions to the received investment, we similar pre and post-intervention, so the higher returns were produced despite participants having the same emotional reaction. This might indicate that the intervention achieved its goal of encouraging participants to respond in a non-impulsive, considered way, overriding the emotional urge to retaliate.

What was also notable was that the intervention led to generally higher returns post-intervention, even before the defection trial. This could be for a number of reasons. One possibility is that participants simply learned that by returning more, the investment on the next trial would be higher. However, this explanation is unlikely, as we did not see a similar increase in percentage returns in the control condition. Another possibility is that, on average, the second player they faced invested more, which prompted positive reciprocity in the human trustees. As all players were programmed in the same way, any difference would be due to participant's actions. Further, participants rated both players similarly on relevant attributes of cooperativeness and trust. The higher returns are thus unlikely to be driven by different beliefs about the investor. A likely explanation for overall higher trustworthiness, as measured by higher percentage return post investment, is that participants, on average, interpreted the intervention message as an argument for more pro-social behavior, irrespective of the investor's actions.

It is also noteworthy that there were important individual differences in the percentage return changes post vs. pre intervention, which can be seen as a proxy for the intervention effectiveness. Some participants might not have been convinced by the intervention's message and decided to reduce their returns both pre and post defection in the second trust game, while others increased their returns in both phases. This raises important questions for the measurement of intervention effectiveness. Recent work has shed light on the important heterogeneity inherent in how disorders are categorised: This heterogeneity arises from the view that mental health problems should be viewed as complex systems, or interactions between neuro-computational processes and socio-environmental contexts evolving over time (Fried and Cramer 2017). This view was used to justify computational psychiatry's difficulty in establishing differential and reliable predictors of likely treatment responses (Hitchcock, Fried, and Frank 2022). But if a healthy group's reaction to a relatively explicit intervention is itself heterogeneous as we have shown in this experiment, then the issue of variable treatment responses might the result of 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.

In our case, judging by the inter-individual heterogeneity in responses, some people may not have been convinced that a coaxing behavior was a good way to establish long term cooperative outcomes, and their need to "punish" the other player for their low investment may have been more pertinent than what we suggested. This was also evident from the participants' replies to a question about whether they would change their behavior, just after seeing the intervention manipulation. An important avenue is to explore the role of emotion in decision making in such situations. We could aim to measure emotional reactions more accurately and explore whether specific emotions mediate the relationship between the investment received and the decision of what proportion to return. Measuring the emotions using the two axes of valence and arousal could be improved: Results indicate that these concepts may not have been well understood by participants since we would not expect to see low arousal after the pre-programmed defection of the investor.

The effect of this short intervention was not transferred to the Repeated Prisoner's Dilemma game. In this game, the rate at which the cooperative option was chosen was not significantly different between the control and intervention groups, both pre and post defection. Since the prisoner's dilemma is a very popular economic game, it is possible that participants had strong prior preferences towards which strategy they would adopt, irrespective of whether or not they received the intervention. As such, this paradigm might not be the best test case for knowledge transfer. For those that took on the intervention message and showed coaxing behavior in the second trust game, the fact that the investor still defected in the final rounds might have reinforced the idea that not reciprocating negative behavior is a losing strategy after all.

Overall, it is remarkable that such a short intervention, consisting of reading a short text detailing a non-impulsive reaction to low investments can lead to such differentiated behavior. In future studies, we aim to explore the effects of different cognitive interventions and improve the experimental design in multiple ways. First, the intervention could benefit from being more interactive medium, with visual inputs such as cartoons and videos, rather than pure text which can be cumbersome to read and lead to lower engagement. Second, we selected trustees from the general population, which might not suffer from the inability or unwillingness to repair relationships due to accidental breakdown of trust that characterises some mental health disorders such as BPD. As such, it would be interesting to contrast these results with findings from experiments involving trustees that are selected from patient populations known to suffer from difficulties in maintaining or repairing cooperative interactions. Third, as we explained above, the choice of the task to measure transfer of intervention learning could be made better by involving less popular paradigms. The high popularity of the Prisoner's Dilemma and the strategy of playing tit-for-tat may have resulted in a strong prior on which strategy to adopt in this game irrespective of the intervention. We believed that asking people about how they felt in the control condition might have affected how they behaved and might constitute an intervention in itself. However, being able to compare the differential impact of the intervention on the emotional interpretation of the opponent action between an intervention and control conditions could lead to insights on the mechanism through which the intervention affects the emotional reaction to the opponent's actions.

6 Conclusion

We explored the effect of a short cognitive intervention on the behavior of human trustees facing adaptive artificial agents endowed with multiple latent behavioral states. Each state defines different levels of a cooperative response with the agent able to transition between these states based on the behavior of the human opponent. Feedback from participants indicated that these agents were sometimes perceived as humans. Their strategy led to emergent cooperative behavior when playing the repeated trust game with human players. The intervention, promoting a less impulsive reaction to transgressive actions, led to coaxing behavior and less negative reciprocity when the investor sent a very low investment. It also led to more trustworthy behavior prior to the pre-programmed defection trial and to coaxing behavior after defection. Whilst this intervention effect varied between participants and generally was not transferred to a new game, an HMM analysis of participant's play post intervention showed differentiated patterns of transitions between

latent states, indicating a change in the effect of the opponent action on the probability of transitioning between latent mental states.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lieb, Klaus, Mary C Zanarini, Christian Schmahl, Marsha M Linehan, and Martin Bohus. 2004. "Borderline

Personality Disorder." *The Lancet* 364 (9432): 453–61. https://doi.org/10.1016/S0140-6736(04)16770-6.

Linehan, Marsha M. 1993. *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. Cognitive-Behavioral Treatment of Borderline Personality Disorder. New York, NY, US: Guilford Press.

———. 2015. *DBT® Skills Training Manual, 2nd Ed.* DBT® Skills Training Manual, 2nd Ed. New York, NY, US: Guilford Press.

Meng, Tianguang, and He Chen. 2014. "A Multilevel Analysis of Social Capital and Self-Rated Health: Evidence from China." *Health & Place* 27 (May): 38–44. https://doi.org/10.1016/j.healthplace.2014.01.009.

Rabiner, L. R., C. H. Lee, B. H. Juang, and J. G. Wilpon. 1989. "HMM Clustering for Connected Word Recognition." In *International Conference on Acoustics, Speech, and Signal Processing,* 405–408 vol.1. https://doi.org/10.1109/ICASSP.1989.266451.

Reiter, Andrea MF, Nadim AA Atiya, Isabel M Berwian, and Quentin JM Huys. 2021. "Neuro-Cognitive Processes as Mediators of Psychological Treatment Effects." *Current Opinion in Behavioral Sciences*, Computational cognitive neuroscience, 38 (April): 103–9. https://doi.org/10.1016/j.cobeha.2021.02.007.

Rudge, Susie, Janet Denise Feigenbaum, and Peter Fonagy. 2020. "Mechanisms of Change in Dialectical Behaviour Therapy and Cognitive Behaviour Therapy for Borderline Personality Disorder: A Critical Review of the Literature." *Journal of Mental Health* 29 (1): 92–102. https://doi.org/10.1080/09638237.2017.1322185.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2). https://doi.org/10.1214/aos/1176344136.

Singmann, Henrik, Ben Bolker, Jake Westfall, Frederik Aust, Mattan S. Ben-Shachar, Søren Højsgaard, John Fox, et al. 2022. "Afex: Analysis of Factorial Experiments."

Visser, Ingmar, and Maarten Speekenbrink. 2021. "depmixS4: Dependent Mixture Models - Hidden Markov Models of GLMs and Other Distributions in S4."