Can Exposure To Forgiving AI Foster Cooperative Play?

Ismail Guennouni, Georgia Koppe, Christoph Korn

1 Introduction

Trust is fundamental to human social interactions, enabling smooth relationships at both interpersonal and intergroup levels. The study of psychopathology has linked deficits in trust-based constructs to the development of mental health disorders (Fonagy and Campbell 2017). Individuals with personality disorders often struggle to form and maintain social connections, a difficulty reflected in uncooperative behaviors – a marker for the severity of PD symptoms (Herpertz and Bertsch 2014; Mulder et al. 1999).

One explanation for such social challenges lies in early caregiver experiences. Attachment theory (Bowlby 1978) suggests that the quality of these relationships shapes our capacity for secure attachments and trust. Individuals with higher levels of insecure attachment may recall negative trust-related experiences more easily, report fewer positive trust experiences, and use less constructive coping strategies when trust is broken (Mikulincer 1998). Similarly, learners exposed to unreliable communicators could develop mistrust of social knowledge as a protective strategy (Fonagy and Allison 2014).

If this adaptive mistrust is the source of social dysfunction, we can ask whether exposing those who exhibit it to cooperative and forgiving interaction partners might correct this bias. Research in the fields of behavioral economics and psychology has explored how positive social interactions influence trust and cooperation. The use of the repeated trust game (RTG), a well-established experimental approach, has allowed for the analysis of the development of trust through ongoing interactions (Joyce, Dickhaut, and McCabe 1995). In this paradigm, cycles of mutual trust, where each party's trust is reciprocated with trustworthiness, have the effect of enhancing cooperative behaviors and trust levels, even among individuals who are initially inclined to be distrustful (King-Casas et al. 2005). Fowler and Christakis (2010) studied behavior in social networks interacting in a public goods game and found that cooperative behavior tends to cluster, suggesting that exposure to cooperative peers can lead to more cooperative behavior. Similarly, research on social learning theory (Bandura 1977) has long demonstrated that individuals learn and model the behavior of those around them, indicating that if someone is consistently exposed to cooperative and positive individuals, they're likely to emulate this behavior. These insights highlight that engaging with compassionate and forgiving others can be an effective method for mitigating deeply ingrained mistrust.

In this study, we use a randomized control trial to test a manipulation aimed at repairing potential breakdowns in RTG cooperation. Participants are exposed to agents designed with a limited propensity for retaliation, potentially mitigating ingrained mistrust. We stratify our sample based on rejection sensitivity (RS): a tendency to anxiously expect, readily perceive, and intensely react to rejection. It has been identified as a potential mechanism linking early interpersonal trauma to negative mental health outcomes (Downey, Khouri, and Feldman 1997). Extensive research demonstrates strong associations between rejection sensitivity and various mental health conditions, including depression, anxiety, personality disorders, and self-harm (for a review, see Gao et al. (2017)). However little is known about whether rejection sensitivity is linked to mistrust and difficulties in cooperation in social dilemmas. Finally, the computerized investor in the RTG was programmed to play according to a hidden Markov model estimated from real players' data in prior research. A key aspect of this model is that the actions of the investor depend on a latent "trust state" which reacts dynamically to the trustees' returns simulating real-life trust-building scenarios. Such a model, informed by empirical data and integrating a responsive trust mechanism, represents a novel approach to study interactive behavior in multi-player games whilst keeping a high degree of experimental control.

2 Methods

2.1 Participants

A total of 206 participants were recruited on the Prolific Academic platform (prolific.co). Participants were paid a fixed fee of £6 plus a bonus payment dependent on their performance that averaged £XXX. Participants were pre-screened on Prolific using the Rejection Sensitivity Questionnaire to form two similarly sized groups: One with high (RSQ score > 15) and the other with low rejection sensitivity (RSQ score < 10).

2.2 Design and Procedure

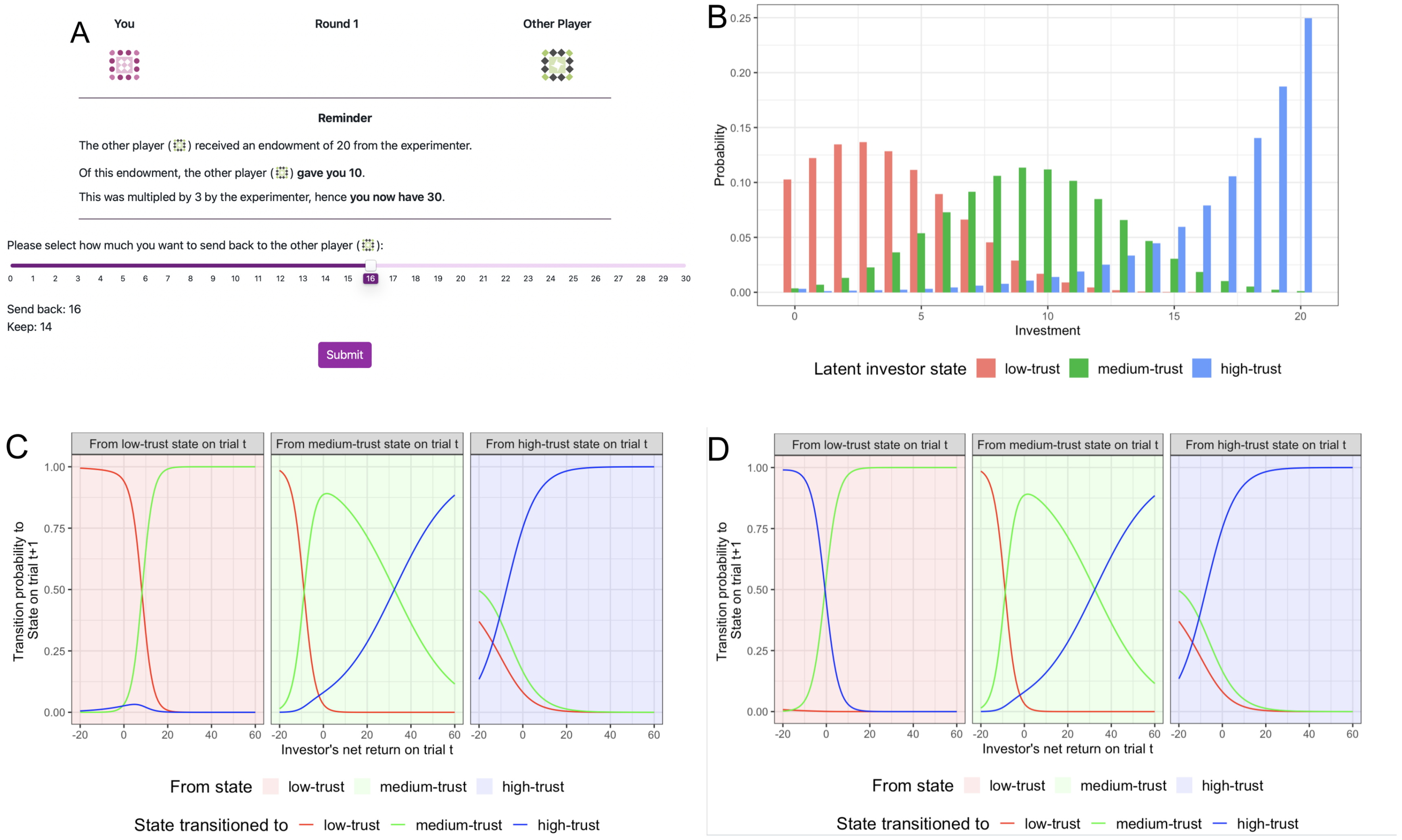
The experiment had a 2 (Condition: Manipulation or Control) by 2 (Rejection Sensitivity : high or low) by 2 (Game: Trust-Game Pre Manipulation, Trust-Game Post Manipulation) design, with repeated measures on the third factor. Participants within each pre-screened group were randomly assigned to one of the two levels of the first factor. The games were designed and implemented online using Empirica v1 (Almaatouq et al. 2021).

2.3 Tasks and Measures

2.3.1 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 is endowed with 20 units and decides how much of that endowment to invest. This investment is tripled and the trustee then decides how to split this tripled amount between them and the investor. If the trustee returns more than one third of the amount, the investor makes a gain. Each player was represented with an icon with the participant always on the left of the screen and the co-player on the right (Figure 1.A). The participants were able to choose the icon that represents them at the start of the experiment. The icon representing the co-player changed at the start of each new game, to simulate a new interaction partner. Participants were not told they were facing computerised co-players.

The strategy of the computerised investor was modelled on behavior of human investors in the Repeated Trust Game (RTG) over 10-rounds with the same (human) co-player. Full detail on the datasets used in the Supplementary Information. Using this data, we estimated a hidden Markov model (HMM) on investors' behavior with three latent states. Each latent state was associated with a state-conditional distribution over the possible investments from 0 to 20 (Figure 1.B). These distributions reflect "low-trust", "medium-trust", or "high-trust". Over rounds, the investor can move between states, and the probability of these transitions was modelled as a function of their net return (i.e return - investment) in the previous round (see Figure 1.C). In order to instigate a potential breakdown of trust, thereby allowing us to probe efforts to repair trust, the computerised agent was programmed to provide a low investment on round 12 (pre-manipulation) or round 13 (post-manipulation). On all other rounds, the investor's actions were determined by randomly drawing an investment from the state-conditional distribution, with the state over rounds determined by randomly drawing the next state from the state-transition distribution as determined from the net return on the previous round (disregarding the net return immediately after the pre-programmed low investment rounds). The initial state for the HMM investor in each instance of the game was the "mid-trust" state.



Reminder

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

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

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

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

Send back: 16
Keep: 14

Figure 1: A: 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 who 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 decides how much to send back to the investor. Panels B - D: We construct the artificial investor agent by fitting a three-state hidden Markov model to data of human investors engaged in the 10 round Repeated Trust Game against human trustees. From the fitted HMM, we get the distribution of investments by the artificial investor agent conditional on its latent state as shown in Panel B. The fitted HMM also yields the transition probability of the agent to a state on trial t+1 as a function of the net return (difference between the investment sent and the amount received in return) on trial t as shown in Panel C. Each plot in Panel C represents a different starting latent state on trial t, and each line represents the probability of transitioning to a particular state in trial t+1. Panel D shows the transition probability of the forgiving HMM agent, where we can see on the left plot that the agent always transitions out of the low-trust state.

2.4 Manipulation

To design a forgiving and ultimately more cooperative agent, we change the transition function of the investor HMM to make it impossible to remain in a low trust state once the agent transitions there. This is achieved by choosing the parameters of the transition function to make the probability of remaining in the "low-trust" state, when the agent is in that state, effectively nil. The policies conditional on the state and the transition function in the other states remain unchanged. The resulting transition function is shown in Figure 1.D.

2.5 Procedure

At the start of the experiment, 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 decided 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, forgiving they perceived the investor to be, and whether they would like to play with them again (all on a scale from 1 to 10).

After phase one, participants in the manipulation condition played three games of 7 rounds each against the forgiving HMM agent, whilst those in the control condition faced the human-like HMM agent. To keep the experience similar to the pre-manipulation game, the agent in the control condition was also designed to send a very low investment in round 5 of each of the three games. Subsequent phase two was similar to phase one, with participants being told they would face a new player. Participants then completed the Levels of Personality Functioning Scale (LPFS) questionnaire (see the supplement for details). Finally, participants were asked whether they thought the other players were human or computer agents, then debriefed and thanked for their participation.

2.6 Statistical analysis

To explore whether participants behaved differently in the RTG after the manipulation compared to the control group, we model the percentage return (percentage of tripled investment returned to investor) using a linear mixed-effects model as described below:

$$\begin{aligned}
\mathrm{R}_{ij} = & \beta_0 + \beta_1 \text{ Phase}_i + \beta_2 \text{ Condition}_i + \beta_3 \text{ Investment}_i + \beta_4 \text{ RS}_i + \\
& \beta_5(\text{Phase} \times \text{Condition})_i + \beta_6(\text{Phase} \times \text{Investment})_i + \beta_7(\text{Phase} \times \text{RS})_i + \\
& \beta_8(\text{Condition} \times \text{Investment})_i + \beta_9(\text{Condition} \times \text{RS})_i + \beta_{10}(\text{Investment} \times \text{RS})_i + \\
& \beta_{11}(\text{Phase} \times \text{Condition} \times \text{Investment})_i + \beta_{12}(\text{Phase} \times \text{Condition} \times \text{RS})_i + \\
& \beta_{13}(\text{Phase} \times \text{Investment} \times \text{RS})_i + \beta_{14}(\text{Condition} \times \text{Investment} \times \text{RS})_i + \\
& \beta_{15}(\text{Phase} \times \text{Condition} \times \text{Investment} \times \text{RS})_i + \\
& b_{0j} + b_{1j} \ (\text{Phase})_i + \epsilon_{ij}
\end{aligned}$$

where:

- R_{ij}: percentage of tripled investment returned to investor for participant j in observation i
- β_0: intercept
- β_1 to β_4: main effects of Phase (RTG game pre vs. post-manipulation), Condition (manipulation vs. control), Investment, and RS (High vs Low RS), respectively

- β_5 to β_{10}: interaction effects between each pair of the four factors, showing how the relationship between one factor and the return percentage not available changes depending on the level of another factor
- β_{11} to β_{14}: three-way interaction effects among the four factors, indicating how the interaction between two factors is further modified by the third factor
- β_{15}: four-way interaction effect between Phase, Condition, Investment, and RS, describing how the interaction among three factors is modified by the fourth factor
- b_{0j}: player-wise random intercept for player j
- b_{1j}: player-wise random slope for Phase for player j
- ϵ_{ij}: error term for player j in observation i

The model was estimated using the `afex` package (Singmann et al. 2022) in R. More complex models with additional random effects could not be estimated reliably, and as such the estimated model can be considered to include the optimal random effects structure (Matuschek et al. 2017). A similar process was used to establish the random effects structures of other linear mixed-effects models used throughout the statistical analyses. 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.

3 Behavioral Results

3.1 Player ratings

Figure 2 shows participants’ ratings of each player they faced by condition. Compared to the first RTG game, participants in the manipulation condition rated the investors they faced in the exposure phase (the forgiving AI) as more Cooperative $\Delta M = 1.37$, 95% CI $[0.14, 2.60]$, $t(816) = 2.18$, $p = .030$. No difference in ratings on forgiveness and whether they would like to face them again. Those in the control condition rated the investors faced in the exposure group (same HMM) as less cooperative $\Delta M = -1.88$, 95% CI $[-3.08, -0.67]$, $t(816) = -3.05$, $p = .002$, less forgiving $\Delta M = -1.14$, 95% CI $[-2.41, 0.13]$, $t(816) = -1.77$, $p = .078$, and were less keen on facing them again $\Delta M = -3.02$, 95% CI $[-4.50, -1.53]$, $t(816) = -3.99$, $p < .001$.

Comparing post-exposure to exposure ratings, participants in the manipulation condition rated the investors they faced in the exposure phase as less cooperative$\Delta M = -2.76$, 95% CI $[-3.99, -1.53]$, $t(816) = -4.40$, $p < .001$, less forgiving $\Delta M = -3.60$, 95% CI $[-4.90, -2.31]$, $t(816) = -5.47$, $p < .001$, and were less willing to face them again $\Delta M = -3.33$, 95% CI $[-4.84, -1.81]$, $t(816) = -4.31$, $p < .001$. Ratings for those in the control group did not differ on how cooperative, forgiving the investors were, and on willingness to face them again.

Finally, comparing ratings of the investors before and after the exposure phase, those in the manipulation condition rated them similarly on cooperation, lower on forgiveness $\Delta M = -1.01$, 95% CI $[-1.54, -0.48]$, $t(816) = -3.75$, $p < .001$ and lower on willingness to face them again $\Delta M = -1.30$, 95% CI $[-1.92, -0.68]$, $t(816) = -4.12$, $p < .001$. Those in the control condition rated the investors post the exposure phase lower on all three attributes.

When asked during debrief whether they thought the investors they faced were Human or not, 41% of participants thought they were either facing a human or were not sure of the nature of the co-player. When asked to justify their choice, many answers reflected participants projecting human traits such as “spitefulness” or “greed” onto the artificial co-player’s behavior.

3.2 Analysis of participant returns

On average, investments and returns, as shown in Figure 3, fell within the documented range of 40-60% of the endowment for investments and 35-50% of the total yield for returns, as reported in previous studies

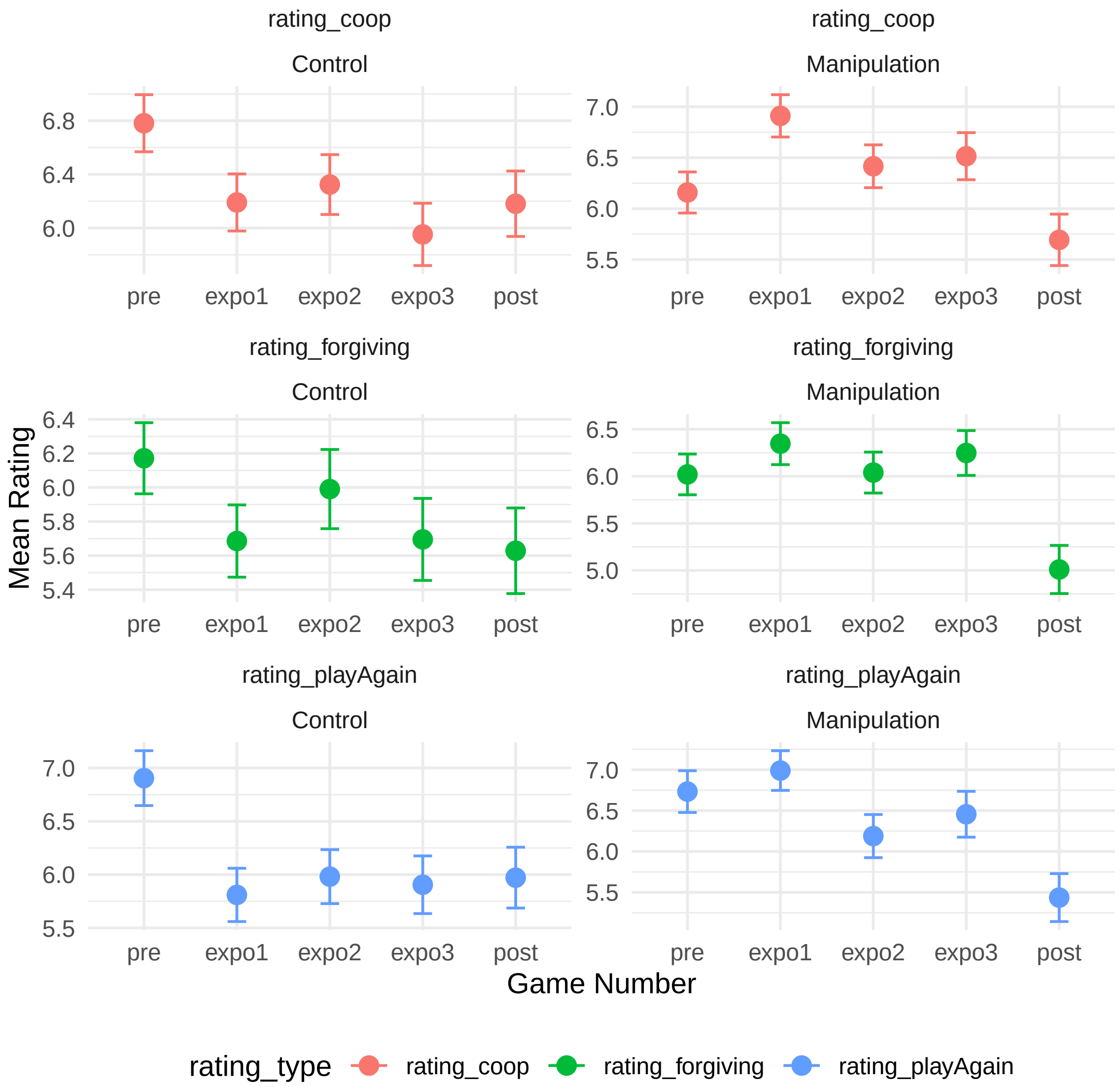


Figure 2: Averages and standard errors of the participants ratings of the opponent for each game and condition. Pre and post are the 15 round repeated trust games before and after the exposure phase respectively. The games titled expo 1 to 3 are the three 7 round games during the exposure phase. We note that absent the exposure to the forgiving AI, the ratings get worse on aggregate even through the participant faces the same human like HMM. In the manipulation condition, participants rate the nicer HMM as more cooperative but not more forgiving. When they face the human-like HMM again, its rating are considerably worse.

(Charness, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 2011).

Mixed-effects analysis on the percentage returns shows a significant main effect of Phase (Pre vs. Post RTG game), $F(1, 201.63) = 5.81$, $p = .017$, with higher percentage returns in the first RTG compared to the second. Importantly, we also find an interaction between Condition and Phase (RTG pre- vs. post-intervention), $F(1, 201.63) = 4.38$, $p = .038$. Post-hoc tests show a decrease in the percentage returned only in the intervention condition, pre - post, $\Delta M = 0.03$, 95% CI $[0.01, 0.05]$, $t(201.50) = 3.15$, $p = .002$, but no change in the control condition. Looking at this interaction effect for the two levels of RS, we find a three way interaction between Phase, Condition and RS, such that this decrease in returns in the manipulation condition is present in the low RS group $\Delta M = 0.05$, 95% CI $[0.02, 0.08]$, $t(200.61) = 3.20$, $p = .002$ but not in the high RS group.

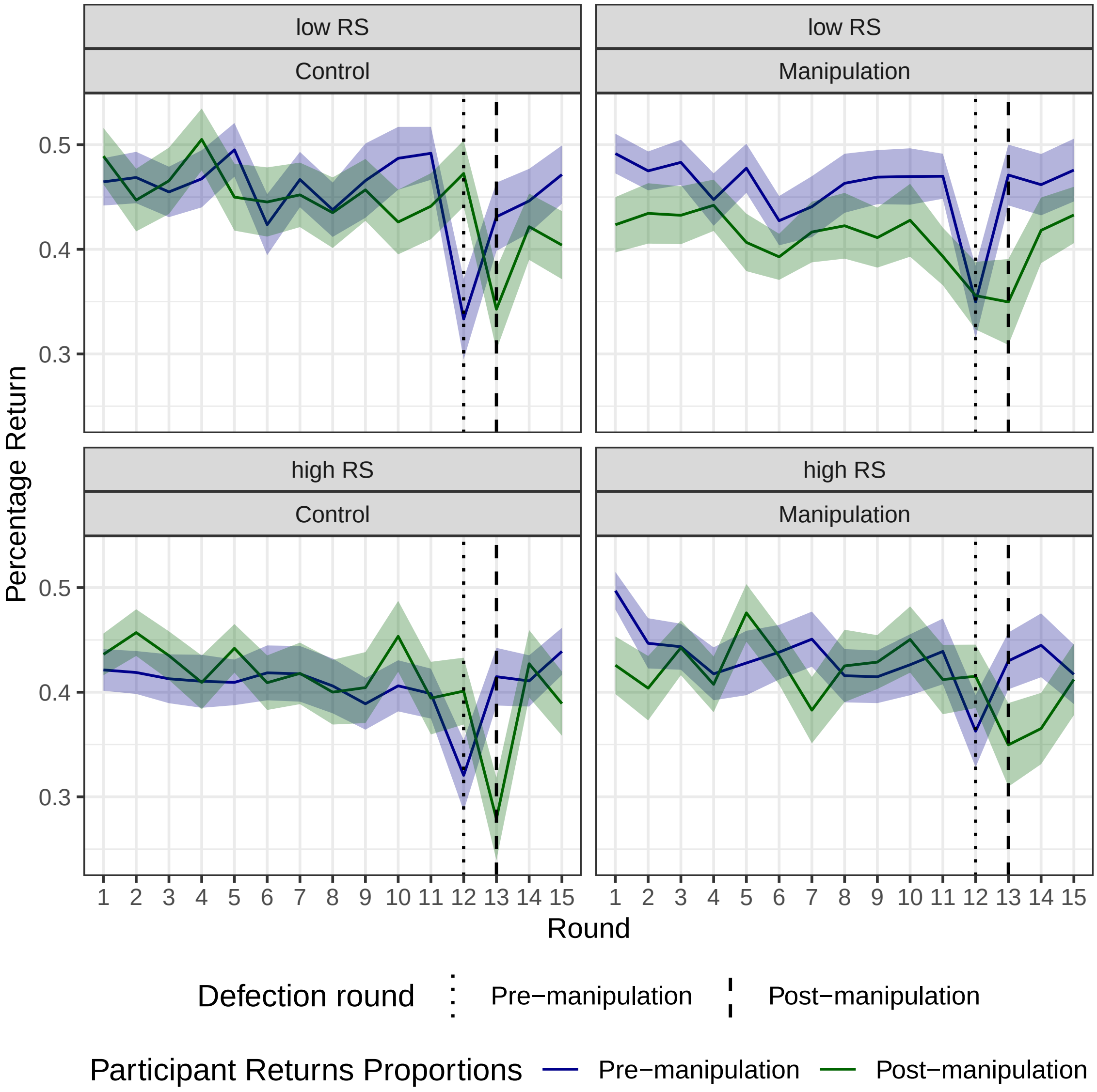


Figure 3: Averages and standard errors of the trustee's return as a percentage of the multiplied investment received by Condition, Phase, and game round. The blue line shows the returns pre-manipulation and the green line post-manipulation. We note a different reaction to the pre-programmed one-off low 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 manipulation condition compared to the dip in returns seen in the control condition in the right panel

There was also a significant main effect of Investment, $F(1, 5955.67) = 325.35$, $p < .001$, such that higher investments were associated with higher percentage returns indicating positive reciprocity. An Investment by Condition interaction, $F(1, 5955.67) = 13.92$, $p < .001$, indicates that returns were more affected by investments in the control condition. We also find a three way interaction between Phase, Investment and RS , showing that the differentiated effect of the investment on the proportion returned by RS group is itself moderated by the Phase (pre- vs post intervention).

3.2.1 Post Defection Trials

Running the same mixed effects model only on the trials following the pre-programmed defection by the HMM agent, we find a significant main effect of Phase $F(1, 224.64) = 4.99$, $p = .026$ with returns lower in the second game post defection trials compared to the first. We also find a main effect of Investment $F(1, 1289.08) = 154.11$, $p < .001$ where participants continued to return higher proportions when receiving higher investments. Finally, we still find an Investment by Condition interaction $F(1, 5955.67) = 13.92$, $p < .001$ showing a lower effects of investment on the manipulation condition compared to the control condition in post-defection trials.

3.2.2 HMM investments

To explore the HMM investors' behavior across games and conditions, we estimate a linear mixed-effects model of investments sent by the computerised HMM agent with Condition, Phase and RS and their interaction as fixed effects, and a similar random effects structure to the returns model. This shows no main or interaction effects, indicating HMM behavior was on aggregate similar across Phase, Conditions and RS groups.

4 Discussion

The use of HMM-based artificial agents in economic games led to similar investment and returns to those recorded in human dyadic interactions. Participants were often uncertain whether they interacted with human or artificial investors, highlighting the agents' realism. This validates the use of these artificial agents to probe the effectiveness of interventions whilst keeping a high degree of experimental control. Following the exposure intervention, participants reduced their returns overall. When breaking down the groups into RS subgroups, the reduction in returns was mostly from the group with low Rejection Sensitivity. The returns of those in the control group did not change between the pre and post phase of the experiment. Why did participants reduce their returns even though they were repeatedly exposed to a more cooperative and more forgiving AI? A look at how the participants rated their co-players might shed some light on what is driving this reduction in returns for those exposed to the forgiving AI.

Those exposed to the forgiving AI rated their opponent in the post-exposure phase lower on all attributes even though they faced the same dynamic human-like HMM as pre-exposure. Indeed, the biggest ratings change in the manipulation group was for the "post" player which were lower than ratings for both the pre-exposure and exposure players. This is likely due to a *negative contrast effect.* The Contrast Effect occurs when the evaluation of a person, object, or situation is influenced by comparisons with recently encountered contrasting objects or people. If we've recently interacted with someone exceptionally nice, our perception of a normal level of niceness might be skewed, making normal behavior seem less favorable or even negative by comparison (Kobre and Lipsitt 1972). As the most recently faced opponents were highly cooperative, this negative contrast effect may have compensated any learning transfer from being repeatedly exposed to cooperative and forgiving AI (Zentall 2005). If this contrast effect is indeed replicable, then an avenue for future research would be to use it to our benefit by making the participants play agents with low cooperation perception. Participants also perceived the forgiving AI during the exposure phase as more cooperative, but not more forgiving. The latter might be the result of the participants not having many opportunities to test the co-player's forgiveness propensity since most decided to continue cooperating.

4.1 Other notable points:

Those in the control group perceived the human-like HMM as less cooperative, forgiving and were less willing to face it after the first interaction. This might indicate either *satiation/hedonic adaptation* (the process by which a novel interaction becomes less enjoyable after too much exposure), or a *negativity bias* (humans pay more attention to and give more weight to negative rather than positive experiences. Over time, as we interact more with someone, we might start noticing more of their flaws or negative traits, which could lead to lower overall ratings or more negative judgements). This negative rating did not translate however into lower proportioanl returns in the post-exposure phase.

The study found that RSQ (Rejection Sensitivity Questionnaire) and LPFS (Levels of Personality Functioning Scale) scores were ineffective at distinguishing between high and low performers in the game. This gap highlights a need for further research, despite theoretical connections suggested in the literature review.

References

Almaatouq, Abdullah, Joshua Becker, James P. Houghton, Nicolas Paton, Duncan J. Watts, and Mark E. Whiting. 2021. "Empirica: A Virtual Lab for High-Throughput Macro-Level Experiments." *Behavior Research Methods* 53 (5): 2158–71. https://doi.org/10.3758/s13428-020-01535-9.

Bandura, Albert. 1977. *Social Learning Theory.* Prentice Hall.

Bowlby, John. 1978. "Attachment Theory and Its Therapeutic Implications." *Adolescent Psychiatry* 6: 5–33.

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.

Downey, Geraldine, Hala Khouri, and Scott I. Feldman. 1997. "Early Interpersonal Trauma and Later Adjustment: The Mediational Role of Rejection Sensitivity." In *Developmental Perspectives on Trauma: Theory, Research, and Intervention*, 85–114. Rochester Symposium on Developmental Psychology, Vol. 8. Rochester, NY, US: University of Rochester Press.

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.

Fowler, James H., and Nicholas A. Christakis. 2010. "Cooperative Behavior Cascades in Human Social Networks." *Proceedings of the National Academy of Sciences* 107 (12): 5334–38. https://doi.org/10.1073/pnas.0913149107.

Gao, Shuling, Mark Assink, Andrea Cipriani, and Kangguang Lin. 2017. "Associations Between Rejection Sensitivity and Mental Health Outcomes: A Meta-Analytic Review." *Clinical Psychology Review* 57 (November): 59–74. https://doi.org/10.1016/j.cpr.2017.08.007.

Herpertz, Sabine C., and Katja Bertsch. 2014. "The Social-Cognitive Basis of Personality Disorders." *Current Opinion in Psychiatry* 27 (1): 73–77. https://doi.org/10.1097/YCO.0000000000000026.

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

King-Casas, Brooks, Damon Tomlin, Cedric Anen, Colin F. Camerer, Steven R. Quartz, and P. Read Montague. 2005. "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange." *Science* 308 (5718): 78–83. https://doi.org/10.1126/science.1108062.

Kobre, Kenneth R, and Lewis P Lipsitt. 1972. "A Negative Contrast Effect in Newborns." *Journal of Experimental Child Psychology* 14 (1): 81–91. https://doi.org/10.1016/0022-0965(72)90033-1.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. "Balancing Type I Error and Power in Linear Mixed Models." *Journal of Memory and Language* 94 (June): 305–15. https://doi.org/10.1016/j.jml.2017.01.001.

Mikulincer, Mario. 1998. "Attachment Working Models and the Sense of Trust: An Exploration of Interaction Goals and Affect Regulation." *Journal of Personality and Social Psychology* 74 (5): 1209–24. https://doi.org/10.1037/0022-3514.74.5.1209.

Mulder, R. T., P. R. Joyce, P. F. Sullivan, C. M. Bulik, and F. A. Carter. 1999. "The Relationship Among Three Models of Personality Psychopathology: DSM-III-R Personality Disorder, TCI Scores and DSQ Defences." *Psychological Medicine* 29 (4): 943–51. https://doi.org/10.1017/S0033291799008533.

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

Zentall, Thomas R. 2005. "A Within-trial Contrast Effect and Its Implications for Several Social Psychological Phenomena." *International Journal of Comparative Psychology* 18 (4). https://doi.org/10.46867/ijcp.2005.18.04.08.