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

Determining the other's goals, intentions, and decision-making process is fundamental to successful social interaction. This inference is however fraught with uncertainty, as we cannot directly observe these features, and may need extensive prior knowledge of the person. Absent a history of interaction, one may decide to trust the other's goals and intentions are aligned with one's own. The challenge of trust is that it is by construction a risky endeavour. If we deem a person trustworthy, we might take the risk of investing in the relationship, hoping for a collaborative outcome. If this trust is misplaced, it can come at a high cost. 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 by reinforcing social support networks, maintaining community norms and facilitating collective action. Research into the determinants of psychopathology has linked trust-based constructs to the emergence of mental health disorders. Fonagy and Allison (2014) identified epistemic trust, 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). In this game, the "investor" decides how much of an endowment to send to the other player (the "trustee"). The amount that is sent is tripled and the trustee decides, in return, how much of the tripled amount to send back to the investor. To encourage the emergence and maintenance of trust in this setting,

some studies focused on modifying the game mechanism, for example by introducing a third-party who monitors the actions of the other players (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler and Haruvy 2017). Others chose to intervene directly on the participants. For example, Drążkowski, Kaczmarek, and Kashdan (2017) found that trust 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 do not address how to repair a breakdown of trust that might occur due to intentional or accidental non-cooperative actions by the players. 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 or a return of the trustee below the investment sent (Bendor, Kramer, and Stout 1991). Such ruptures of cooperation appear frequently when the trustee suffers from social disorders such as Borderline Personality Disorder [BPD; Lieb et al. (2004)]. In theses situations, BPD trustees fail to engage in trust repairing behaviours such as coaxing the investor by signalling trustworthiness via sending high returns. This failure may be linked to the misperception that their low returns in the game are not violating 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 social dysfunction in BPD patients. Whilst there is no proven pharmacological therapy for BPD, psychotherapies such as Mentalisation Based Therapy (Allen and Fonagy 2006) and Dialectical Behavior Therapy (Linehan 1993) have been shown to improve various dysfunctional behaviors in BPD patients, including those related to social interaction (Gunderson et al. 2018). However, response to these treatments is highly variable, and determining which interventions are effective for particular patients is challenging (Rudge, Feigenbaum, and Fonagy 2020; Arch et al. 2012). One promising approach is the study of how specific components of psychotherapeutic treatment affect quantitative markers of behaviour such as those inferred through computational models (Huys, Maia, and Frank 2016; Reiter et al. 2021). Combining the use of specific cognitive probes inspired by therapeutic interventions and computational models of behaviour may allow us to uncover the cognitive mechanisms targeted by common forms of psychotherapy. In turn, this may provide the basis for choosing effective psychotherapeutic interventions for given individuals.

In this study, we assess the effectiveness of a cognitive intervention aimed at repairing the potential breakdown of cooperation from a one-off low investment sent by an computerised investor. The intervention focuses on explaining the potential harm from reciprocating non-cooperative actions and suggesting a non-impulsive course of action to coax the investor back into cooperation. Participants are given the role of the trustee and are randomly assigned to either a control or intervention group. They play two instances of the repeated trust game with two different investors. After the first instance of the game, they either receive a cognitive intervention (intervention condition) or perform an unrelated task solving anagrams (control condition). In reality, participants face the same computerised agent in both instances, which is programmed to play according to a hidden Markov model fitted to real players' data. We explore whether the intervention has an effect on the behaviour of the human trustee and whether this effect transfers to a different game (repeated Prisoner's Dilemma) and 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 payment dependent on their performance. The experiment had a 2 (Condition: Intervention or Control) by 3 (Game: Trust-Game Pre Intervention, Trust-Game Post Intervention, Prisoner's Dilemma Post Intervention) design, with repeated measures on the second factor. Participants were randomly assigned to one of the two levels of the first factor.

2.2 Tasks and Measures

2.2.1 Repeated Trust Game

Participants played a 15-round repeated trust game (Joyce, Dickhaut, and McCabe 1995) in the trustee role against a computer-programmed investor. On each round the investor was endowed with 20 units and decided how much of that endowment to invest. This investment is tripled and the trustee then decides how to split this amount between them and the investor. If the trustee returns more than one third of the amount, the investor makes a gain. The Nash equilibrium for a single-round version is for the investor to send nothing. In the repeated version, rewards for both players are maximised if they build trust and share the benefits of the the investment multiplied by three. An investor who has been rewarded for taking the risk of sending an investment is more likely to invest more on future rounds. An investor obtaining a low return on their investment may choose to reduce future investment and thereby reduce both players' gains.

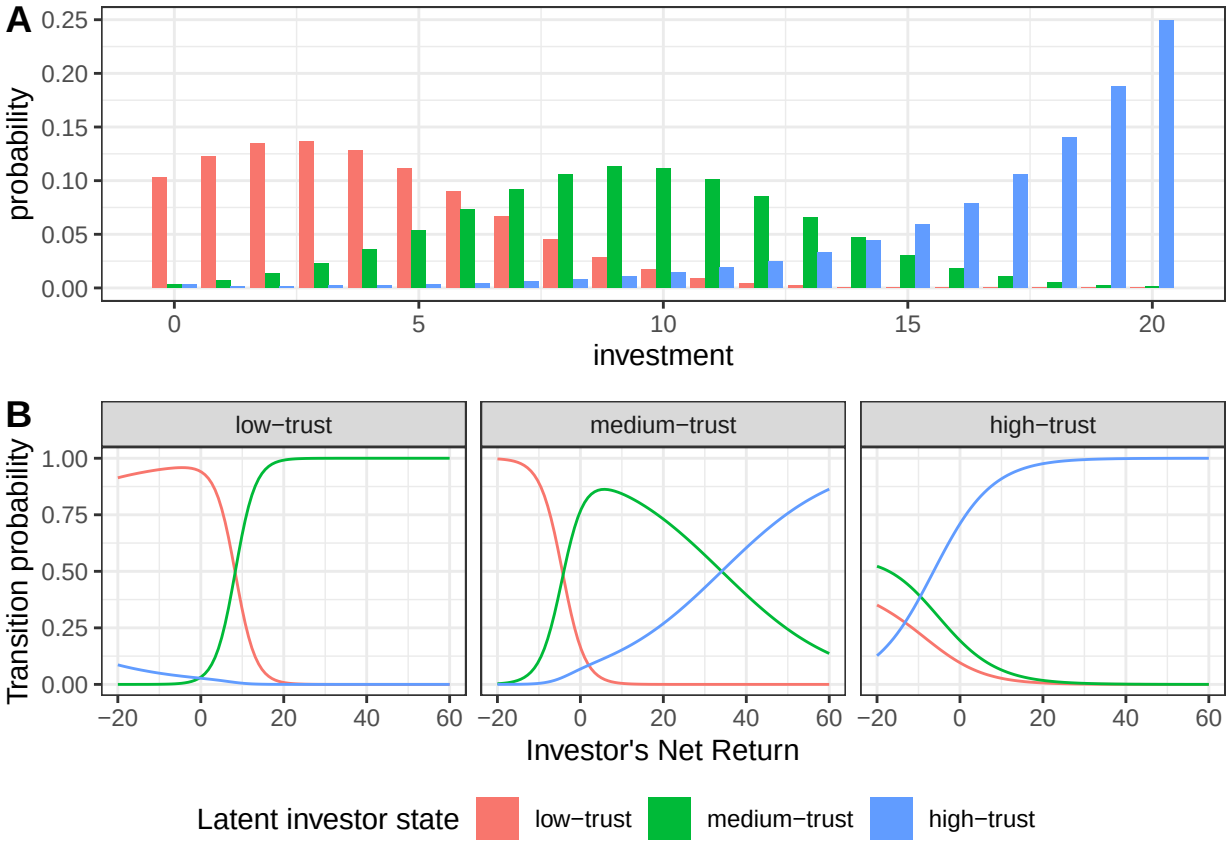

Figure 1: A: Distribution of investments by the artificial investor agent conditional on its latent state as estimated by a three state hidden Markov model fitted to human dyadic play dataset. B: Transition function for the HMM investor conditional on its current state and the net return from the previous round of play.

The strategy of the computerised investor was modelled on behaviour of human investors in a 10-round RTG.. Using this data, we estimated a hidden Markov model (HMM) on investors' behaviour with three latent states, reflecting "low-trust", "medium-trust", or "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 (Figure 1.A). Over rounds, the investor can move between states, and the probability of these transitions was modelled as a function of the net return in the previous round (see Figure 1.B).

In order to probe efforts to repair trust, the computerised agent was programmed to provide a low return on round 12 pre-intervention or 13 post-intervention. On all other rounds, the agent's actions were derived from the HMM (disregarding the participant's response on round 12).

2.2.2 Repeated Prisoner's Dillema

Participants played 7 rounds of a repeated Prisoner's Dilemma. In each round, participants could choose a cooperative action with a a reward of 5 if the player also cooperated and the a reward of 1 if not, or a non-cooperative action that would yield a 7 points if the other person chooses the cooperative action and 2 points if not. The Nash equilibrium for a single-round version is to choose the non-cooperative action.

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

2.3 Intervention

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.

2.4 Procedure

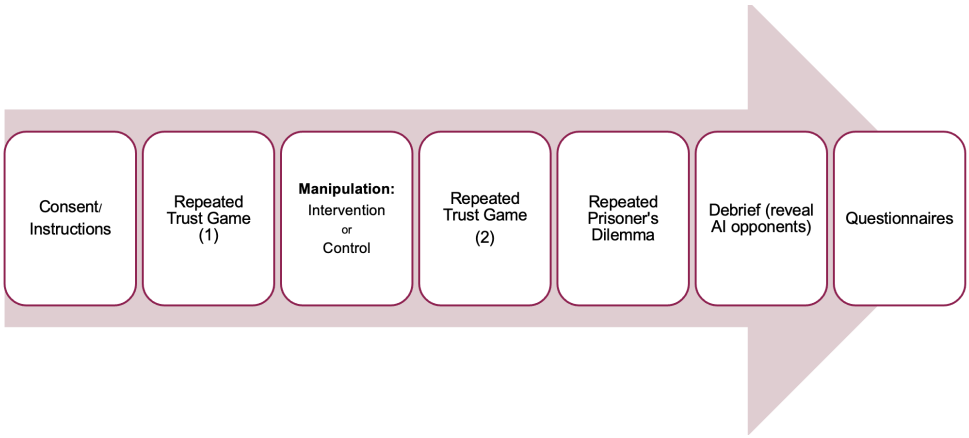

Figure 2: Experiment overview. After playing 15 rounds of a repeated trust game (RTG) as trustee, participants were randomized to either receive the control or active intervention. They then played the second set of 15 rounds of RTG (again as trustee) to examine intervention effects, and 7 rounds of a repeated Prisoner's Dilemma to examine generalization of intervention effects. Finally, participants answered questionnaires and were debriefed.

At the start of the experiment (Figure 2), participants provided informed consent and were instructed the study would consist of three phases in which they would face a different other player. Participants were told their goal was to maximise the number of points in all phases. They were not told the number of rounds of each phase. Phase one was a 15 round repeated Trust Game (RTG) in which participants took the role of trustee, facing the same investor over all 15 rounds. On each round, after being informed about the amount sent by the investor, participants were asked to provide feedback, with participants in the Intervention condition rating their feeling in terms of valence (from negative to positive) and arousal (from low to high), and participants in the Control condition rating the investment in terms of speed (from slow to fast) and magnitude (from low to high). These ratings were made by indicating a point on a two-dimensional grid.

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 their interaction partner (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 (RFQ8), emotion regulation (DERS), and BPD traits (PAI-BOR). They were then asked about the strategy in the games, as well as whether they thought the other players were human or computer agents. They were then debriefed and thanked for their participation.

2.5 Statistical analysis

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) for R. The transition between latent states is assumed to depend on the investment received and a dummy variable to characterise the group that the participant belongs to. Details on how the models are constructed can be found in the supplement. We fit models with different numbers of hidden states, and use the Bayesian Information Criterion (Schwarz 1978) to select the best model.

3 Behavioural results

Average investments and returns prior to the "defection round" (Figure 3) were within the range of reported investments (40-60% of endowment) and returns (35-50% of total yield) in the literature (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011).

Mixed-effects analysis shows a significant main effect of Condition (intervention vs. control), $F(1, 317.20) = 9.53$, $p = .002$, with higher percentage returns in the intervention compared to the control condition. Importantly, we also find an interaction between Condition and Game-number (pre- vs. post-intervention), $F(1, 318.30) = 26.91$, $p < .001$. Post-hoc tests show an increase in the percentage returned in the intervention condition, pre - post $\Delta M = -0.03$, 95% CI $[-0.05, -0.02]$, $t(316.54) = -4.85$, $p < .001$, but a decrease in the control condition, $\Delta M = 0.02$, 95% CI $[0.00, 0.03]$, $t(319.69) = 2.34$, $p = .020$ (see Figure 4.A). This indicates the intervention was effective in increasing cooperative behaviour. There was also a significant main effect of Investment, $F(1, 9, 208.68) = 373.23$, $p < .001$, such that higher investments were associated with higher percentage returns. An Investment by Condition interaction, $F(1, 9, 208.68) = 45.35$, $p < .001$, indicates the positive effect of investment on percentage returns was greater in the control than intervention condition. There was also an Investment by Game-number interaction, $F(1, 8, 990.38) = 4.31$, $p = .038$. Finally, we find a three way interaction between Game-number, Condition and Investment, $F(1, 8, 990.38) = 24.56$, $p < .001$, showing that the differentiated effect of the investment on the proportion returned by condition is itself moderated by the Game-number.

We next analysed HMM agent investments via 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. This shows a main effect of Condition, $F(1, 317) = 8.72$, $p = .003$, and Game-number, $F(1, 317) = 8.32$, $p = .004$. As can

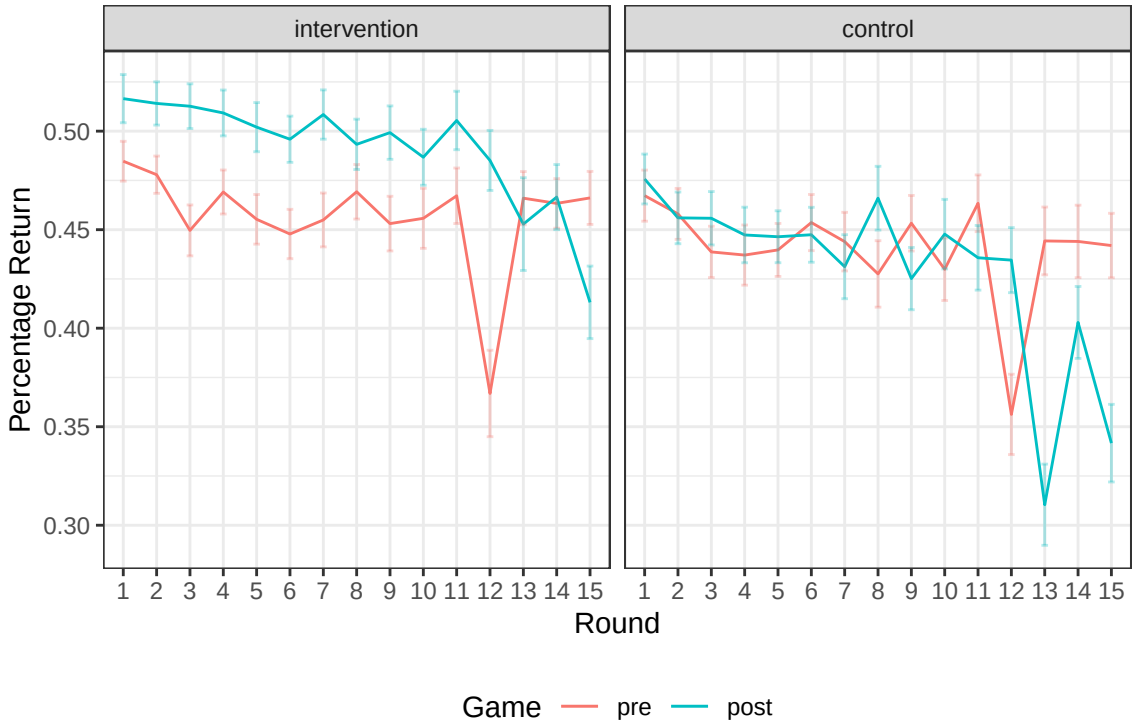

Figure 3: Average and standard errors of the trustee's return as a percentage of the multiplied investment received for each round and for both conditions. The red line shows the returns pre-manipulation and the blue line post-manipulation.We note a different reaction to the pre-programmed one-off low 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

be seen in Figure 4.B, investment was higher in the intervention compared to the control condition across games, and higher in the second game compared to first across conditions.

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 the second 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 group compared to the control group were not purely driven by reciprocity towards higher investments, since we find a Condition by Game-number interaction whilst controlling for investment in the model, and a reduced effect of Investment in the Intervention condition.

We next analysed returns separately for rounds prior to the pre-programmed defection (rounds 1 to 11 pre-intervention and 1 to 12 post-intervention) and rounds after (rounds 12 to 15 pre-intervention and rounds 13 to 15 post-intervention). Applying the same mixed effects model as before to returns before defection, largely replicates results over all trials. Full results are provided in the supplement. 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 two phases of the RTG, $\Delta M = 0.00$, 95% CI $[-0.01, 0.02]$, $t(318.64) = 0.21$, $p = .833$.

The intervention focused on repairing cooperation after a "transgressive action" by the investor. To assess whether the intervention succeeded, we next considered the returns by participants after the pre-programmed transgression (Figure 4.D). A mixed-model analysis shows similar results to those of the model applied to all returns (see SI for full results). The was again a significant interaction between Condition and Game-number and Condition. Participants in the control condition decreased their post-transgression returns from the first to the second phase of the RTG, $\Delta M = 0.07$, 95% CI $[0.04, 0.10]$, $t(325.03) = 5.17$, $p < .001$. There was no change for participants in the intervention condition, $\Delta M = -0.01$, 95% CI $[-0.04, 0.02]$, $t(313.90) = -0.83$, $p = .407$. So, whilst the intervention did not appear to move participants to increase their returns after a transgression by the investor, it may have countered the decrease in returns showed by participants in the control condition.

3.1 Emotion self-reports

Participants in the intervention condition rated their emotion on valence (negative to positive) and arousal (low to high) after each investment. To assess the impact of the intervention on these emotional reactions, we used linear mixed-effects models (one for valence, and one for arousal) with fixed 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. This showed that higher investments were associated with more positive emotions, $F(1, 3, 448.17) = 2,108.08$, $p < .001$, and higher arousal, $F(1, 3, 453.24) = 1,505.03$, $p < .001$. 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-number on valence, $F(1, 3, 419.70) = 1.49$, $p = .222$, or arousal, `papaja::apa_print(mod_emo_y)$full_result$gameNum_f_scaleinvestment`. This indicates that participants in the intervention condition returned higher amounts post-intervention despite their emotional reaction to investments remaining largely the same.

3.2 Evaluation of the investor

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 Game-number and Condition on any of the attributes.

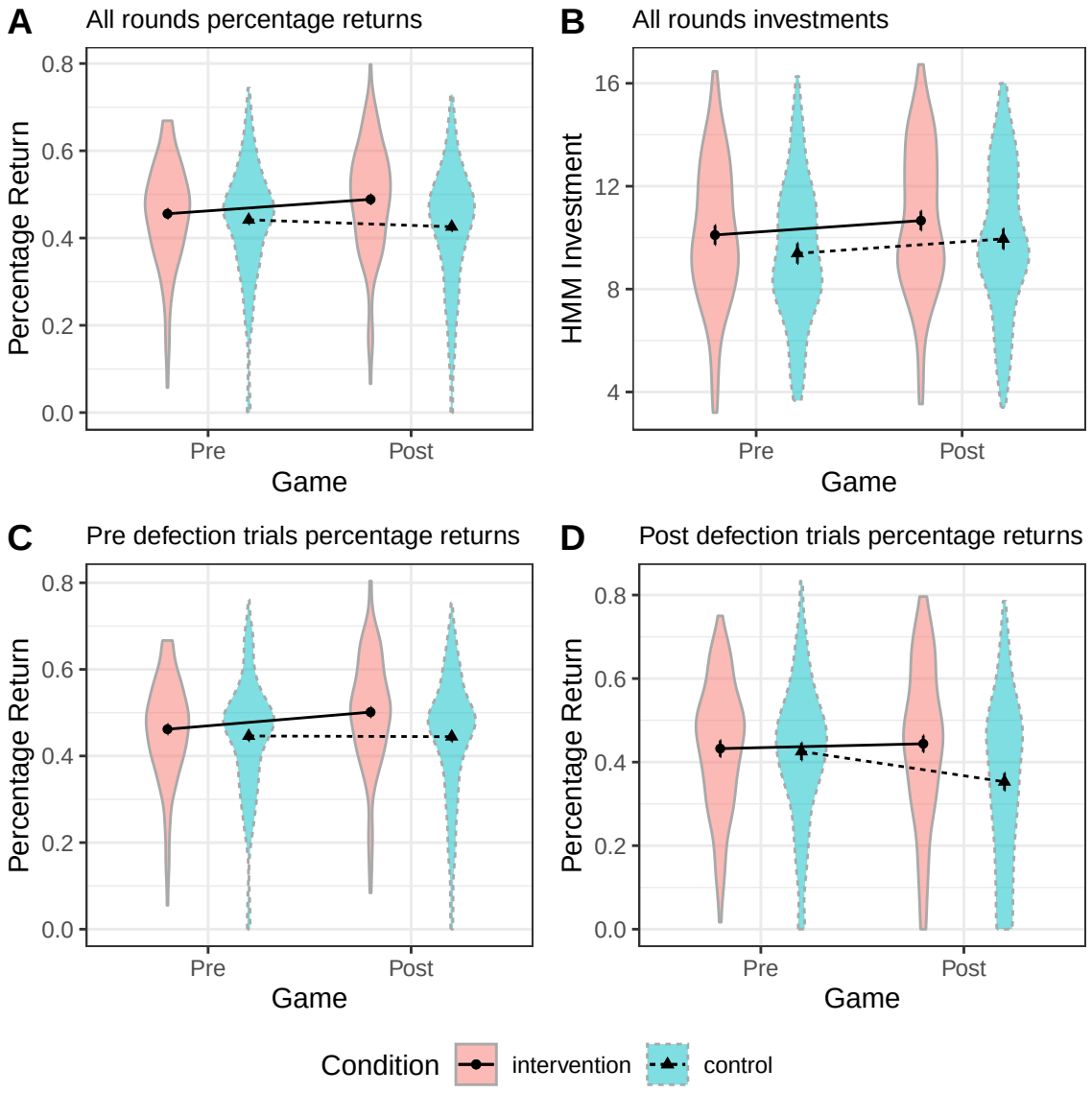

Figure 4: A: Marginal means and distributions of percentage trustee returns over all rounds, shown across participants by Game number and Condition. B: Marginal means and distributions of investments over all rounds for HMM investors, by Game number and Condition. C: Marginal means and distributions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

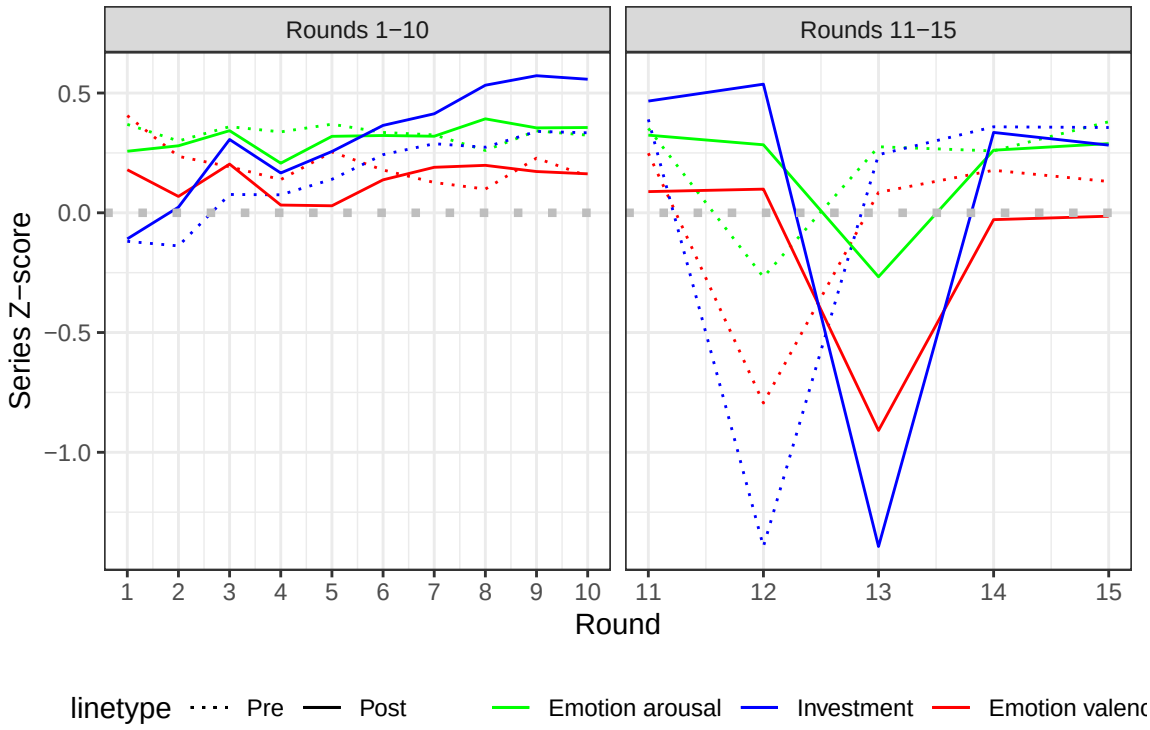

Figure 5: Self-reported emotion valence and arousal as well as investment z-scores for each round of the repeated Trust Game averaged across participants in the intervention condition only.

3.3 Transfer to the repeated Prisoner's Dilemma game

We next asked whether the intervention had any discernible effect on participants' behaviour 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 a different cooperation rate in the intervention condition compared to the control condition, `rpapaja::apa_print(ipd_mod)`$full_result$`condition_f`, or a different response to defection between the conditions, $\chi^2(1) = 0.23$, $p = .635$.

3.4 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 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

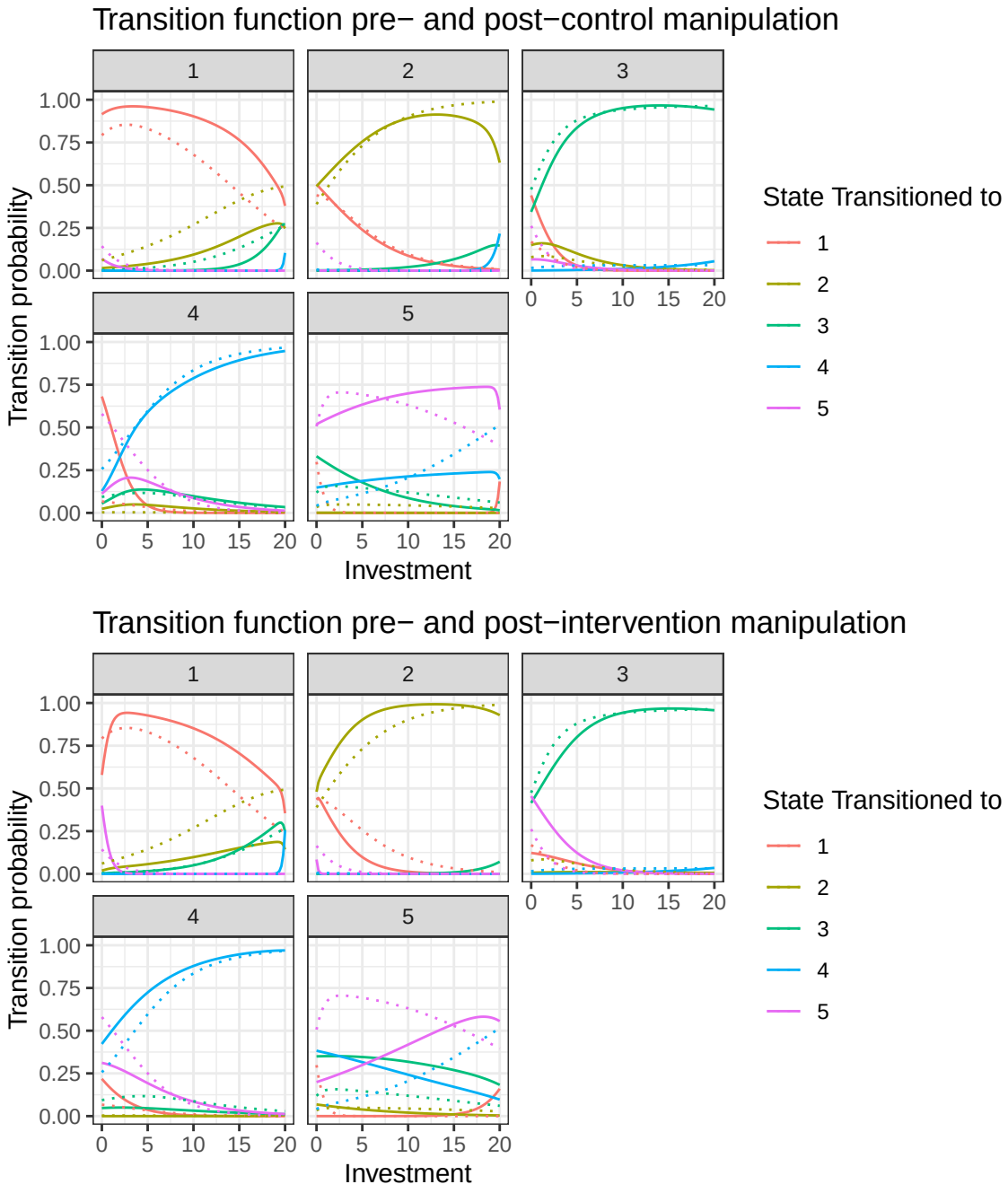

Figure 6: Transition function for the HMM-full trustee model. Each panel represents the state transitioned from, and each color the state transitioned to. Solid lines show estimated transition probabillities post-manipulation. Dotted lines show the same probabilities prior to the manipulation

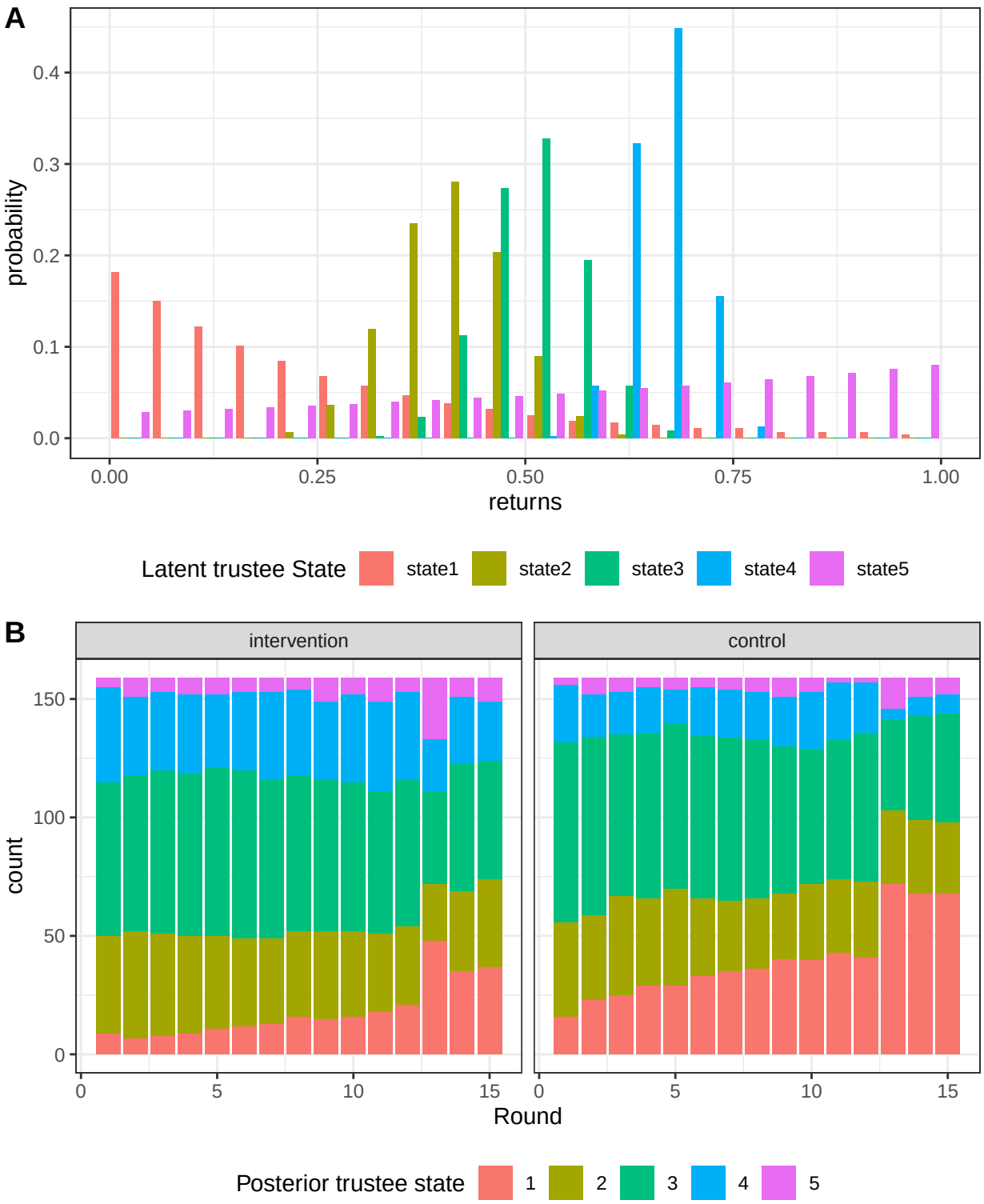

Figure 7: A: Distribution of participants' percentage return for each of the latent states in the 5 state HMM-full model. The latent states are ordered by the mean of the Gaussian that best fits the policy in that state, so higher numbered states are more pro-social. B: Distribution of posterior trustee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

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.

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

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

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

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

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.

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."