

Repairing cooperation through a cognitive intervention in the repeated Trust Game

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 cognitive intervention or not. 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 saw the emergence of cooperative behavior in these games, consistent with what is expected when humans interact. This substantiates the use of these AI agents as credible human opponents whilst affording a high degree of experimental control. 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. Rousseau et al. (1998) define trust as a “psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another”. As such, whilst trusting and being able to signal trustworthiness are necessary conditions for building and maintaining important social bonds, 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 may be foregone, and any substitution for trust, such as excessive suspicion, over-reliance on self and other maladaptive strategies are likely to add burden and costs to a person.

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 as an important function of early attachment relationships. Epistemic trust can be defined as the trust in the authenticity and personal relevance of interpersonally transmitted knowledge. It enables individuals to learn from their social environment in dynamic social and cultural contexts. Importantly, it allows the individual receiving social information to let go of their natural self-protective vigilance, dubbed hyper-vigilance. Fonagy and Campbell (2017) identified hyper-vigilance, or lack of “epistemic trust” which may partially be rooted in adverse childhood experiences, as a key risk factor for the emergence of multiple mental health disorders such as Borderline Personality Disorder and Anti-Social Personality Disorder.

A popular paradigm in the study of the emergence and maintenance of trust is the repeated trust game (Joyce, Dickhaut, and McCabe 1995). In this game, one player takes the role of the “investor” and is provided with a fixed endowment at the start of each trial. They get to decide how much of their endowment to invest with the other player taking the role of the “trustee”. The amount that is sent is tripled and the trustee gets to decide, in return, how much of the tripled amount to send back to the investor. If the investor sends a non-zero amount, they express trust in the other player for that round, as the amount they will receive back is uncertain. Assuming the trustee sends back more than the initial investment, they signal trustworthiness to the investor and both players make gains. This scenario would constitute a cooperative exchange and would be mutually beneficial. However, cooperation in this setting is prone to be broken through intentional or simply misinterpreted actions on behalf of either player (Bendor, Kramer, and Stout 1991). In order to repair a damaged cooperative equilibrium, players need to infer that their behaviour has violated social norms and offer amends through generous actions at potentially a high cost to them. King-Casas et al. (2008) links the breakdown of cooperation in trustees with Borderline Personality Disorder to the absence or reduction of activation in brain regions associated with the perception of norm violations.

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 an intervention 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. In the context of social dilemmas, there have been many attempts at fostering cooperative outcomes through either mechanism design, environment modifications or other.

Mechanism design, pioneered by Vickrey (1961), is a field that studies incentive alignment and looks for ways to promote social welfare or revenue maximization, despite self-interest of the individual actors. In the context of social dilemmas, key mechanisms were explored with good effect to foster and maintain the cooperative outcome. Axelrod (1986) studied the emergence and stability of behavioral norms to regulate non-cooperative actions in social settings. He identifies a behavioral norm as a dominant behavioural strategy which is often punished by others when not adhered to. A meta-norm is the propensity to incur a cost in order to enforce a norm. Axelrod (1986) showed that when playing a social dilemma game, individuals have a strong incentive to enforce punishment of defectors lest they are in turn punished by others. This led to a decline of defection. Thus, meta-norms can be seen as a mechanism to promote and sustain cooperation in a population. In the context of the trust game, Charness, Cobo-Reyes, and Jiménez (2008) explored the effect a third party monitor can have on the amounts sent and received. This third-party’s payoff is unaffected by the decisions made by the investor and trustee. However, the study allowed the third party to punish overly selfish trustees or reward Investors making a loss on the interaction. They found that the actions of both players were materially more cooperative in the presence of this third party. Fiedler and Haruvy (2017) found that the introduction of a third party that monitors the investments and returns of the players led to more cooperative behavior, even when the third party had no ability to reward or punish the players.

Rather than try to modify the mechanism of the game, other approaches focused on intervening directly on the participants. Drażkowski, Kaczmarek, and Kashdan (2017) investigated the effect of expressing gratitude on later behaviour in the trust game. The gratitude intervention consisted of thinking about and writing down 5 things that the participants were grateful for in their life. The exercise was repeated over three sessions. Those completing the gratitude intervention reported more positive emotions and this mediated an increase in their investment in one-shot trust games. Burnham, McCabe, and Smith (2000) primed participants by introducing the investor as either a “friend” or “foe”. They found that this priming produces significant differences in both trust (measured by the proportion of endowment sent) and trustworthiness (measured by the proportion returned) with participants in the “friend” treatment exhibiting over twice as much trustworthiness than those in the “foe” condition.

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 these 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 (MBT) and Dialectical Behavior Therapy (DBT), 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). 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, as well as evaluate whether knowledge about these different mechanisms can improve the 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 the investor. The intervention focuses on explaining the potential harm from reciprocating non-cooperative actions and suggesting a non-impulsive course of action to coax the investor back into cooperation. It explicitly articulates the advantages of adopting a coaxing behavior in this setting, and as such, is directly relevant to the task at hand and is very clear about the desired behavior without being overly directive. The idea being that if participants show no response to this type of direct intervention in this setting, then other forms of intervention that may target general processes such as better emotional regulation or better mentalising are then even less likely to show a benefit. As such this intervention can be seen as a litmus test for the success of more “general” forms of cognitive interventions. Moreover, in the event of the intervention’s failure, analysis of the resistance to an explicit intervention itself might be worthy of further investigation to understand the mechanism behind the lack of success.

In this experiment, 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. 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

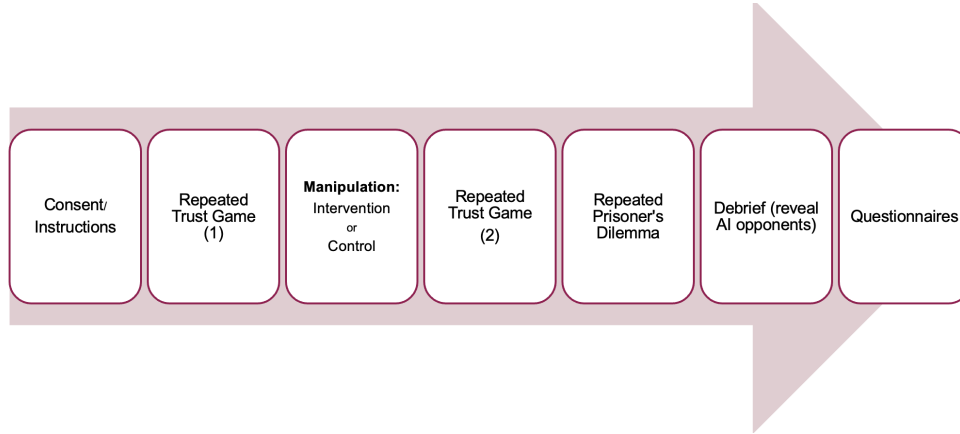


Figure 1: Overview of the timeline of the experiment detailing the various phases participants go through.

Figure 1 shows the timeline of the experiment. After reviewing an information sheet about the experiment, participants were informed that there would be three phases to go through, with them facing a different opponent in each phase. They would face the same player throughout the first phase, consisting of the “Investment Game” with several rounds. This was the moniker used for the repeated trust game, and a total of 15 rounds were played.

After reading instructions on how to play the trust game, players were told that they would be assigned the role of the trustee. A comprehension quiz on the trust game followed the instructions. The first trust game then ensued: Each round started with an investment stage where the investor decided how much to send to the trustee and that choice was revealed. Immediately afterwards, participants were presented with a two-dimensional field to indicate their emotional state on two dimensions: valence, from unpleasant to pleasant (on the horizontal axis) and arousal, from low arousal to high arousal (on the vertical axis). After selecting a point on the field, participants were asked to decide how much to send back to the investor using a slider ranging from sending back nothing to sending back the whole (tripled) investment. Following that choice, an outcome page summarised both players decisions and showed the total pay-off for each player and the round then concluded. At the end of the game, participants were asked to rate the player they were facing on 4 attributes using a scale from 1 to 10: Trustworthiness, friendliness, cooperativeness and selfishness.

After being randomly assigned to either a control or intervention condition, participants were informed they would be paired with a different player in a second phase of the experiment. This second phase was similar to the first phase in terms of the game played (RTG) and the number of rounds. Finally, players were told they would be paired with a new player in the third phase of the experiment. This third phase consisted of 7 rounds of the prisoner’s dilemma game, facing the new player. Throughout all games, players were explicitly told to aim to maximise the number of points as their bonus would depend on the score they accumulated throughout the experiment. The total number of rounds was not communicated to the participants in any of the games played.

After the three phases were completed, participants were asked to complete a series of questionnaires at the end of the experiment. These included: the PAI-BOR measuring Borderline traits (Morey 1991), the DERS measuring Emotion Regulation ability (Gratz and Roemer 2004) and the RFQ8 for mentalising abilities (Fonagy et al. 2016). Finally, participants were asked a series of questions around how they played each phase. First, participants were asked whether they thought they played differently in the second compared to the first phase of the RTG. Second, participants had to select whether they thought the opponents they faced were human or computer agents. Finally, we revealed to the participants that they faced the same computer agent throughout the RTG. We explained that any change in their perception of the opponent or any change in how the opponent played is due to their own change in behavior as the agent was simply

reacting to their actions. Finally, we asked participants to reveal whether this experiment has taught them anything about how they should behave in social situations.

Decision Making on Impulse

When making decisions about how to interact with others, we have found that people may sometimes act on impulses, and this might not serve them well in achieving their goals from the interaction. As such, it is important to slow down, check-in with ourselves and ask whether the urge to act a certain way comes from an impulsive reaction to the events. If it is, then we can check whether this urge is leading us towards sound decisions, and decide to act differently if it isn't.

For instance, in the situation exhibited here, the urge might be to send back very low returns to the investor, to express discontent. However, this is unlikely to make the investor trust us more going forward. It would be more helpful to signal to the investor that we are trustworthy to convince them to trust us with more of their money in future rounds. One way of doing that is to be generous and send them back high returns even when they have sent you low investments.

In the next part, there will be an open ended question. Please take time to reflect on the question before writing down your answers.

Figure 2: Screenshot of the main slide in the intervention condition

2.3 Interventions

After a total of 15 rounds of the trust game, players were given either an “intervention” or a “control” manipulation. The intervention consisted of presenting a hypothetical scenario in which they were playing the repeated trust game and the investor would send a low investment in a new round after having previously sent higher amounts. Participants were then asked how they would react in this situation and what sort of return (low or high) they were thinking of sending back. The players then were presented with an educational slide about the benefits of not resorting to impulsive decisions such as punishment when they feel they have been wronged. In this text, players were told that punishment can create a negative feedback loop where the other player might trust them even less. An alternative action was suggested, whereby players would respond kindly to such a transgression in the hope of gaining trust from the investor. The full text of the intervention slide is presented in Figure 2. Afterwards, participants were asked whether they would send a low or high return in the same hypothetical scenario now that they have read the information on the slide. Players were asked to justify their answer. For each question during this intervention, participants had to wait for a fixed duration of 20 seconds before being able to write their answers, and they were prevented from proceeding before that time was up. This choice was made to allow participants to engage with the questions, think about their answers and provide meaningful feedback.

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 Measures

2.4.1 Repeated Trust Game

Participants played two iterations of the standard version of the repeated trust game (Joyce, Dickhaut, and McCabe 1995). The game is played in dyads, with one player assigned the role of the “investor” and the other player that of the “trustee”. In the variant of the game we chose, the investor is endowed with 20 units at the


start of each round. They need to decide how much of that endowment they want to invest with the trustee. The investment is then multiplied by a factor of 3 and sent to the trustee. As shown in Figure 3, the trustee in turn needs to decide how much of the (multiplied) investment they want to send back to the investor. If they send back more than a third of what they received, then both the investor and the trustee make a gain. Otherwise, the trustee would benefit but the investor would lose on their investment. If the interaction lasts for only one round, the Nash equilibrium of this game is for the investor to send nothing, as there is no incentive for the trustee to send back any return. In the repeated version, to maximise their rewards over time, both players need to build trust so that they can share the benefits of the bigger pie (the investment multiplied by three). If during the course of the interaction, the investor has been rewarded for taking the risk of sending an investment, then they would be more likely to invest more, and the multiplied investment would give both a bigger pie to share. If however the investor felt that they were not getting a return on their investment, then they would refrain from investing and neither would gain from the interaction.

You



Round 1

Other Player



Reminder

The other player (🟡) received an endowment of 20 from the experimenter.

Of this endowment, the other player (🟡) **gave you 10**.

This was multiplied by 3 by the experimenter, hence **you now have 30**.

Please select how much you want to send back to the other player (🟡):

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

16

Send back: 16
Keep: 14

Submit

Total payoff: 0

Figure 3: Screenshot of the repeated Trust Game. The game is played in dyads, with one player assigned the role of the investor and the other player that of the trustee. The investor is endowed with 20 units at the start of each round. They need to decide how much of that endowment they want to invest with the trustee. The investment is then multiplied by a factor of 3 and sent to the trustee. The trustee in turn needs to decide how much of the multiplied investment they want to send back to the investor. Shown here is the stage at which the trustee makes a decision of how much to send back the investor



2.4.2 Repeated Prisoner's Dilemma

The other game the participants played is the repeated Prisoner's Dilemma (RPD). Over multiple rounds, participants could choose one of two actions: A cooperative action that would yield a high pay-off if the other person also cooperated, and the lowest possible pay-off if they did not cooperate. Or a non-cooperative option that would yield a high pay-off if the other person chooses the cooperative action and a lower pay-off if they also defect. Figure 4 shows the payoff of each combination of actions as presented to the participants.


		Their decision	
		A	B
Your decision	A	You receive: 2 points	You receive: 7 points
		They receive: 2 points	They receive: 1 points
	B	You receive: 1 points	You receive: 5 points
		They receive: 7 points	They receive: 5 points


Figure 4: Screenshot of the Repeated Prisoner's Dilemma game. Over multiple rounds, participants could choose one of two actions: A cooperative action that would yield a high pay-off if the other person also cooperated, and the lowest possible pay-off if they did not cooperate. Or a non-cooperative option that would yield a high pay-off if the other person chooses the cooperative action and a lower pay-off if they also defect. Shown here is the table explaining the payoffs of each combination of actions the participant and their opponent choose

2.4.3 Feedback post investment in the repeated Trust Game

You	Round 1	Other Player
		

Reminder

The other player () received an endowment of 20 from the experimenter.

Of this endowment, the other player () **gave you 10**.

This was multiplied by 3 by the experimenter, hence **you now have 30**.

The Emotion Grid

Please indicate how **you are feeling** right now by clicking the appropriate point on the grid below:

High Arousal

Unpleasant		Pleasant

Low Arousal

Figure 5: Screenshot of the two axis grid in the Coaxing condition where participants were asked to report the valence and arousal of their emotional response

Whilst playing the RTG, participants were asked to provide some feedback, in each round of the game, after seeing the amount sent by the investor. This feedback took the form of the answer to a question using a two dimensional field. What the axes of the field represented differed by condition. In the intervention condition, participants were asked to rate their emotional state with regard to the other player's choice of investment, with the x-axis representing the valence of the emotion they felt (positive or negative) and the Y-axis representing the emotional "arousal" they experienced (low or high). In the control condition, players were asked to rate attributes of the investment that were not related to their emotion, with the X-axis rating of how fast the investor is (slow or fast) and the Y-axis the magnitude of the investment (low or high). Figure 5 shows a screenshot of the field in the intervention condition where participants were asked about their own emotional response to the investment. At the beginning of each game, participants were provided with detailed explanations of the meaning of the two axes as well as the opportunity to provide a baseline emotional state through using the field prior to the start of the game.

2.4.4 Design of the AI agents in the RTG and RPD

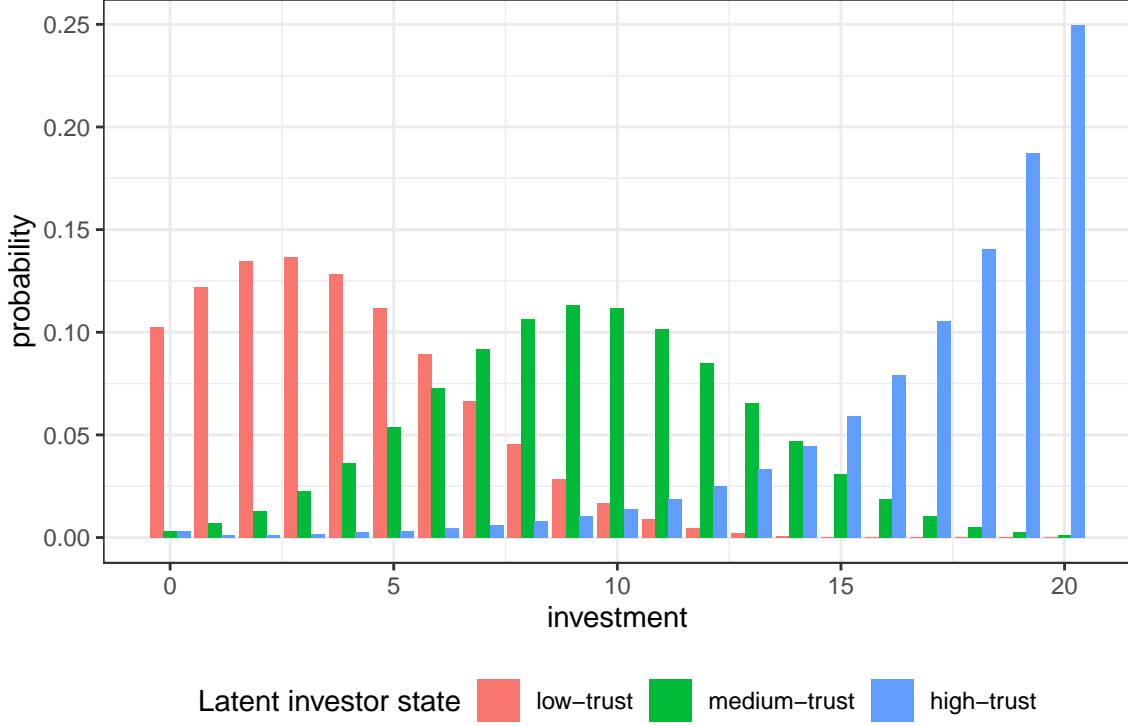


Figure 6: Distribution of investments by the artificial investor agent conditional on its latent state as estimated by a three state hidden Markov model with a discretised truncated Gaussian as a response function

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 6.

The HMM assumes that the probability of each investment $I_t = 0, \dots, 20$, at each trial t , conditional on the current state of the investor S_t , is dependent on an underlying normal distribution with mean μ_s and standard deviation σ_s . The probability of each discrete investment was determined by the cumulative normal distribution Φ . For instance, the probability of an investment $I_t = 2$ is defined as:

$$P(I_t = 2 | S_t = s) = \frac{\Phi(2.5 | \mu_s, \sigma_s) - \Phi(1.5 | \mu_s, \sigma_s)}{\Phi(20.5 | \mu_s, \sigma_s) - \Phi(-0.5 | \mu_s, \sigma_s)}$$

Note that the denominator truncates the distribution between 0 and 20. To estimate the transition probability between states for the investor, a multinomial logistic regression model was fitted to the investor’s data such as:

$$P(S_{t+1} = s' | S_t = s, X_t = x) = \frac{\exp(\beta_{0,s,s'} + \beta_{1,s,s'}x)}{\sum_{s''} \exp(\beta_{0,s,s''} + \beta_{1,s,s''}x)}$$

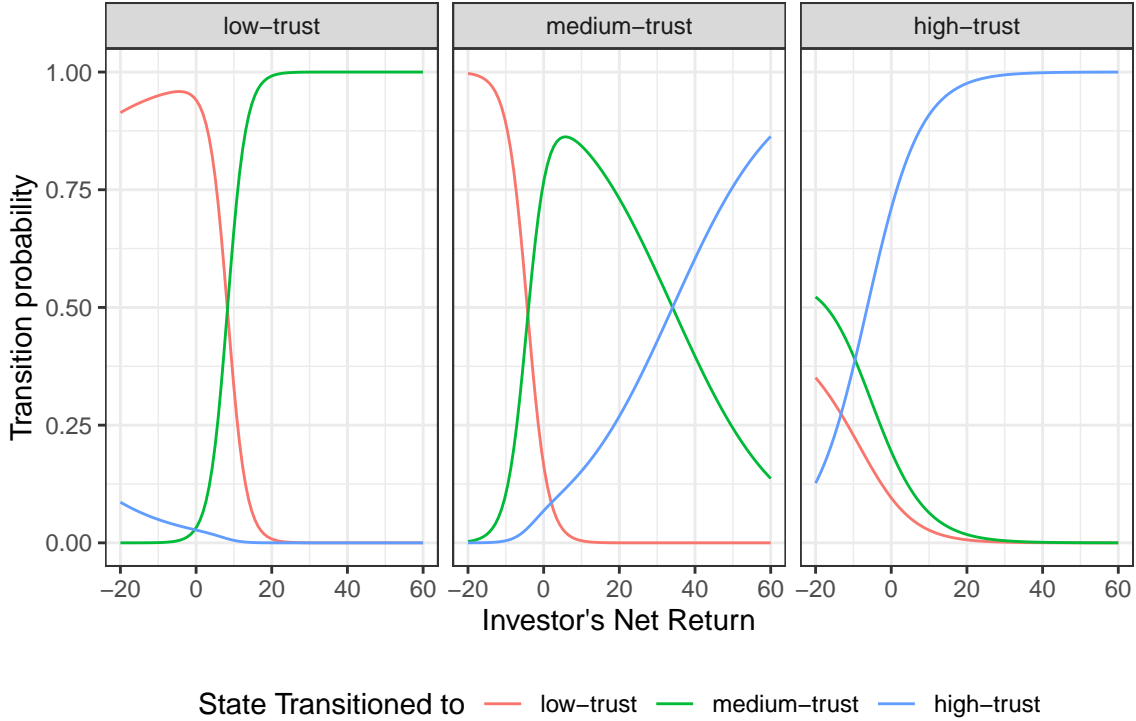


Figure 7: Transition function for investor conditional on current state

where $X_t = R_t - I_t$ is the net return to the investor with R_t the amount returned by the trustee and I_t is the Investment sent. Figure 7 shows these transition probabilities as a function of the net return conditional on the investor’s state.

For the repeated prisoner’s dilemma, we used an artificial agent that was playing according to Tit-For-Tat (Axelrod and Hamilton 1981). It started by choosing the cooperative action and then it mirrored whatever the other player played in the previous round.

2.4.5 AI agents pre-programmed defection

In order to probe efforts to repair trust, we introduced rounds where the investor deviated from the policy provided by the HMM, and chose to send a low investment in the Trust Game, or chose the non-cooperative action in the Prisoner’s Dilemma. In the trust games, we programmed the HMM agent to send a very low investment (2 out of 20) to the trustee on specific rounds (round 12 pre-intervention and 13 post-intervention). For the repeated Prisoner’s Dilemma game, we programmed the agent to choose the non-cooperative action on round 4. After each “defection” round, the agents were programmed to proceed their policy from before that round, effectively ignoring the participants’ response to their defection. For instance, the HMM investor would send an investment in the post defection round that was consistent with the outcome of the round immediately prior to that round. For the RPD, the agent would continue to play tit-for-tat, but also ignoring what happened in the defection round, and basing its action on mirroring what the other player did in the round immediately prior to the defection round. This ignorance of reactions to the defection rounds was chosen to reflect an accidental defection, with the Investor being open to repair subsequent interaction.

2.4.6 HMM analysis of participants returns in the RTG

To model participants returns in the RTG across games and conditions, we used an HMM response function based on a discretised Gaussian distribution that takes into account what proportion the trustee would ideally like to return, and what returns are possible given the investment. For instance, if the investor sends an amount of 2, the trustee would receive 6 and they can send back any amount between 0 and 6. As such, we assume that the response is a distribution over proportions that can be calculated from these possible returns, i.e. $\{0, 1/6, 2/6, \dots, 1\}$. The model assumes an underlying Normal distribution for each possible proportional return, predicting the probability of each via the cumulative Normal distribution with cut-off points set halfway between the proportions (e.g. the probability of returning $1/6$ is determined as the probability of returning anything between $1/12$ and $3/12$). The transition between states is assumed to depend on the investment through a multinomial logistic function such as:

$$P(S_{t+1} = s' | S_t = s, X_t = x) = \frac{\exp(\beta_{0,s,s'} + \beta_{1,s,s'} \text{inv} + \beta_{2,s,s'} x)}{\sum_{s''} \exp(\beta_{0,s,s''} + \beta_{1,s,s''} \text{inv} + \beta_{2,s,s''} x)}$$

where inv is a variable representing the investment received, x is a dummy variable to characterise the group that the participant belongs to. We define four contrast codes for these dummy variables: pre-post (comparing pre and post games), post-coax (compares the post-intervention group to all others), post-control (compares the post-control group to all others) and full-contrast (a three level dummy variable: post-intervention compared to post-control and all pre games).

Having defined these contrast codes, we then fit HMM models where the transition function between latent states depends both on the investment received and depending on the model, one of the aforementioned contrast codes. To select the number of hidden states governing participants' policies, 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

For the outcomes of the RTG, we first check the benchmark results for the average investment and return to see if they are comparable to results in the literature. We focus on the first 10 rounds pre-manipulation (the first RTG) and pre-defection trial to exclude the effects of the investor's transgression and those of the intervention and to have a comparable number of rounds to the standard version of the RTG. The average investments and returns were within the range of reported investments (40-60% of endowment) and returns (35-50% of total yield) in the literature, and hence comparable to other implementations of this task reported in the literature (Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011).

Figure 8 shows the average and standard error of trustee returns as a percentage of the total yield (3 x investment) over the full 15 rounds for both conditions and both RTG games. 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.

To explore whether participants behaved differently after the intervention compared to the control group over all rounds, we estimate a linear-mixed effects model, 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. We Z-transform the Investment variable as centering is beneficial to interpreting the main effects more easily in the presence of interactions. More complex models with additional random effects provided could not be estimated reliably.

Estimated marginal means of percentage returns as well as the distribution of the returns for participants across all trials are presented in panel A of Figure 9. If the intervention has an effect, then we should

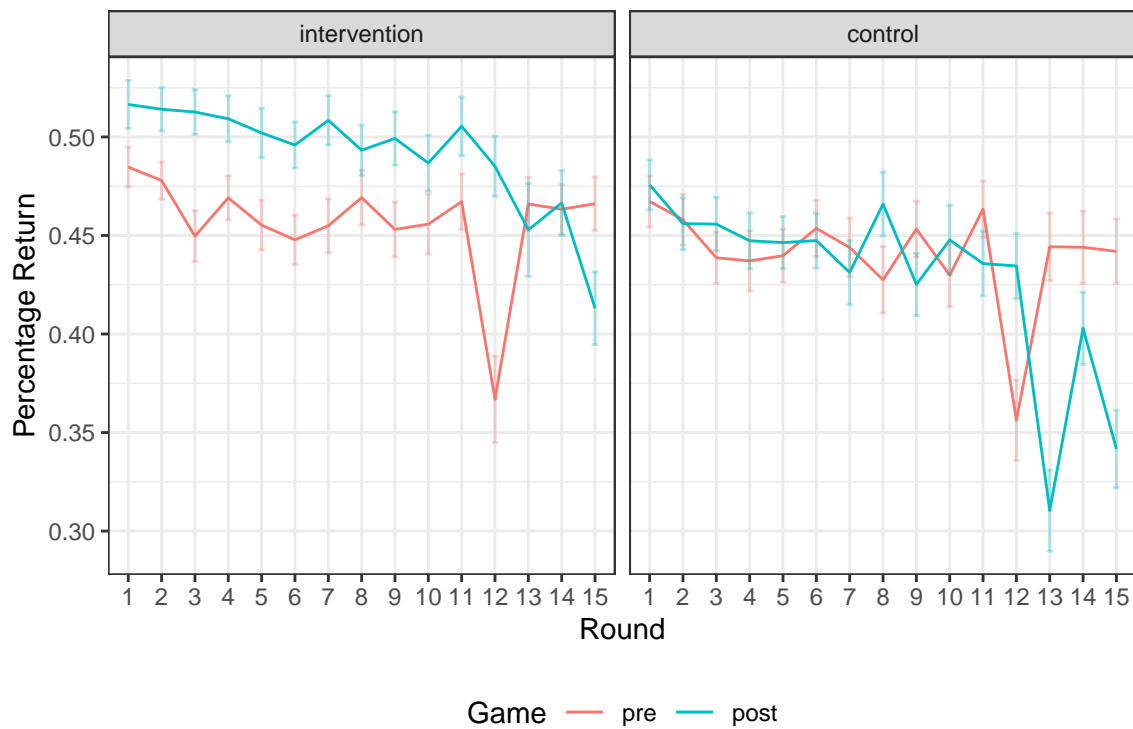


Figure 8: Average and standard errors of the trustee's return as a percentage of the multiplied investment received for each round and for both conditions. The red line shows the returns pre-manipulation and the blue line post-manipulation.

expect a higher increase in returns after the intervention compared to the control condition, which should be evident from an interaction between Condition (intervention vs control) and Game-number (pre- vs post-manipulation). Therefore, the variable of interest here is the interaction between Condition and Game-number. We do find an interaction effect due to an increase in the percentage returned in the intervention but not in the control group ($F(1, 314.2) = 26.9, p < 0.001$). We also find a significant main effect for Condition ($F(1, 315.31) = 9.52, p = 0.002$), which is due to participants returning higher percentage returns, in the post game, in the intervention group ($p < 0.001$), and lower returns in the control group ($p = 0.02$). We also find a main effect of Investment due to higher percentage returns being sent back for higher investments ($F(1, 9208) = 373.6, p < 0.001$), and an interaction effect of Investment by Condition ($F(1, 9207) = 45.38, p < 0.001$), with investments having less of a positive effect on the percentage returns in the intervention group compared to the control group. The latter effect indicates that the higher returns we see in the intervention group are unlikely to be purely driven by 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.

Since the artificial agent taking the role of the investor adapts its behavior to the actions of the participants, we can analyse whether the HMM agent exhibits a differentiated behavior between games and conditions. We use 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 panel D of Figure 9, 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** 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 absolute 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. We can rule out that these higher returns in the intervention group compared to the control group were purely driven by higher investments. If that was the case, then we would expect the presence of the Investment variable in our model of returns to lead to a non-significant interaction between Game-number and Condition. Since we find the interaction effect whilst controlling for Investment, this suggests this effect is not purely explained by the change in investments.

3.1 Pre-defection trials

Since we observed higher returns pre-defection trials, we can also explore whether there was differentiated behavior if we restrict our analysis to the rounds prior to the pre-programmed one-off investment (rounds 1 to 11 pre-manipulation and 1 to 12 post-manipulation). We 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 panel B of Figure 9. 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 the differentiated behavior we saw, where participants were sending back higher returns in the intervention condition but not in the control condition, was not simply due to their reaction post the low-investment trial. The higher returns were consistent with the message of the intervention, but have also happened, on average across participants, before the participants experienced the event similar to the one described in the intervention.

3.2 Defection and post defection trials

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. We now restrict 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).

Panel C of Figure 9 shows the marginal means of trustee returns and their distribution by Game-number and Condition. We found a main effect of Condition with returns on the intervention group higher than returns in the control group ($\Delta M = 0.05$, 95% CI [0.02, 0.08], $t(318.26) = 3.00$, $p = .003$). We also found a main effect of Game-number, with returns higher for the pre-manipulation game compared to the post-manipulation game ($\Delta M = 0.03$, 95% CI [0.01, 0.05], $t(319.44) = 3.09$, $p = .002$). In addition, there was a significant interaction between Game-number and Condition. We found a significant decrease in 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], $t(313.90) = -0.83$, $p = .407$).

3.3 Emotion self-reports

The results of the mixed-effects model of returns suggest that returns of participants in the intervention condition were less associated with the magnitude of the investment compared to those in the control condition. We can explore whether the emotional reaction to the investment was also lessened in the intervention condition through analysing the self-reported emotions on the two-axes feedback field. Figure 10 shows the average, across participants *in the intervention condition*, of emotion valence, the degree of arousal as well as the HMM’s investment z-scores. We find a significant positive correlation between the two emotion axis ($r = 0.52$, $p < 0.001$) as well as between valence and Investment ($r = 0.47$, $p < 0.001$), and arousal and Investment ($r = 0.50$, $p < 0.001$). Unsurprisingly, the low investment in the pre-determined defection round leads to negatively valenced emotions. Perhaps less expected is that they also lead to low arousal. We fit 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 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.

For emotional valence, we find a main effect of Investment ($F(1, 4664) = 2902$, $p < 0.001$), with higher investment leading to more positive emotions. We also find a main effect for Game-number, due to the reported emotional valence of investments decreasing significantly after the intervention ($\Delta M = 0.11$, 95% CI = [0.05, 0.18], $t(\infty) = 3.72$, $p < .001$). For emotional arousal, fitting a similar mixed effects model we find only a main effect of Investment ($F(1, 4668.4) = 1919$, $p < 0.001$) with higher investment leading to higher arousal. These effects are still present if we exclude the defection trial. We did not find any interaction between emotional valence or arousal and the investment, which would indicate a moderating effect of the intervention on the reaction to investments.

3.4 Player rating

One possible reason of the differentiated returns post manipulation between conditions is that participants had a different perception of the investor, post-manipulation, in the control condition compared to the intervention condition. To explore this possibility, we can determine whether participants rated their artificial opponents differently between conditions (intervention vs control) and games (pre- vs post-intervention) using the players ratings they had to submit at the end of each game. To analyse the participants rating of their opponents, we used a linear mixed-effects model for each of the four players ratings (cooperativeness, selfishness, trustworthiness, friendliness) with fixed effects for Game-number (pre or post manipulation)

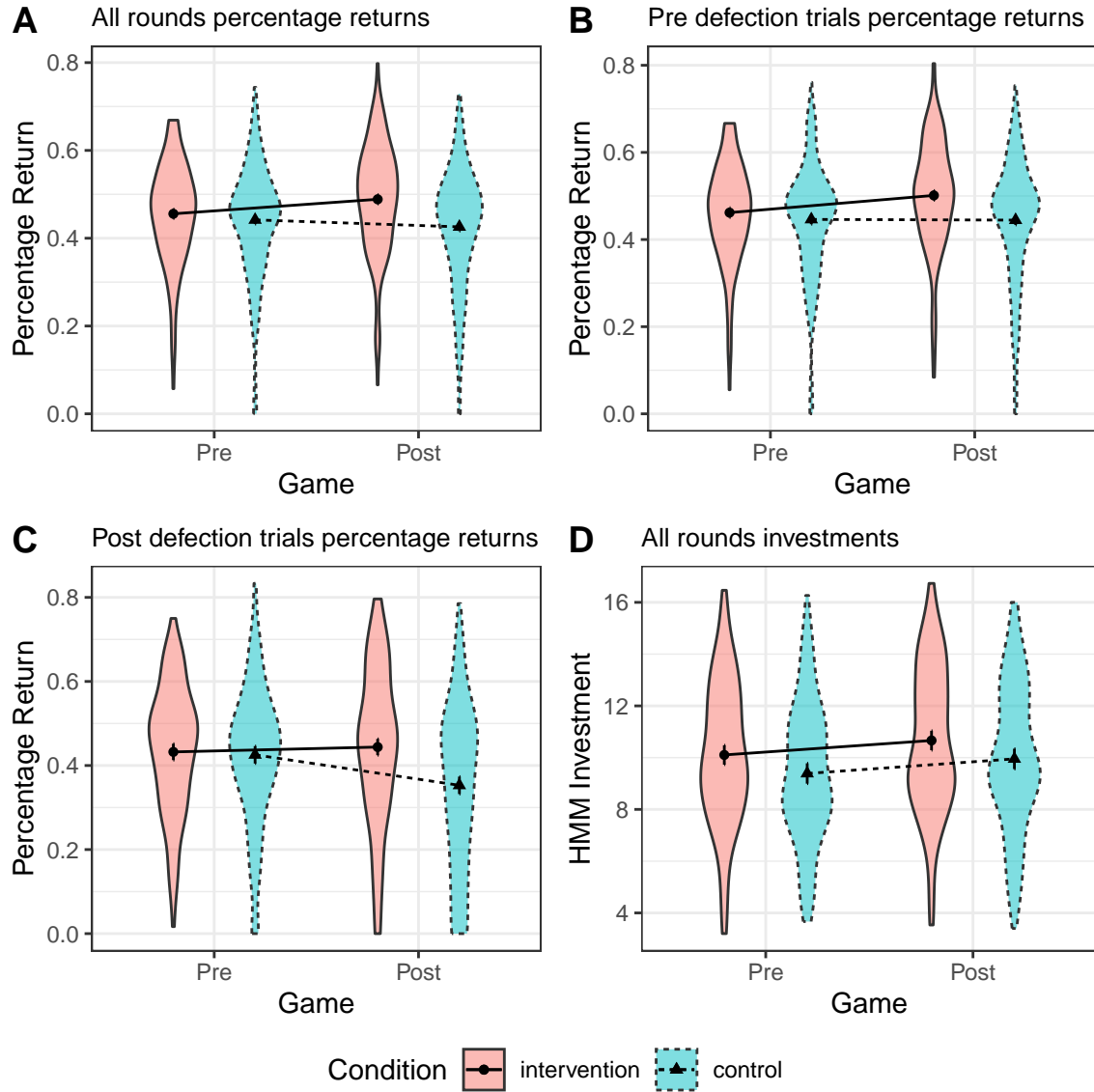


Figure 9: Panel A: Marginal means and distributions of percentage trustee returns over all rounds, shown across participants by Game number and Condition. Panel B: Marginal means and distributions of percentage trustee returns across all participants for pre-defection trials only, by Game number and Condition. Panel C: Marginal means and distributions of percentage trustee returns across all participants for post-defection trials only, by Game number and Condition. Panel D: Marginal means and distributions of investments over all rounds for HMM agents playing the role of the investor, by Game number and Condition

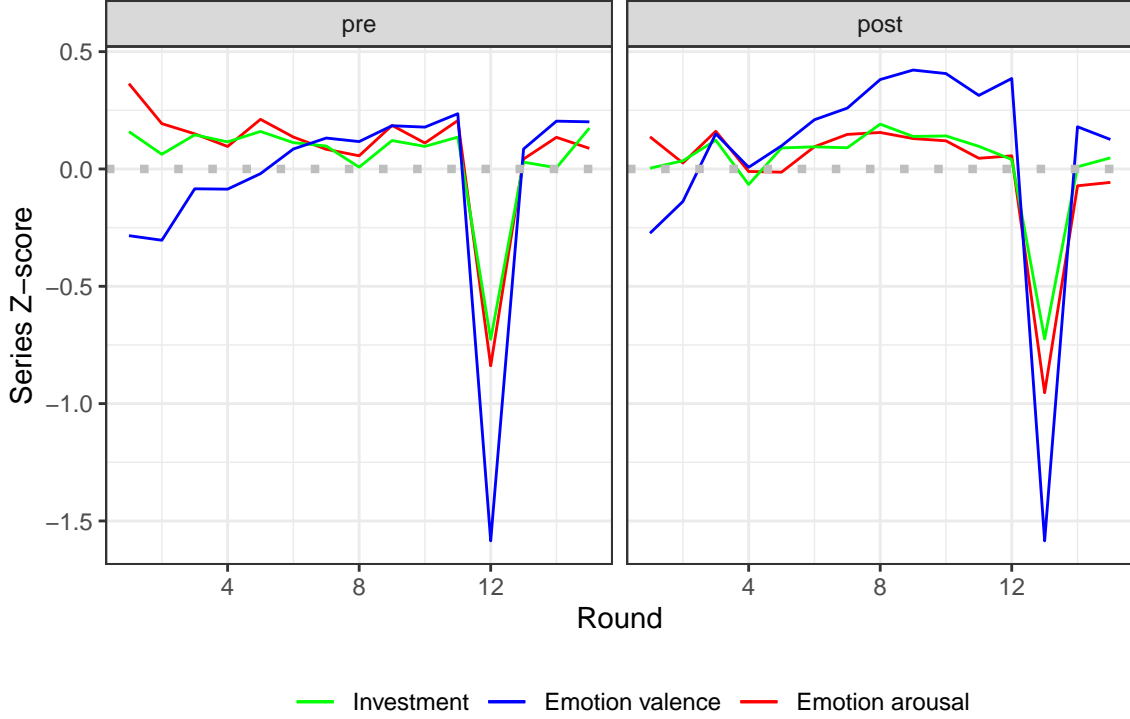


Figure 10: Self-reported emotion valence and arousal as well as investment z-scores for each round of the repeated Trust Game averaged across participants in the intervention condition only.

and Condition (intervention or control), as well as interaction between condition and Game-number, and participant-wise random intercepts.

We find 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], $t(317.00) = 3.10$, $p = .002$), less trustworthy ($\Delta M = 0.43$, 95% CI [0.16, 0.70], $t(317.00) = 3.19$, $p = .002$), less friendly ($\Delta M = 0.40$, 95% CI [0.17, 0.64], $t(317.00) = 3.36$, $p = .001$) and more selfish ($\Delta M = -0.36$, 95% CI [-0.61, -0.10], $t(317.00) = -2.76$, $p = .006$). 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], $t(317.00) = 1.95$, $p = .052$) and lower on selfishness ($\Delta M = -0.41$, 95% CI [-0.80, -0.02], $t(317.00) = -2.04$, $p = .042$). There was no evidence for an interaction effect between Game-number and Condition on any of the attributes.

3.5 Transfer to the Prisoner's Dilemma Game

If the intervention successfully increased coaxing behaviour in participants, we would hope they generalized this to other games beyond the Trust Game. Across participants, we first look at the rate at which the cooperative action was chosen in each round. Figure 11 shows the mean and standard error of the cooperation rate by round. Round 4 is where we programmed the Tit-for-Tat agent to defect, which explains the lower cooperation rate we see in round 5 onwards. 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.

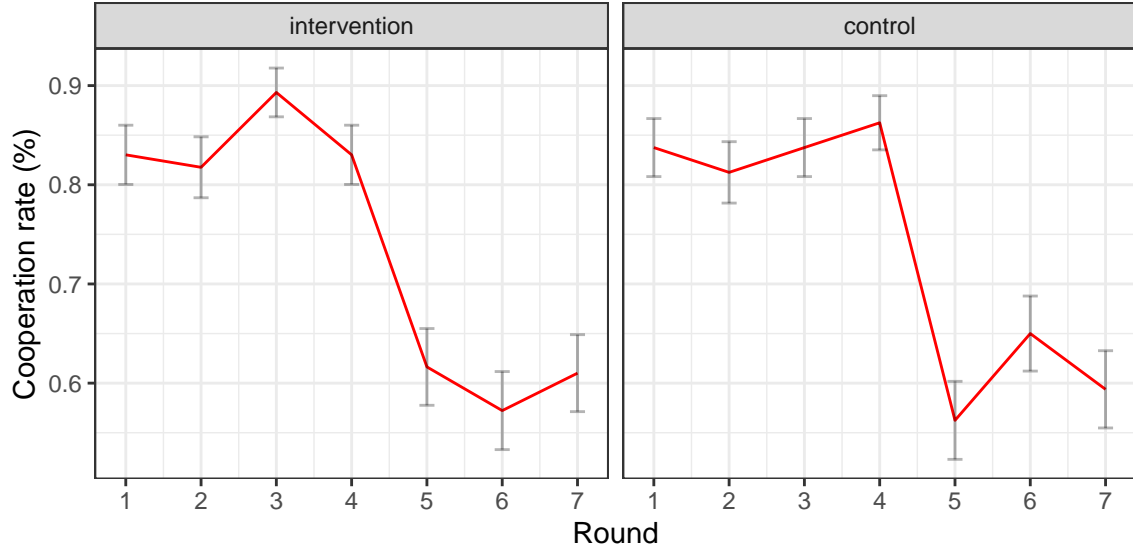


Figure 11: Mean and standard error of the rate at which the cooperative action was chosen by the participants for each round of the repeated Prisoner’s Dilemma game. The decrease of cooperation seen from round 5 onwards is likely due to a pre-determined defection of the agent in round 4

3.6 Self report questionnaires

Failure to repair a breakdown in trust in the repeated trust game has been associated with trustees with BPD traits (King-Casas et al. 2008). Theories of social dysfunction in BPD have focused on dysfunction in the patients’ mentalising ability (Allen and Fonagy 2006) as well as difficulties in emotional regulation (Rudge, Feigenbaum, and Fonagy 2020). The questionnaires we included in the experiment tried to assess borderline traits (PAI-BOR), emotional regulation capabilities (DERS) and mentalising ability (RFQ8). As such, we wanted to test whether there was an association between scores in these questionnaires and the effect of the intervention. We fit a linear mixed effect model to the percentage return of trustees with fixed effects for Condition (intervention or control), Game-number (pre or post manipulation), Investment, and questionnaire score as well as all interactions between the fixed effects. We assume participant-wise random intercepts. We Z-transform the questionnaire scores and Investment as centering would be beneficial to interpreting the main effects more easily. We find no interaction effect of the questionnaire scores with Condition variable, nor a main effect of the questionnaire scores.

3.7 Debrief Questions

In post experimental debrief questions, when asked whether they thought they were facing humans or machines, participants were divided in their answers. Although these were open questions, not easily amenable to quantitative analysis, we noted some important trends. Reading through participants answers, we note that most of those who correctly deduced that their opponents were artificial agents mentioned cues such as the speed of response of the agent or the duration it took to match up with a game partner to justify their answers. This means that this deduction was not associated with the way the agent was perceived to play the game. In terms of the agent’s behavior, participants expressed they felt it was either human or possessed human-like characteristics such as “spitefulness” or “greed”. In summary, participants did not systematically detect that their opponents were AI agents reacting to their returns, and were sometimes projecting human traits onto the observed behavior of their artificial opponents.

Table 1: Table of BICs for each of the estimated HMM models for assumed number of latent states between 2 and 7

Number of states	HMM-inv	HMM-ctrl	HMM-coax	HMM-prepost	HMM-full
2	50,321	50,365	50,393	50,363	50,383
3	46,676	46,743	46,763	46,766	46,833
4	45,371	45,523	46,585	46,533	46,707
5	45,220	45,298	45,425	45,284	45,525
6	44,813	45,124	45,400	46,235	46,199
7	45,068	45,383	45,392	45,363	46,026

4 HMM analysis of participant returns

To characterise differences between participants’ behavior in the intervention compared to the control condition following the manipulation, we analyse their observed returns using hidden Markov models. Here we present the results from fitting the HMM models to trustee data using various contrasts that distinguish between pre and post-manipulation and/or between conditions.

We distinguish between five types of models: First, we fit a simple model where we assume that the transition function between states depends only on the investment (HMM-inv). This specification does not distinguish between pre- and post-manipulation games, nor does it distinguish between control and intervention conditions. It simply assumes that the way the investment affects the transition function of the trustee is the same regardless of which Game number and condition the data is from. Second, we specify a model for the transition function where we contrast two groups: Pre and Post Intervention, irrespective of the condition (HMM-prepost). Alternatively, we also specify three other models with contrasts distinguishing between pre and post intervention and the condition. A model that contrasts the post-intervention condition to the pre-intervention and both control conditions (HMM-coax). Another model that contrasts the post-control condition to the pre-control and both intervention conditions (HMM-ctrl). Finally, in a full contrast model, both pre-control and pre-intervention groups are coded as one group, another group consists of post-control and a third group of post-intervention condition (HMM-full). For all five specifications, we fit various HMMs with a number of states varying between 2 and 7 and select the model with the lowest BIC.

Table 1 shows the BICs of the various fitted models for an assumed number of states between 2 and 7. For a simple model without any contrasts (HMM-inv) and a model with post-control only contrast (HMM-ctrl) we find a 6-state model to be best fitting. If the contrast is between the post-intervention group and all the other groups (HMM-coax), then a 7 state model is best fitting. When the contrast is comparing only pre and post Intervention groups (HMM-prepost), a 5 state model fits best. Finally, when we distinguish between pre-manipulation, post-control and post-intervention (HMM-full), we find that a 5 state model fits best. Since we only fit models between 2 and 7 states, it is possible that for those where we find the 7 state model to be best fitting, models with a higher number of states could fit the data better. We decided to stop at 7 states for computational cost reasons and because the interpretation of models with a higher number of states becomes complex.

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 of 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.

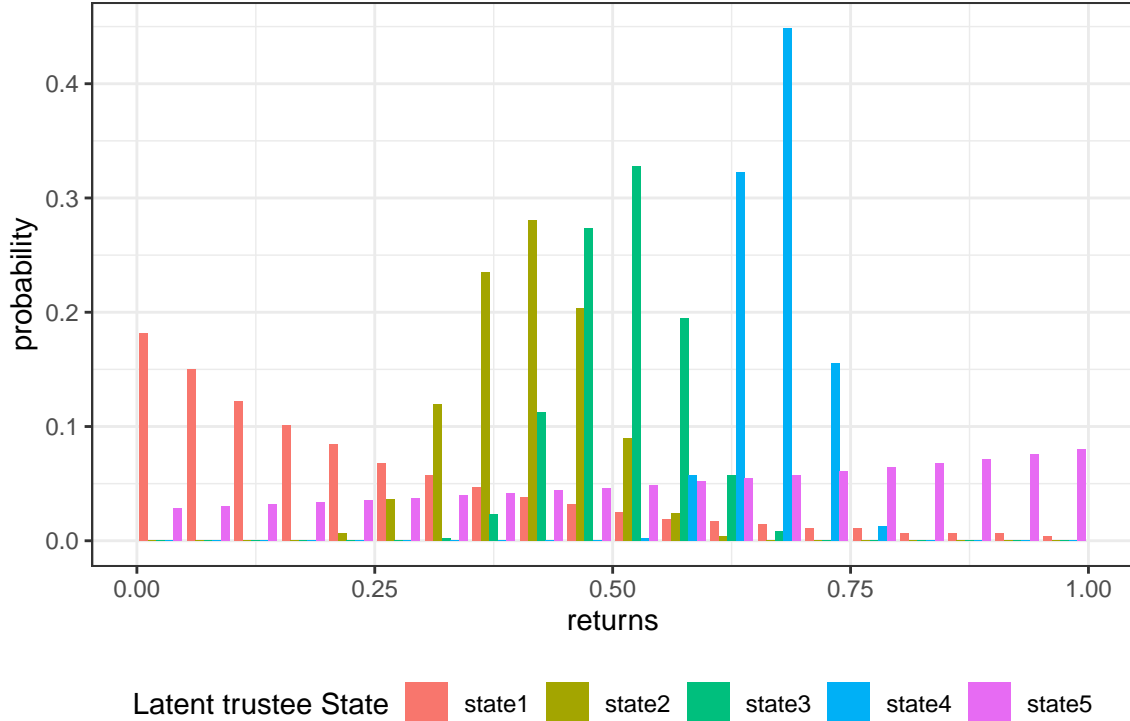


Figure 12: Distribution of participants' percentage return for each of the latent states in the 5 state HMM-full model

Using the best fitting model (HMM-full), Figure 12 shows the distribution of the participants' returns conditional on the latent state they are in. The states are ordered based on the mean of the Gaussian distribution fitted to the response (percentage return). State 1 represents the state in which the returns have the lowest underlying mean, and state 5 is the state in which the returns have the highest underlying mean. The higher the state number the more pro-social the policy adopted. More specifically, state 1 can be thought of as a non-cooperative state in which returns are low and close to 0, meaning that the trustee is keeping most of the tripled investment. State 2 is a state where the average return is around the investment sent and can be interpreted as a cautious state in which the return on trust is small. States 3, 4 and 5 are increasingly cooperative states. For instance, in state 5, the average return is close to two thirds of the tripled investment, meaning the trustee is keeping an amount similar to the investment that was sent and returning to the investor double the investment.

Figure 13 shows the transition between states as a function of the investment received for each group (pre-manipulation, post-control and post-intervention) using results from the best fitting model: HMM-full. The best fitting HMM model has multiple states and a high number of parameters. We can nonetheless focus on particular states linked to the breakdown and repair of cooperation. For instance, focusing on the transition functions to state 1 (the lowest return state with returns close to 0) from higher return states, we can compare the post-control group to the post-intervention group. The red line representing the probability of transitioning to the low-return state 1 when the investment is close to 0 is lower in the intervention group compared to the control group when the trustee is in a relatively pro-social state (states 3 to 5). This is pointing towards more "forgiving" behavior where participants in the intervention group were less likely to transition to this anti-social state 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, the most likely posterior state of the trustee participants by round given the actions they have taken. Figure 14 shows the distribution of these

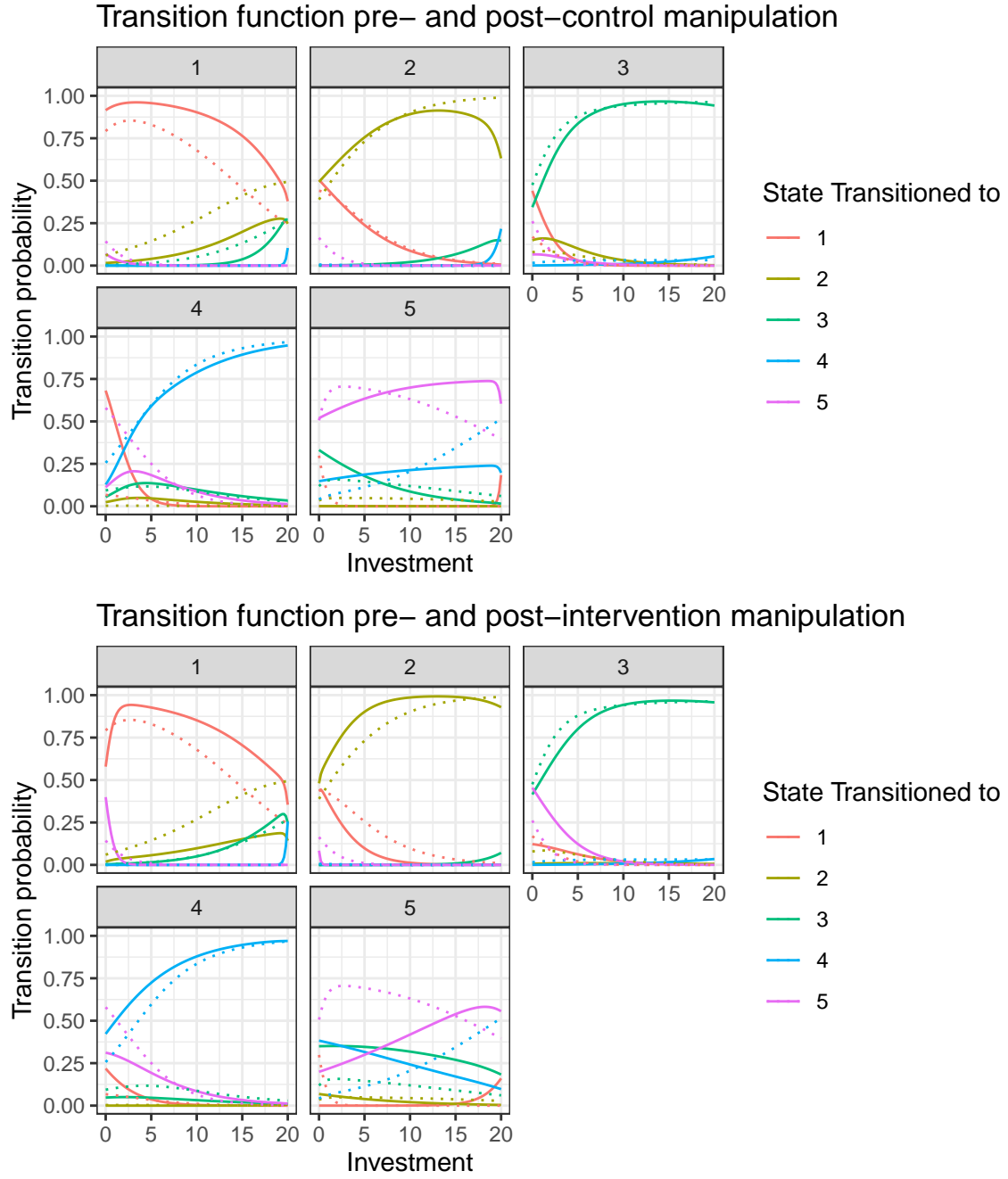


Figure 13: Transition function for the HMM-full trustee model. Each panel represents the state transitioned from, and each color the state transitioned to. Solid lines show estimated transition probabilities post-manipulation. Dotted lines show the same probabilities prior to the manipulation

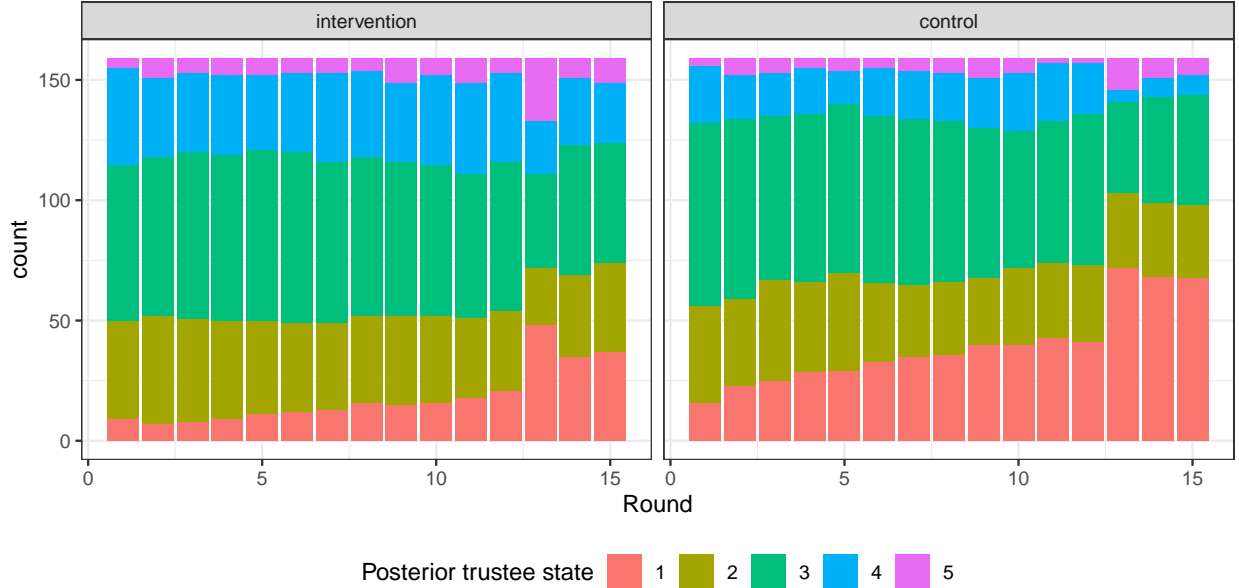


Figure 14: Distribution of posterior trustee states post manipulation by condition for all rounds, as estimated by the most likely posterior state in the best fitting HMM model (HMM-full) using a local decoding procedure.

states by round for both the intervention and control groups post manipulation. We can see that participants were more likely to be in a lower return state in the control condition compared to the intervention condition. For instance, in round 5, state 1 was the most likely posterior state for 7% of participants in the intervention condition compared to 24% in the control condition, indicating a higher proportion of participants were in the lowest trustworthiness state without the intervention, and the differences between proportions were significant ($\chi^2(1) = 8.26, p < 0.01$). For the post defection trial after the intervention (round 14), state 1 was the most likely state for 22% of participants in the intervention condition compared to 43% in the control condition, and the differences between proportions were significant ($\chi^2(1) = 14.70, p < 0.001$).

It is also noteworthy that there is an important heterogeneity in participants' most likely posterior state when faced with a situation like the one described during the intervention manipulation. Focusing on round 13 (the round where the low investment was sent post manipulation), we can see that a high proportion of participants (30.2%) in the intervention condition were in a low-return state. This group does not exhibit a behaviour consistent with the goal of the intervention. By contrast, another group exhibited behavior that is likely to be driven by assimilating the message from the coaxing manipulation, through transitioning to a high-return state even in the face of a low investment, which would be consistent with a coaxing behavior. Whilst only 11.3% in the control condition were in states 4 and 5 following the pre-programmed low investment, 30.3% of participants in the intervention condition were estimated to be in those states. The difference between the proportions is statistically significant ($\chi^2(1) = 16.08, p < 0.001$). 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 comments projecting human attributes on the opponent from the answers to the debrief questions, 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. It explained that reciprocating this non-cooperative action might signal untrustworthiness and lead to a more permanent breakdown of cooperation. It also suggested an alternative strategy of coaxing the investor back into cooperation. When the agent defected as programmed, yielding a situation similar to the one presented in the intervention, we saw that the intervention led to higher percentage returns post defection compared to the control intervention, as intended. What was also notable, is that the intervention led to generally higher returns in post intervention, indicative of more cooperative behavior even before the defection trial. The question therefore is: what are the reasons for participants to engage in higher cooperative actions even in the absence of a sudden low investment? One possibility is that they have simply learned that by returning more, the investment on the next trial would be higher. If that was the case, we would expect a similar increase in percentage returns in the control condition, which we didn’t find. 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. Likewise, 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 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 the 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 be 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, with visual inputs such as cartoons or short videos, rather than only text, which can be more 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 opined 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. 2006. *The Handbook of Mentalization-Based Treatment*. John Wiley & Sons.
- 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. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80 (4): 1095–1111. <https://doi.org/10.2307/1960858>.
- 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](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 Tudományos Folyoirata* 32 (3): 283–87.
- Fonagy, Peter, Patrick Luyten, Alesia Moulton-Perkins, Ya-Wen Lee, Fiona Warren, Susan Howard, Rosanna Ghinai, Pasco Fearon, and Benedicte Lowyck. 2016. "Development and Validation of a Self-Report Measure of Mentalizing: The Reflective Functioning Questionnaire." Edited by Keith Laws. *PLOS ONE* 11 (7): e0158678. <https://doi.org/10.1371/journal.pone.0158678>.
- 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>.
- Gratz, Kim L., and Lizabeth Roemer. 2004. "Multidimensional Assessment of Emotion Regulation and Dysregulation: Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale." *Journal of Psychopathology and Behavioral Assessment* 26 (1): 41–54. <https://doi.org/10.1023/B:JOBA.0000007455.08539.94>.
- Gunderson, John G., Sabine C. Herpertz, Andrew E. Skodol, Sverre 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](https://doi.org/10.1016/S0140-6736(04)16770-6).
- 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>.
- Morey, Leslie Charles. 1991. *The Personality Assessment Inventory TM: Professional Manual*. PAR,

- Psychological Assessment Resources, Incorporated.
- 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>.
- Rousseau, Denise M., Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. "Introduction to Special Topic Forum: Not so Different After All: A Cross-Discipline View of Trust." *The Academy of Management Review* 23 (3): 393–404.
- 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): 461–64.
- Vickrey, William. 1961. "Counterspeculation, Auctions, and Competitive Sealed Tenders." *The Journal of Finance* 16 (1): 8–37. <https://doi.org/10.2307/2977633>.