When Forgiveness Backfires: Rejection Sensitivity and Cooperative Behavior Following Exposure to Adaptive Forgiving Agents

Ismail Guennouni[1,2,3,4*], Georgia Koppe[1,2,3†], Christoph Korn[4†]

[1] *Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*

[2] *Interdisciplinary Center for Scientific Computing, Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany*

[3] *Hector Institute for AI in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim Germany*

[4] *Department of General Psychiatry, Section Social Neuroscience, Heidelberg University, Germany*

[†] *Joint last author*

[*] *Corresponding author. Address: Central institute of Mental Health, J5, Mannheim, Germany. Email: ismail.guennouni@zi-mannheim.de. ORCID: 0000-0002-1096-4714*

Abstract:

Can exposure to forgiving partners improve interpersonal cooperation? Attachment theory suggests positive relational experiences can correct negative internal working models, but individuals high in Rejection Sensitivity (RS)—characterized by anxious expectations of rejection—may be resistant to such corrective experiences due to stable negative expectations. We tested this using a randomized experiment (N = 206) in which participants played repeated trust games with HMM-based artificial agents that simulate human-like trust dynamics. After a baseline game, participants were exposed to either forgiving agents with no pre-programmed trust withdrawal (Manipulation) or human-like agents maintaining typical trust violation patterns (Control), then played a final game with a standard agent. Overall, forgiveness exposure *reduced* subsequent cooperation—participants appeared to perceive the standard post-exposure agent as less cooperative by comparison (a negative contrast effect). Critically, RS moderated specific behavioral patterns but not overall cooperation levels: high RS participants failed to recover cooperation after trust violations and became *less* responsive to partner behavior following exposure, whereas low RS participants showed normal recovery and became *more* responsive. These findings suggest that positive relational experiences do not universally promote cooperation, and that high RS individuals may require interventions targeting their capacity to update expectations rather than simply providing positive experiences.

Keywords: Interpersonal functioning; Rejection Sensitivity; Forgiveness Intervention; Trust-based Cooperation; Hidden Markov Models

General Scientific Summary:

This study found that exposing people to forgiving partners (with no trust withdrawal) in economic games decreased their subsequent cooperation—likely because human-like partners seemed less cooperative by comparison. Individuals high in rejection sensitivity showed a distinct pattern: while they detected trust violations as readily as others, they failed to restore cooperation afterward and became less responsive to their partner's behavior. In contrast, those low in rejection sensitivity appeared to learn from the positive exposure and became more reciprocal. These findings suggest that simply providing positive social experiences may not benefit everyone equally, and that individuals prone to rejection sensitivity may need targeted support to translate positive experiences into lasting behavioral change.

1 Introduction

Trust is fundamental to human social interactions, facilitating seamless relations 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 & Campbell, 2017). Individuals with personality disorders (PD) often struggle to form and maintain social connections, a difficulty reflected in uncooperative behaviors – a marker for the severity of PD symptoms (Herpertz & 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). These insecure attachment patterns are often associated with heightened rejection sensitivity (RS), a tendency to anxiously expect, readily perceive, and intensely react to rejection (Downey et al., 1997; Downey & Feldman, 1996). RS has been linked to the development of various mental health conditions, including depression, anxiety, personality disorders, and self-harm (Gao et al., 2017). Individuals high in RS show attentional biases towards social threat cues, which may contribute to difficulties in social interactions (Berenson et al., 2009). A recent meta-analysis revealed prosocial behavior and interpersonal trust as two key processes of interpersonal functioning that are markedly impaired in PDs and which are likely to contribute to interpersonal dysfunction in this population (Hepp & Niedtfeld, 2022). The interaction of RS and trust-based constructs has been explored, particularly in Borderline Personality Disorder (BPD). Miano et al. (2013) and Richetin et al. (2018) found that RS mediated the relationship between BPD features and lower trust appraisal. Abramov et al. (2022) found that higher baseline feelings of rejection in individuals with BPD predict slower trust formation and less pronounced declines in trust following trust violations during the trust game. However, the interaction between *reciprocity* and RS hasn't been studied as extensively, leaving a gap in our understanding of how these constructs might interplay.

Given that RS may be a manifestation of maladaptive attachment styles, it is important to explore whether exposure to consistently forgiving and reliable interaction partners could reshape interpersonal expectations and behaviors. The *corrective experience hypothesis*, rooted in attachment theory, suggests that new positive relational experiences can modify internal working models of relationships (Bowlby, 1988). Research on social learning (Bandura, 1977) similarly demonstrates that individuals model the behavior of those around them, and exposure to cooperative peers promotes cooperative behavior (Fowler & Christakis, 2010). In the repeated trust game (RTG) paradigm, cycles of reciprocated trust enhance cooperative behaviors even among initially distrustful individuals (King-Casas et al., 2005). This perspective predicts that exposure to forgiving partners should increase subsequent cooperation, as participants internalize more positive expectations about social interactions.

However, an alternative perspective suggests high RS individuals may be *resistant* to such corrective experiences. RS is characterized by stable negative expectations that operate through self-fulfilling prophecies (Downey & Feldman, 1996)—high RS individuals interpret ambiguous social cues negatively, which elicits rejection, thereby confirming their expectations. Research on belief updating in depression and personality pathology has documented "cognitive immunization" processes whereby negative schemas resist modification despite contradictory evidence (Kube et al., 2020). From this perspective, positive exposure might fail to update expectations in high RS individuals, or might even produce paradoxical effects if the contrast between positive exposure and subsequent "normal" interactions confirms their belief that trustworthy partners are rare.

In this study, we use a randomized controlled online experiment to test whether exposing participants with varying RS levels to forgiving and more cooperative co-players results in more trustworthy behavior and a repair of potential breakdowns in RTG cooperation. To simulate realistic social interaction while maintaining a high degree of experimental control, we take a novel paradigmatic approach: We use generative models of how humans play the RTG to design an agent that plays the role of the investor, based on Hidden Markov Models (HMMs) fitted to real players' data. A key aspect of these agents is that their actions depend on a latent "trust state" which reacts dynamically to the trustees' returns, simulating real-life trust-building scenarios. An advantage of having such a generative model of behavior is the possibility of controlling different aspects of the agent's strategy such as its general policy, the propensity to cooperate actively, or the propensity to trust again after breakdowns of cooperation. To further mimic real-world interactions and examine participants' responses to one-off breakdowns of cooperation, we incorporate occasional pre-programmed low investments by the agent.

We pre-screened participants for high or low RS using a validated questionnaire, then assigned them exclusively to the trustee role in a series of trust games. After playing a 15-round RTG with a human-like HMM investor, they were randomly assigned to either a Control or Manipulation condition. In the Manipulation condition, participants were exposed over three RTGs to HMM investors designed with a limited propensity for retaliation—agents that were both forgiving of low returns AND free from pre-programmed trust violations, providing a consistently positive relational experience. In the Control condition, participants played three RTGs against the same human-like HMM that maintained the occasional low-investment pattern from the pre-exposure phase, representing continuity with typical social interactions. This design tests whether exposure to partners combining forgiveness with behavioral consistency transfers to subsequent interactions with standard partners. After this exposure phase, all participants played another 15-round RTG with a human-like HMM investor, similar to the one in the pre-exposure phase.

Based on the corrective experience hypothesis, we predicted that forgiveness exposure would increase subsequent cooperation, with high RS individuals potentially benefiting from positive relational experiences that challenge their negative expectations. We examined both overall effects of the manipulation and differential responses based on RS, with particular attention to how participants respond to trust violations before and after the exposure phase.

2 Methods

2.1 Participants

To have participants with large differences in RS, a total of 1195 participants were pre-screened on the Prolific Academic platform (prolific.co) using the Rejection Sensitivity Questionnaire (RSQ) to finally select two similarly sized groups: One with high RS (RSQ score $>$ 15, N=103) and the other with low RS (RSQ score $<$ 10, N=103) totalling 206 participants (56% female). These were then invited through prolific to take part in the main experiment. The required sample size was determined using an *a priori* power analysis to have an 80% probability to detect a small effect size (Cohen's f = 0.10) for a within-between interaction with a 5% type I error rate in a repeated measures ANOVA. The sample size calculation assumed two groups, two measurements per group and was performed using the G*Power software (Faul et al., 2009). The mean age of participants was 34.6 years, with an 11.9 years standard deviation. The majority of participants identified ethnically as White (80%). The online cohort registered 30 unique countries of birth with the most frequent being the U.K (33%) followed by Poland (10%) and Portugal (10%). Participants were paid a fixed fee of £6 plus a bonus payment dependent on their performance that averaged £0.5. Data was collected over multiple sessions between December 2023 and February 2024.

2.2 Design and procedure

The experiment had a 2 (Condition: Manipulation or Control) by 2 (RS : High or Low) by 2 (Phase: Trust-Game Pre-Exposure, Trust-Game Post-Exposure) design, with repeated measures on the Phase factor (Figure 1.A). Participants within each pre-screened RS group were randomly assigned to one of the two levels of the Condition factor, resulting in 101 participants in the Manipulation condition and 105 in the Control condition. The games were designed and implemented online using Empirica v1 (Almaatouq et al., 2021). The planned experiment received approval from the University of Heidelberg's Medical Faculty ethics commission (ID:S-708/2023) and the experiment was performed in accordance with the ethics board guidelines and regulations. All participants provided informed consent prior to their participation.

2.3 Tasks and measures

2.3.1 Repeated trust game and HMM investor

Participants played a 15-round RTG (Joyce et al., 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. 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. We chose to simulate the behavior of a human interaction partner through allowing for a delay whilst pairing with new opponents as the start of each game as well as programming the agents to respond during each round after a varying time lapse (randomly chosen between 5 and 10 seconds).

The computerised investor consisted of a hidden Markov model (HMM) trained on an independent existing behavioral RTG data set of human investors. This data-driven approach thus sought to learn an investor strategy that mimics human-like interactions. The data set used for training consists of 388 ten round games with the same player (full details can be found in the Supplementary Information). On this data set, the HMM was inferred with three latent states that could be interpreted as reflecting a “low-trust”, a “medium-trust”, and a “high-trust” state. A separate output distribution, that maps each HMM state onto possible investments from 0 to 20 separately, is learned (Figure 1.B). In analogy to the latent states, these distributions can be interpreted as reflecting “low-trust”, “medium-trust”, or “high-trust” dispositions. Finally, the HMM is specified by transition probabilities that describe the transition between states. 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)). The initial state for the HMM investor in each instance of the game was set to the “mid-trust” state. Details on how the HMM state conditional probabilities and transition functions are specified can be found in the supplement.

In order to instigate a potential breakdown of trust, thereby allowing us to probe efforts to repair it, the computerised agent was programmed to provide a low investment on round 12 (pre-exposure) and round 13 (post-exposure). 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).

2.4 Manipulation

In all phases of the RTG other than the ‘Exposure phase’ (Figure 1.A), participants interacted with this human-like HMM. In the ‘Manipulation’ Condition of the exposure phase, however, the parameters of this HMM were adjusted to design a ‘forgiving’ and ultimately more cooperative agent. To achieve this, we changed the state transition probabilities of the HMM such that it becomes impossible for it to remain in a low trust state, effectively setting the transition probability for remaining in a “low-trust” state to 0. The resulting transition function is shown in Figure 1.D. The policies conditional on the latent states and the transition function in the other latent states remain unchanged.

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. Participants were randomly assigned to either a Control or Manipulation condition. The timeline of the experiment is shown in Figure 1.A. Phase one (“pre”) consisted of a 15 round 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. Phase 2 (“exposure”) consisted of three seven-round RTGs. Participants in the Manipulation condition faced the forgiving HMM investor and rated the agent on the same attributes as in the pre-exposure phase. Those in the Control condition faced the same human-like HMM agent as in the “pre” phase and rated each co-player on the same attributes. To keep the experience similar to the “pre” phase, the agent in the Control condition was also designed to send a very low investment in round 5 of each of the three games. In the post-exposure phase (“post”), participants in both conditions faced the same human-like HMM as in “pre” phase.

At the beginning of each game in all three phases, participants were told they would face a new player and had to wait to be paired with an available co-player. This simulated the waiting time in real social interaction tasks. After completing each RTG in each phase, participants rated how cooperative and forgiving they perceived the co-player to be, and whether they would like to play with them again (all on a scale from 1 to 10 with 10 being the most positive rating). After completing the three game phases, participants then completed the Levels of Personality Functioning Scale Brief-Form (LPFS-BF) questionnaire (Weekers et al., 2019). This is a self-report measure designed to assess core elements of personality functioning as defined in the Alternative Model for Personality Disorders in the DSM-5 (American Psychiatric Association, 2013), and provides a dimensional assessment of personality functioning, which complements the categorical approach of RS. Finally, participants were asked whether they thought the other players were human or computer agents, to probe how well the agent can mimic human behavior, then debriefed and thanked for their participation.

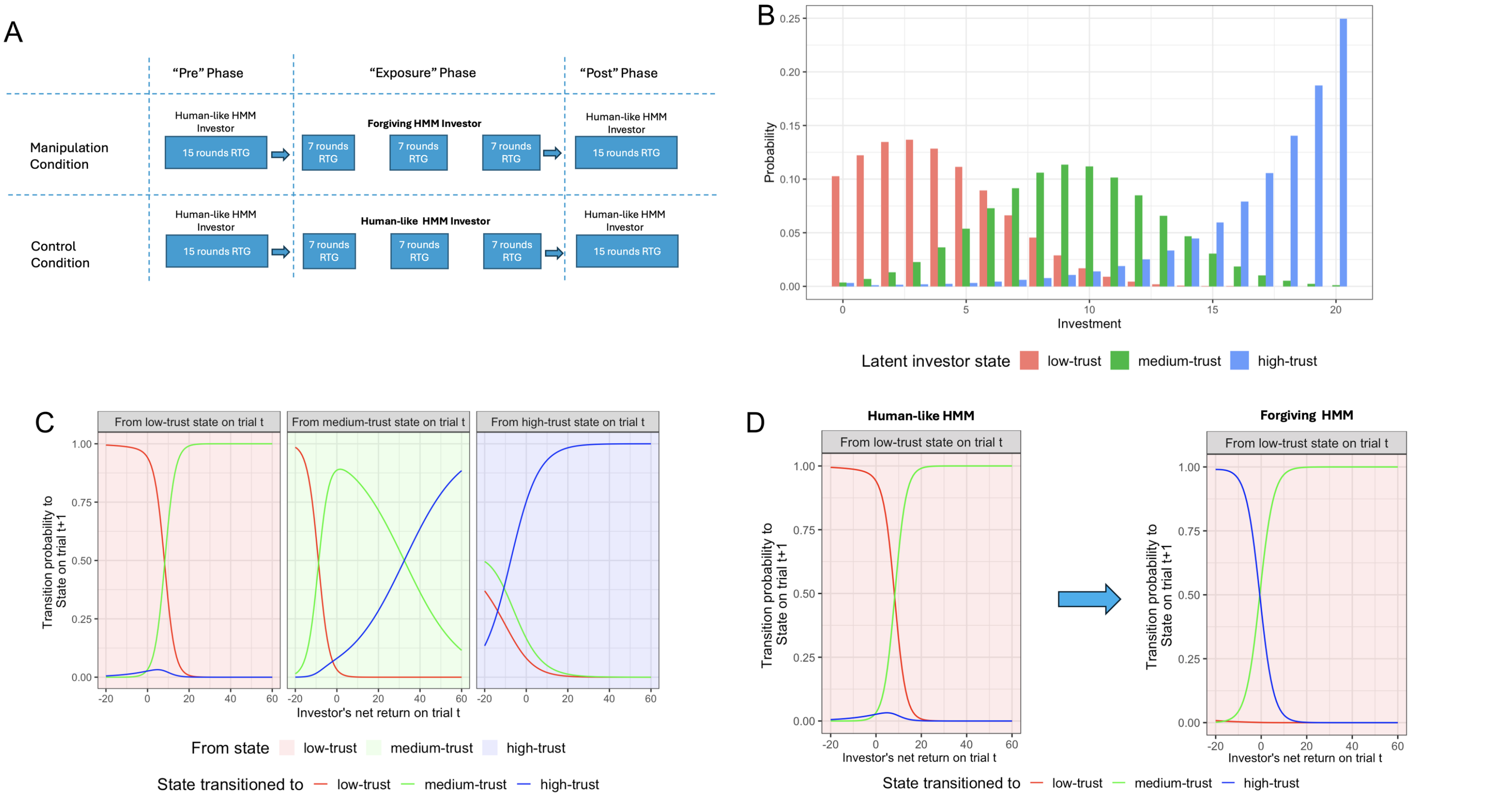

Figure 1: A: Experiment timeline. Participants (trustees) played RTGs with HMM investor agents. The investor sends investments (multiplied by 3) and participants decide returns. Conditions differ in exposure phase agents. B-D: The artificial investor is a three-state HMM fitted to human data. B: Investment distributions by latent state. C: Transition probabilities to states at t+1 as a function of net return at t; each panel shows a different starting state. D: Forgiving HMM transitions from low-trust state—unlike the human-like HMM, it always exits low-trust and favors high-trust transitions.

2.6 Statistical analysis

We analyzed participants' behavior in the RTG using linear mixed-effects models. First, to examine the effect of the manipulation, we modeled the percentage return (percentage of tripled investment returned to investor) as a function of Phase (RTG game pre vs. post-exposure), Condition (Manipulation vs. Control), Investment, and RS (High vs Low RS group), including all interactions as fixed effects. This model included player-wise random intercepts and slopes for Phase. Second, we analyzed behavior during the Exposure phase specifically, modeling returns with Condition, Investment, and RS and their interactions as fixed effects, along with player-wise random intercepts. Third, to verify the consistency of the HMM agent, we modeled the investments sent by the computerized agent using Condition, Phase, and RS and their interactions as fixed effects. To isolate effects occurring prior to any pre-programmed low investment, we also analyzed returns in rounds preceding the low investment trials only (rounds 1-11 in the pre-exposure phase and rounds 1-12 in the post-exposure phase) using the same model specification. To test whether reduced returns reflected strategic exploitation, we examined investor-harming returns (those below one-third of the tripled investment, which cause the investor to incur a net loss) using a mixed logistic regression with Phase, Condition, Investment, and RS as predictors, and player-wise random intercepts. Trustee payoffs were compared between phases using paired t-tests. Finally, to rigorously assess participants' reactions to and recovery from the specific instance of pre-programmed low investment, we conducted an event study analysis centered on the low investment round ($t = 0$). We analyzed percentage returns in a three-round window ($t - 1$ to $t + 1$) using a linear mixed-effects model with Phase, Condition, Time Point, and RS Group as fixed effects. We specifically examined two key behavioral responses: the Drop (change in return from $t - 1$ to the low investment round) and the Recovery (change in return from the low investment round to $t + 1$). The full specification of all statistical models can be found in the supplement.

All models were estimated using the `afex` package (Singmann et al., 2022) in R. We determined the random effects structure by starting with the maximal model and simplifying until convergence was achieved, ensuring the optimal structure (Matuschek et al., 2017). A similar process was applied to the models analyzing HMM agent investments and participant ratings. We report differences in marginal means rather than effect sizes, as there is no consensus on effect size calculation for mixed models. F-tests used the Kenward-Roger approximation for degrees of freedom. The Investment variable was Z-transformed to facilitate the interpretation of main effects in the presence of interactions. Significant interactions were probed using planned contrasts with the `emmeans` package. We applied the "Sidak" correction for multiple comparisons to control the familywise error rate while maintaining statistical power.

3 Behavioral Results

3.1 Analysis of participant returns

On average, investments and returns, as shown in Figure 2, 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 (Charness et al., 2008; Fiedler et al., 2011).

Participants returned higher percentages in the Pre phase compared to the Post phase ($F(1, 201.63) = 5.81$, $p = .017$). This effect was moderated by Condition ($F(1, 201.63) = 4.38$, $p = .038$): contrary to our expectations, participants in the Manipulation condition decreased their returns from pre to post ($\Delta M = 0.03$, 95% CI $[0.01, 0.05]$, $t(201.50) = 3.15$, $p = .002$), while those in the Control condition showed no change (Figure 3). RS did not moderate this Condition × Phase interaction.

Higher investments elicited higher percentage returns, indicating positive reciprocity ($F(1, 5955.67) = 325.35$, $p < .001$). This relationship was stronger in the Control condition than in the Manipulation condition ($F(1, 5955.67) = 13.92$, $p < .001$). The effect of investment on returns varied by RS group and Phase ($F(1, 5864.62) = 7.84$, $p = .005$), and a four-way interaction indicated that these patterns further differed across Conditions ($F(1, 5864.62) = 9.24$, $p = .002$).

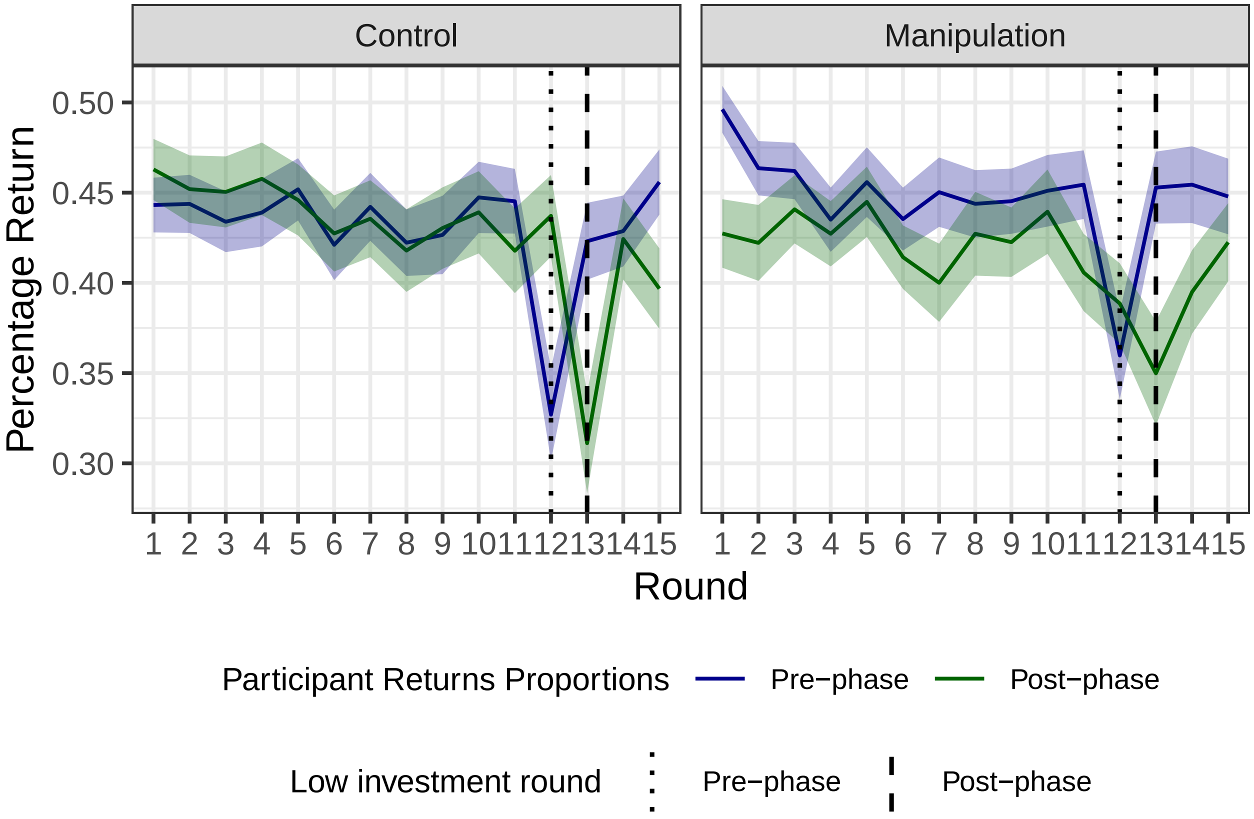

Figure 2: Averages and standard errors of the trustee's return as a percentage of the multiplied investment received (y-axis) by Condition, Phase, and game round (x-axis) averaged across RS groups. The blue line shows the returns in the Pre phase and the green line those in the Post phase. The left Panel shows returns in the Control condition and the right one those in the Manipulation condition. The dotted lines identify the rounds where the pre-programmed one-off low investment occurs. We note lower average returns post vs pre in the Manipulation condition, whilst returns in the Control condition are similar between the two phases.

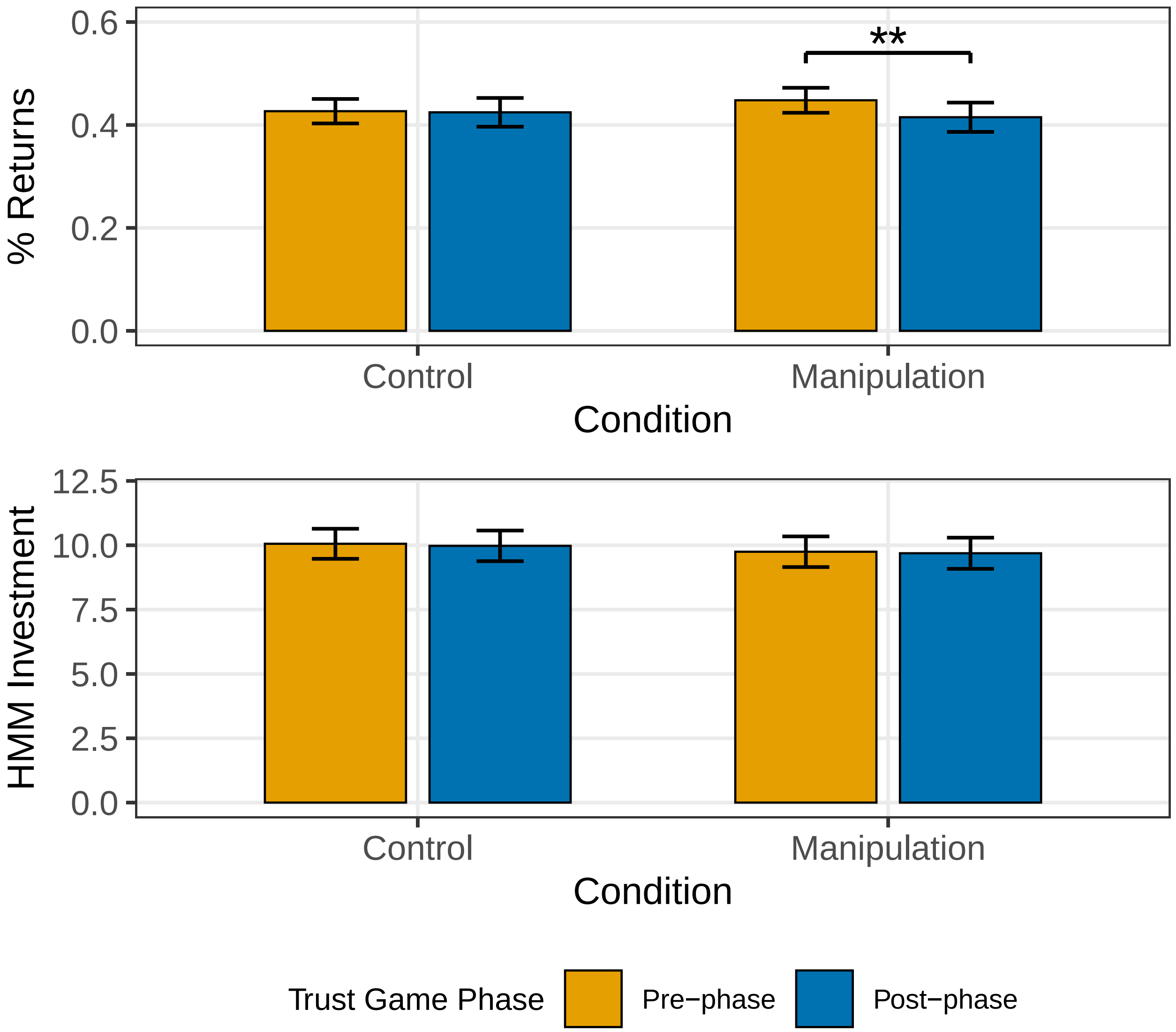

Figure 3: Marginal means of percentage returns (top) and HMM investments (bottom) by Phase and Condition. Bars show estimated marginal means; error bars represent 95% confidence intervals. Participants in the Manipulation condition returned lower proportions post-exposure compared to pre-exposure (** p $<$ .01), while Control participants showed no change. HMM investment did not differ across Phases or Conditions.

3.1.1 Exploitation diagnostic

A mixed logistic regression predicting exploitation probability (returns below one-third of the tripled investment) revealed that exploitation increased from pre to post in both conditions (main effect of Phase, $z = 3.18$, $p = .001$). The Phase × Condition interaction was not significant ($z = 1$, $p = .316$). The Manipulation condition showed lower exploitation rates than Control at both time points (pre: 14.9% vs. 17.9%; post: 20.4% vs. 21.8%). Trustee payoffs in the Manipulation condition did not significantly change from pre ($M = 15.7$) to post ($M = 16.2$; $t(100) =$ -1.04, $p =$ .299).

To examine this four-way interaction, we conducted a contrast analysis of how the effect of investment on returns changed from pre- to post-exposure for different RS groups in both conditions (Figure 4). Starting with the Manipulation condition, for participants with low RS, the effect of investment on returns increased significantly from pre- to post-phase, $\Delta M = 0.03$, 95% $\mathrm{CI}_{\mathrm{Sidak}(3)}$ $[0.01, 0.05]$, $t(5881.28) = 3.15$, $p_{\mathrm{Sidak}(3)} = .005$. This suggests that after the manipulation, low RS participants became more responsive to their co-player's investments, returning proportionally more as investments increased. In contrast, for participants with high RS, the effect of investment on returns decreased significantly from pre- to post-exposure, $\Delta M = -0.02$, 95% $\mathrm{CI}_{\mathrm{Sidak}(3)}$ $[-0.04, 0.00]$, $t(5891.81) = -2.67$, $p_{\mathrm{Sidak}(3)} = .023$. This indicates that high RS participants became less responsive to their co-player's investments after the manipulation, with smaller increases in returns as investments increased. The difference in these pre-post changes between high and low RS groups was significant, $\Delta M = -0.05$, 95% $\mathrm{CI}_{\mathrm{Sidak}(3)}$ $[-0.08, -0.02]$, $t(5887.33) = -4.11$, $p_{\mathrm{Sidak}(3)} < .001$. This result suggests that the manipulation had significantly different effects on how low and high RS participants responded to their co-player's investments.

In the Control condition, we observed no significant changes in how participants responded to their co-player's investments between the pre and post phases, regardless of their RS level.

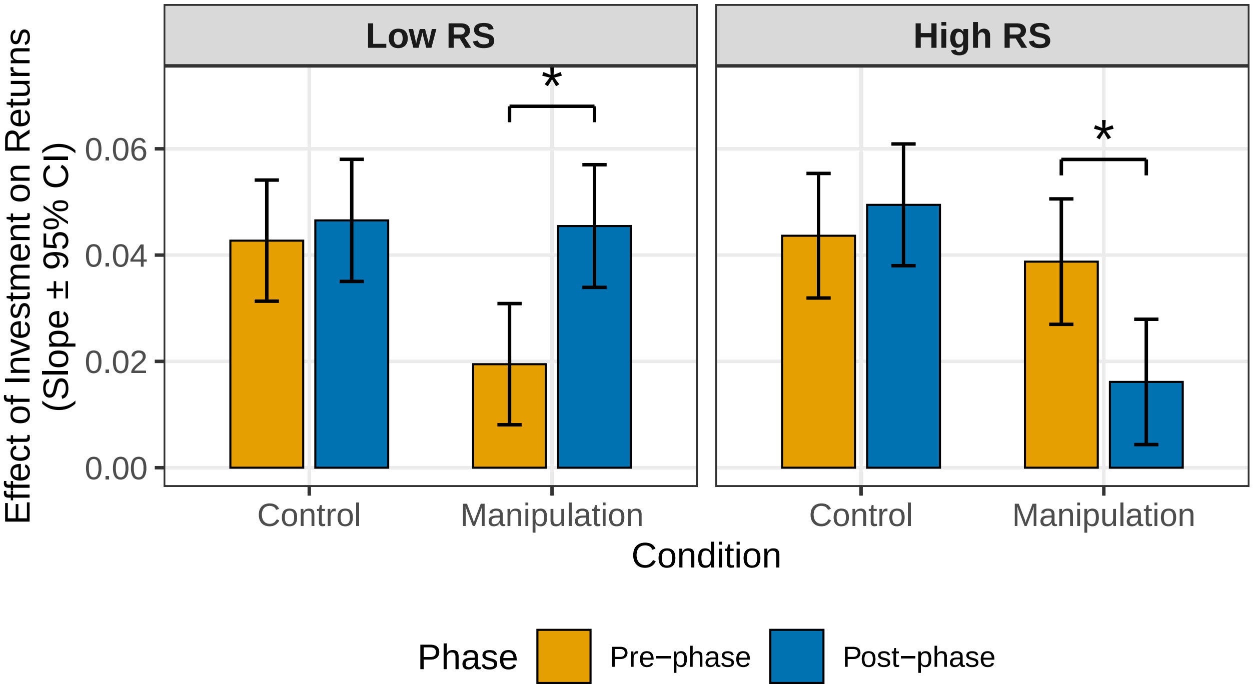

Figure 4: Marginal effect of investment on percentage returns by Phase, Condition, and RS group. Bars show estimated slopes (change in returns per SD increase in investment) from the mixed model; error bars represent 95% confidence intervals. In the Manipulation condition, Low RS participants became more responsive to investments post-exposure (* p < .05), while High RS participants became less responsive (* p < .05). No significant changes were observed in the Control condition.

3.1.2 Returns prior to pre-programmed low investment trials

To distinguish between contrast effects and betrayal aversion as explanations for reduced cooperation in the Manipulation condition, we examined returns in the rounds preceding the pre-programmed low investment. If contrast effects were operating, participants in the Manipulation condition should already show reduced returns before experiencing any low investment in the post-exposure phase. Conversely, if betrayal aversion were the primary mechanism, group differences should only emerge after the low investment.

The Phase × Condition interaction was significant in rounds prior to the low investment ($F(1, 204.05) = 4.86$, $p = .029$). Participants in the Manipulation condition significantly decreased their returns from pre- to post-exposure phase even before encountering the low investment ($\Delta M = 0.03$, 95% CI $[0.01, 0.06]$, $t(202.02) = 2.83$, $p = .005$), whereas those in the Control condition showed no change ($\Delta M = 0.00$, 95% CI $[-0.02, 0.02]$, $t(201.17) = -0.15$, $p = .878$). The four-way interaction also remained significant ($F(1, 4454.34) = 7.12$, $p = .008$), suggesting that the differential responsiveness to investments observed in the full analysis was likewise present before the low investment occurred.

3.1.3 Reaction to pre-programmed low investment: event study analysis

To understand how participants reacted to and recovered from the pre-programmed low investment, an event study analysis was conducted centered on the low investment round (Figure 5). Two behavioral responses were examined: the Drop (change in returns at the low investment round relative to $t-1$, where negative values indicate reduced returns reflecting punishment of low trust) and Recovery (change in returns at $t+1$ relative to the low investment round, where positive values indicate restored cooperation).

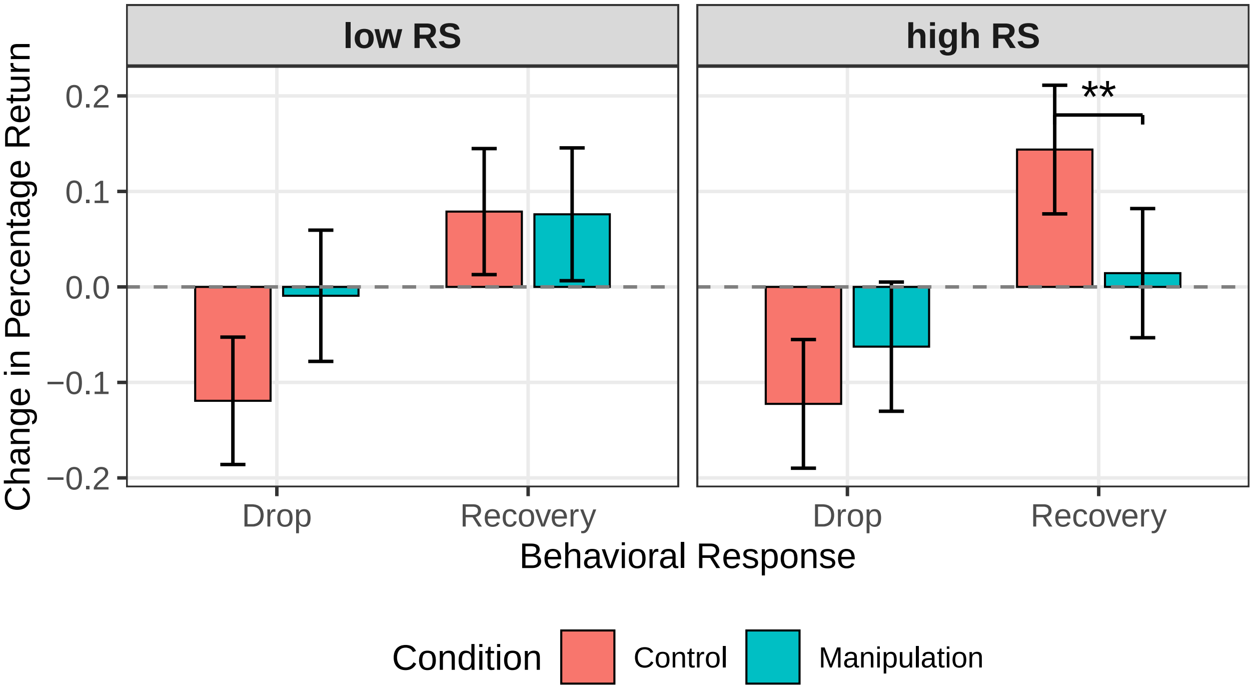

Figure 5: Drop and Recovery responses to pre-programmed low investment in the post-exposure phase (pre-exposure phase not shown as no between-condition differences were observed). Drop = change in returns from low investment round minus t-1 (negative values indicate reduced returns); Recovery = change from t+1 minus low investment round (positive values indicate restored returns). Bars show estimated marginal means from the mixed model; error bars represent 95% confidence intervals. The bracket shows the significant between-condition difference in Recovery for High RS participants. ** p < .01.

In the pre-exposure phase, there were no significant differences between conditions in either the Drop ($M = -0.02$, 95% CI $[-0.09, 0.05]$, $t(1773.82) = -0.61$, $p = .543$) or Recovery ($M = 0.00$, 95% CI $[-0.06, 0.07]$, $t(1773.56) = 0.11$, $p = .914$), confirming that both groups started with equivalent behavioral patterns. In the post-exposure phase, a divergence emerged. The Control group showed a significantly larger Drop than the Manipulation group ($M = -0.08$, 95% CI $[-0.15, -0.02]$, $t(1773.93) = -2.46$, $p = .014$), indicating that participants exposed to the forgiving agent showed a blunted immediate reaction to the low investment and did not reduce their returns as sharply. The subsequent Recovery did not differ significantly between conditions overall ($M = 0.07$, 95% CI $[0.00, 0.13]$, $t(1774.02) = 1.92$, $p = .055$).

When examining moderation by RS, the pattern appeared to be driven primarily by high RS participants. Low RS participants showed no significant difference in Recovery between conditions ($M = 0.00$, 95% CI $[-0.09, 0.10]$, $t(1774.24) = 0.06$, $p = .954$), with both groups displaying modest, similar recovery patterns after the low investment. The larger Drop observed in the Control condition for this group was partly attributable to their elevated cooperation level at $t-1$.

High RS participants showed a different pattern. In the Control condition, they demonstrated trust repair by significantly increasing their returns after the low investment ($\Delta M = 0.14$, 95% CI $[0.08, 0.21]$, $t(1774.05) = 4.19$, $p < .001$). However, those in the Manipulation condition failed to recover, showing no significant increase in returns at $t+1$ ($\Delta M = 0.01$, 95% CI $[-0.05, 0.08]$, $t(1773.54) = 0.42$, $p = .676$). The difference in Recovery between conditions was significant ($M = 0.13$, 95% CI $[0.03, 0.22]$, $t(1773.79) = 2.66$, $p = .008$).

In summary, the forgiveness intervention appeared to dampen reciprocal responsiveness, hindering the re-establishment of cooperation following a temporary withdrawal of trust. This effect was more pronounced among high RS individuals. While high RS participants in the Control condition demonstrated active reciprocity by reducing returns sharply when trust was withdrawn and increasing them when trust was restored, those exposed to the forgiving agent exhibited a disengaged pattern characterized by a muted reaction to the low investment and a failure to reinstate high returns afterward.

3.1.4 HMM investor in pre and post phases

Was the HMM's strategy similar between pre and post phases in the control condition? Was participants' behavior post exposure differentiated enough to induce a different reaction from the HMM? To answer these questions, we test for differences in the HMM agent's investment by Phase, Condition and RS using a linear mixed-effects model as described in the methods section. As seen in Figure 3, we find no main or interaction effects, indicating the HMM's behavior was on aggregate similar across levels of Phase, Condition and RS. This consistency in the investor's behavior is a desirable feature of the HMM agent when the participants' behavior is largely similar between phases. More importantly, it indicates that the lower returns of participants in the post phase of the manipulation condition were not differentiated enough to make the HMM react by transitioning to lower latent trust states. It is also noteworthy that the HMM agent was relatively successful in imitating human behavior in this paradigm: 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.

3.1.5 Exposure phase trials

So far we focused on analysing behavior for the pre and post phases. Here, we look at returns and investments in the exposure phase. The linear mixed effects model of participants' returns in the exposure phase does not show a main effect of Condition on returns. There was a main effect of Investment, $F(1, 4117.20) = 233.19$, $p < .001$, with participants positively reciprocating higher investments, an interaction effect between Condition and Investment $F(1, 4117.20) = 45.93$, $p < .001$, showing a stronger positive reciprocity in the Control condition, and finally a three way interaction between the RS group, Condition and Investment $F(1, 4117.20) = 4.21$, $p = .040$, showing that this stronger positive reciprocity to investment in the Control condition is higher for participants with high RS. The linear mixed effects model of the HMM investments shows a main effect of Condition $F(1, 202) = 197.64$, $p < .001$, suggesting higher overall investments for the forgiving HMM compared to the human-like HMM, but no difference in investments when facing low and high RS groups.

In summary, despite the forgiving HMM sending overall higher investments in the exposure phase, participants returned similar proportions of the multiplied investments as those facing the human-like HMM. The positive reciprocity of returns to investments was higher in the Control condition with this relationship stronger for the high RS group.

3.1.6 Questionnaire scores and performance

Whilst we found a significant correlation between participant's Levels of Personality Functioning Score (LPFS) and the Rejection Sensitivity Questionnaire score (RSQ), Spearman's $r_s = .52$, $p < 0.001$, there was no correlation between these questionnaire scores and participant's return or overall task performance.

3.2 Player ratings

Figure 6 shows participants' ratings of co-players across phases. We examined two contrasts: pre-exposure versus exposure phase ratings, and pre-exposure versus post-exposure ratings.

High RS participants showed more differentiated perceptions of the agents. In the Manipulation condition, they rated the forgiving agents as more cooperative during exposure ($\Delta M = 2.57$, 95% CI $[0.84, 4.30]$, $t(808) = 2.91$, $p = .004$). In the Control condition, however, high RS participants rated the same human-like HMM progressively more negatively—lower on cooperation ($\Delta M = -2.65$, 95% CI $[-4.37, -0.94]$, $t(808) = -3.04$, $p = .002$), forgiveness ($\Delta M = -2.19$, 95% CI $[-4.00, -0.39]$, $t(808) = -2.38$, $p = .017$), and willingness to play again ($\Delta M = -3.62$, 95% CI $[-5.73, -1.50]$, $t(808) = -3.36$, $p < .001$)—despite the agent's strategy remaining unchanged. Low RS participants showed largely undifferentiated perceptions between pre and exposure phases regardless of condition.

Comparing pre to post-exposure ratings revealed a contrast effect: after experiencing the forgiving agent, both RS groups in the Manipulation condition rated the post-exposure agent (identical to pre-exposure) more negatively on forgiveness (High RS: ΔM = -0.88, SE = 0.38, $t(808.0)$ = -2.33, p = .020; Low RS: ΔM = -1.14, SE = 0.38, $t(808.0)$ = -2.98, p = .003) and willingness to play again (High RS: ΔM = -1.29, SE = 0.44, $t(808.0)$ = -2.92, p = .004; Low RS: ΔM = -1.30, SE = 0.45, $t(808.0)$ = -2.90, p = .004). Low RS participants in the Control condition showed stable ratings, accurately perceiving the consistent agent strategy, while high RS participants in the Control condition continued their negative drift (Cooperation: ΔM = -0.96, SE = 0.36, $t(808.0)$ = -2.70, p = .007). These rating patterns converge with the behavioral findings, suggesting high RS individuals are particularly sensitive to relative comparisons between interaction partners.

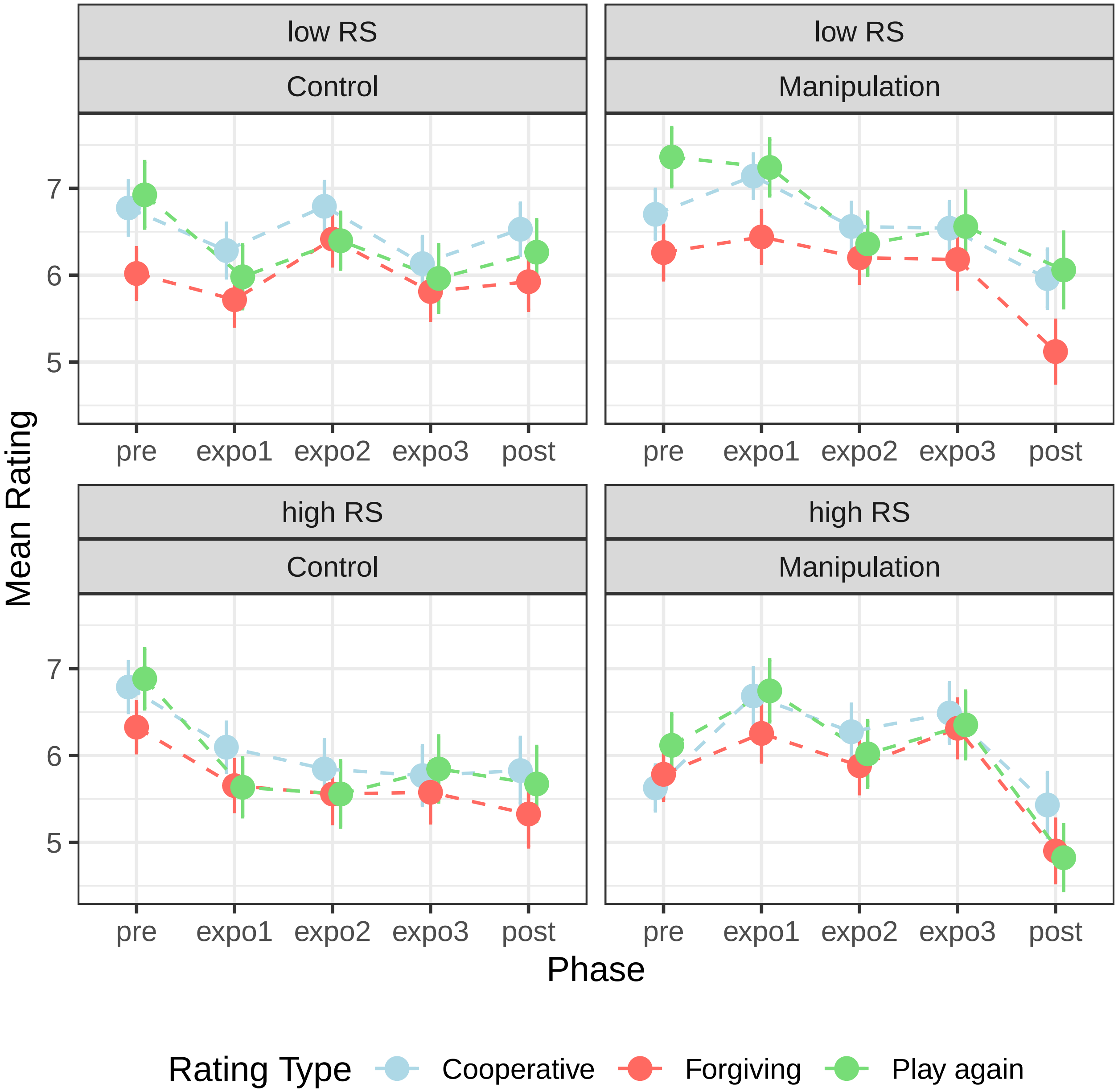

Figure 6: Participants' ratings of co-players by phase, condition, and RS group. Blue: perceived cooperation; red: perceived forgiveness; green: willingness to play again. Low RS participants showed stable ratings in the Control condition. High RS participants showed declining ratings in the Control condition despite unchanged agent strategy, and more differentiated perceptions in the Manipulation condition.

4 Discussion

We used a randomized controlled online experiment where participants played a RTG with artificial agents designed to simulate human-like trust-building scenarios. Participants were then exposed to either forgiving HMM agents (which, by design, were also more cooperative due to their inability to remain in a low-trust state) or standard human-like HMM agents before playing another RTG. We found that RS did not moderate participants' returns as trustees in the repeated trust game. While previous research has shown that RS affects *trust* formation, appraisal and repair, its impact on *reciprocity* in repeated economic exchanges has been less explored. Our results suggest a potential dissociation between RS's known effects on broader social behavior and its limited influence on reciprocity in structured, repeated interactions, challenging assumptions about the pervasive influence of RS on social behavior and highlighting the complexity of factors influencing reciprocity in economic exchanges.

Contrary to our hypothesis, exposure to forgiving agents did not increase participant's reciprocity or cooperation, nor did it prompt the artificial agent to increase their trust in participants through higher investments. Instead, participants reduced their returns overall whilst 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 agents designed to be more forgiving? A look at how the participants rated their co-players might shed some light on what might be driving this reduction in returns for those in the Manipulation condition. Those exposed to the forgiving agent 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. One possible explanation for this drop in rating is that participants exhibited a negative contrast effect. This occurs when the evaluation of a person, object, or situation is influenced by comparisons with recently encountered contrasting objects or people. If we've repeatedly interacted with someone exceptionally nice, our perception of a normal level of niceness might be skewed, making typical behavior seem less favourable or even negative by comparison (Kobre & Lipsitt, 1972). As the most recently faced opponents were more forgiving (and consequently more cooperative), this negative contrast effect may have trumped any learning transfer from being repeatedly exposed to forgiving agents (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.

The design deliberately created two qualitatively different exposure experiences: Manipulation participants interacted with partners who were both forgiving (unable to remain in low-trust states) AND behaviorally consistent (no pre-programmed trust violations), while Control participants interacted with partners who maintained the same pattern of occasional low investments seen in the pre-exposure phase. This design reflects the multidimensional nature of secure relational experiences in attachment theory—a secure base provides both responsiveness to distress (forgiveness) and consistent availability (Bowlby, 1988). The Control condition thus represents continuity with typical relationship patterns, while the Manipulation condition tests whether exposure to an idealized partner—one who is positive on both dimensions—produces lasting change.

One might argue that the Manipulation condition's absence of trust violations during exposure could produce heightened betrayal aversion when participants later encountered the post-phase low investment. However, the analysis of returns prior to the pre-programmed low investment provides evidence against this account. If betrayal aversion were driving the effect, group differences should only emerge after participants encountered the low investment in the post-exposure phase. Instead, Manipulation participants had already significantly reduced their returns in rounds 1-12 of the post-exposure phase—before any low investment occurred. This pattern is consistent with contrast effects operating from the beginning of the post-exposure phase, as participants immediately perceived the human-like agent as less cooperative compared to the forgiving agent they had just experienced. While betrayal aversion may contribute to specific aspects of the observed patterns, such as the impaired recovery following the low investment in high RS participants, it cannot account for the overall reduction in cooperation that was already evident before any low investment.

A third alternative interpretation is that exposure to forgiving agents reduced deterrence, promoting strategic exploitation (Thielmann et al., 2020). From this perspective, participants in the Manipulation condition may have learned during exposure that they could return less without consequence, then carried this exploitation strategy forward. However, the exploitation diagnostic analysis argues against this account. If participants were strategically exploiting, we would expect an increase in investor-harming returns (those below the one-third threshold that cause the investor to incur a loss) and higher trustee payoffs in the Manipulation condition. Instead, the Manipulation condition showed *lower* exploitation rates than Control at both time points, and the Phase × Condition interaction for exploitation probability was not significant. Furthermore, trustee payoffs did not increase in the Manipulation condition. This pattern is inconsistent with learned exploitation: participants were not extracting more resources or causing greater harm to their partners. Rather, the reduced returns appear to reflect altered perception of partner cooperativeness, as evidenced by the decline in explicit ratings of the post-exposure agent. The trustee role in the trust game is also fundamentally reactive—trustees respond to investments already received—making proactive exploitation tendencies less relevant than in games where participants initiate exchanges (Thielmann et al., 2020).

While the negative contrast effect operated across RS groups, the pattern of responses to trust violations differed in ways that align with clinical models of rejection sensitivity. In the event study analysis, high and low RS participants showed comparable immediate reactions to the low investment (the Drop), indicating intact detection of trust violations regardless of RS level. However, the groups diverged in their subsequent recovery patterns: high RS participants in the Manipulation condition failed to restore cooperation following the low investment, whereas low RS participants and Control participants showed recovery. This dissociation between intact rejection detection and impaired relationship repair is consistent with research on social learning difficulties in individuals with elevated RS and related clinical presentations. Studies of borderline personality disorder, where RS is characteristically elevated, have documented specific deficits in updating social expectations following positive interpersonal experiences (Schuster et al., 2021; Staebler et al., 2011). Similarly, research on depression has identified "cognitive immunization" processes whereby negative schemas resist modification despite contradictory evidence (Kube et al., 2020). The high RS participants' failure to recover cooperation, despite prior exposure to consistently forgiving behavior, may reflect analogous difficulties in leveraging positive social experiences to update expectations and restore trust.

The four-way interaction findings further support this interpretation. High RS participants showed decreased responsiveness to their co-player's investments following the forgiveness manipulation, a pattern suggestive of withdrawal from contingent social exchange. This reduced sensitivity to partner behavior parallels the self-silencing and social withdrawal documented in high RS populations, where anticipatory self-protection can paradoxically undermine relationship maintenance (Ayduk et al., 2000; Romero-Canyas et al., 2010). In contrast, low RS participants showed increased responsiveness to investments post-phase, suggesting they internalized the cooperative norms experienced during exposure and carried this forward to subsequent interactions. This differential capacity to benefit from positive social experiences maps onto broader findings that RS impedes the acquisition and transfer of adaptive interpersonal strategies (Pietrzak et al., 2005).

The combination of blunted responsiveness to investments and absent recovery in high RS participants suggests a pattern of passive disengagement rather than active retaliation. While overall return levels did not differ between RS groups, these specific behavioral signatures indicate that RS does modulate particular aspects of cooperative behavior. The ratings data complement these findings: high RS participants showed more negative explicit evaluations of their co-players, rating them lower on forgiveness and willingness to interact again. This convergence between explicit ratings and behavioral patterns suggests that the effects of RS on social exchange are expressed across multiple response systems. The structured nature of the trust game may constrain RS effects to specific behavioral signatures (such as contingent responding and recovery) rather than overall cooperation levels, while the more open-ended nature of rating tasks allows for broader expression of RS-related evaluative biases (Lieberman, 2007).

These findings have implications for interventions aimed at promoting trust and cooperation. The present results suggest that exposure to positive social models alone may be insufficient for high RS individuals, and may even produce iatrogenic effects through negative contrast. The specific deficits observed—impaired recovery from trust violations and reduced sensitivity to partner behavior—point to potential intervention targets. Approaches that focus on enhancing the capacity to update expectations following interpersonal ruptures may be more effective than simply providing positive experiences. This could include explicit training in recognizing repair attempts, practicing graduated trust restoration, or developing metacognitive awareness of the tendency toward disengagement following perceived rejection. Future research should examine whether these behavioral patterns generalize to naturalistic social contexts and whether targeted interventions can modify the updating and recovery deficits observed in high RS participants (Balliet et al., 2011). However, given the minimal nature of the trust game paradigm and the brief exposure period, these findings should be interpreted as proof-of-concept demonstrations rather than direct evidence for clinical intervention design. The corrective experience hypothesis in attachment theory typically refers to sustained, emotionally significant relationships; the present findings suggest that even minimal positive exposure produces measurable effects, though whether such effects scale to therapeutic contexts requires investigation.

4.1 Limitations

While this study offers valuable insights into trust and cooperation dynamics, several limitations warrant consideration. First, the Manipulation condition combined two features of positive relational experiences: forgiveness (agents could not remain in low-trust states following low returns) and behavioral consistency (no pre-programmed low investments during exposure). This combination was deliberate—secure relationships typically involve both dimensions—but it means we tested a "strong" version of the corrective experience hypothesis rather than isolating forgiveness specifically. Control participants, by contrast, experienced realistic continuity: partners who showed the same pattern of occasional low investments across all phases. Although the analysis of pre-low-investment returns favors contrast effects over betrayal aversion as the primary mechanism, future replications could include additional conditions (e.g., forgiving agents that still deliver occasional low investments) to isolate the unique contribution of each relational dimension. Second, our extreme groups design for RS (selecting participants with RSQ scores > 15 or < 10) maximized power to detect moderation effects but may inflate effect sizes and limits generalizability to individuals with moderate RS. Future research should examine RS as a continuous variable. Third, the brief exposure phase (three seven-round games) may have been insufficient to induce lasting changes. Fourth, the online format eliminates social cues present in face-to-face interactions. Notably, while 41% of participants believed they faced humans and a similar proportion were unsure, the observed effects emerged even among those who suspected AI, suggesting robustness of the findings. Despite these limitations, subsequent studies could address these constraints by incorporating face-to-face interactions, longer exposure periods, and continuous RS measurement.

4.2 Constraints on generality

The results may be specific to adults with high or low RS recruited from online platforms, and may not generalize to clinical populations, children, or older adults. Our use of a computerized Repeated Trust Game with HMM agents, while allowing for high experimental control, may limit generalizability to face-to-face interactions or games with different economic structures. The brief exposure phase and pre-programmed low investments are specific to our design and may not reflect real-world trust-building scenarios. The online context may not capture all aspects of high-stakes or information-rich interactions. We believe the core finding of decreased cooperation after exposure to forgiving agents should generalize across different populations and contexts, though the effect's strength and its interaction with RS may vary. While the specific economic game, agent representation, and perception assessment questions could be varied, the use of artificial agents with consistent behavior, an exposure phase with more forgiving behavior, and assessment of both behavior and perceptions should remain constant to preserve the results. Future studies could systematically vary these factors to establish the boundaries of generalizability for our findings. Additionally, cultural differences in norms of cooperation and trust may influence the generalizability of these findings, necessitating cross-cultural replications to establish the universality of the observed effects.

5 Conclusion

This randomised controlled experiment enabled us to uncover unexpected effects of exposure to forgiving behavior on subsequent cooperation, particularly in relation to RS. These findings challenge existing assumptions about fostering cooperative behavior and suggest the need for more nuanced interventions. Importantly, the use of HMM-based artificial agents in this study represents a significant methodological advancement. By providing a balance between experimental control and realistic, adaptive behavior, these agents allowed for a nuanced exploration of trust dynamics that would be challenging to achieve with human confederates or simplistic computer algorithms. This approach opens up new possibilities for studying complex social interactions in controlled settings, potentially bridging the gap between laboratory experiments and real-world social dynamics.

Author contributions statement

I. Guennouni, G. Koppe and C. Korn. designed and developed the study concept. Experiment design, testing and data collection were performed by I. Guennouni. I. Guennouni analysed and interpreted the data under the supervision of G. Koppe and C. Korn. All authors jointly wrote and approved the final version of the manuscript for submission.

Funding

This study was supported by the Federal Ministry of Science, Education, and Culture (MWK) of the state of Baden-Württemberg within the AI Health Innovation Cluster Initiative, the German Research Foundation (DFG) within the Excellence Strategy EXC 2181/1 – 390900948 (STRUCTURES), and the Hector II foundation. I. Guennouni was supported by a research fellowship from the AI Health Innovation Cluster.

Competing interests statement

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

Acknowledgements

We are grateful to Tobias Nolte, Andreas Hula and Read Montague's team at Virginia Tech for sharing with us the data on the RTG allowing us to fit the HMM to investor data. We're also grateful for Samuel Dupret for help designing the online platform used for the experiment and Maarten Speekenbrink for his guidance on the use of HMM models.

Additional information

Correspondence

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

Transparency and data availability

Preregistration: The hypotheses and methods were not preregistered. The primary hypothesis, based on attachment theory, predicted that forgiveness exposure would increase cooperation. The alternative predictions regarding contrast effects and RS-specific updating biases were incorporated into the theoretical framework following initial data analysis, though both perspectives were grounded in existing literature. The analysis plan was not preregistered. Materials: All study materials are publicly available (https://github.com/ismailg/exposure-public). Data: All primary data are publicly available (https://github.com/ismailg/exposure-public). Analysis scripts: All analysis scripts are publicly available (https://github.com/ismailg/exposure-public).

References

Abramov, G., Kautz, J., Miellet, S., & Deane, F. P. (2022). The Influence of Attachment Style, Self-protective Beliefs, and Feelings of Rejection on the Decline and Growth of Trust as a Function of Borderline Personality Disorder Trait Count. *Journal of Psychopathology and Behavioral Assessment*, *44*(3), 773–786. https://doi.org/10.1007/s10862-022-09965-9

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158–2171. https://doi.org/10.3758/s13428-020-01535-9

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Ayduk, O., Downey, G., & Kim, M. (2000). Rejection sensitivity and depressive symptoms in women. *Personality and Social Psychology Bulletin*, *26*(8), 909–919. https://doi.org/10.1177/0146167200269001

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Bandura, A. (1977). *Social Learning Theory*. Prentice Hall.

Berenson, K. R., Gyurak, A., Ayduk, Ö., Downey, G., Garner, M. J., Mogg, K., Bradley, B. P., & Pine, D. S. (2009). Rejection sensitivity and disruption of attention by social threat cues. *Journal of Research in Personality*, *43*(6), 1064–1072. https://doi.org/10.1016/j.jrp.2009.07.007

Bowlby, J. (1978). Attachment theory and its therapeutic implications. *Adolescent Psychiatry*, *6*, 5–33.

Bowlby, J. (1988). *A secure base: Parent-child attachment and healthy human development.* Basic Books.

Charness, G., Cobo-Reyes, R., & Jiménez, N. (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, G., & Feldman, S. I. (1996). Implications of rejection sensitivity for intimate relationships. *Journal of Personality and Social Psychology*, *70*(6), 1327–1343. https://doi.org/10.1037/0022-3514.70.6.1327

Downey, G., Khouri, H., & Feldman, S. I. (1997). Early interpersonal trauma and later adjustment: The mediational role of rejection sensitivity. In *Developmental perspectives on trauma: Theory, research, and intervention* (pp. 85–114). University of Rochester Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fiedler, M., Haruvy, E., & Li, S. X. (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, P., & Campbell, C. (2017). Mentalizing, attachment and epistemic trust: How psychotherapy can promote resilience. *Psychiatria Hungarica: A Magyar Pszichiatriai Tarsasag Tudomanyos Folyoirata*, *32*(3), 283–287.

Fowler, J. H., & Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, *107*(12), 5334–5338. https://doi.org/10.1073/pnas.0913149107

Gao, S., Assink, M., Cipriani, A., & Lin, K. (2017). Associations between rejection sensitivity and mental health outcomes: A meta-analytic review. *Clinical Psychology Review*, *57*, 59–74. https://doi.org/10.1016/j.cpr.2017.08.007

Hepp, J., & Niedtfeld, I. (2022). Prosociality in personality disorders: Status quo and research agenda. *Current Opinion in Psychology*, *44*, 208–214. https://doi.org/10.1016/j.copsyc.2021.09.013

Herpertz, S. C., & Bertsch, K. (2014). The social-cognitive basis of personality disorders. *Current Opinion in Psychiatry*, *27*(1), 73–77. https://doi.org/10.1097/YCO.0000000000000026

Joyce, B., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (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, K. R., & Lipsitt, L. P. (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

Kube, T., Schwarting, R., Rozenkrantz, L., Glombiewski, J. A., & Rief, W. (2020). Distorted cognitive processes in major depression: A predictive processing perspective. *Biological Psychiatry*, *87*(5), 388–398. https://doi.org/10.1016/j.biopsych.2019.07.017

Lieberman, M. D. (2007). Social Cognitive Neuroscience: A Review of Core Processes. *Annual Review of Psychology*, *58*(1), 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Miano, A., Fertuck, E. A., Arntz, A., & Stanley, B. (2013). Rejection Sensitivity Is a Mediator Between Borderline Personality Disorder Features and Facial Trust Appraisal. *Journal of Personality Disorders*, *27*(4), 442–456. https://doi.org/10.1521/pedi_2013_27_096

Mikulincer, M. (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), 1209–1224. https://doi.org/10.1037/0022-3514.74.5.1209

Mulder, R. T., Joyce, P. R., Sullivan, P. F., Bulik, C. M., & Carter, F. A. (1999). The relationship among three models of personality psychopathology: DSM-III-R personality disorder, TCI scores and DSQ defences. *Psychological Medicine*, *29*(4), 943–951. https://doi.org/10.1017/S0033291799008533

Pietrzak, R. H., Downey, G., & Ayduk, O. (2005). Appearance-rejection sensitivity predicts body dysmorphic disorder symptoms and cosmetic surgery. *Annals of Clinical Psychiatry*, *17*(4), 213–219. https://doi.org/10.1080/10401230500295471

Richetin, J., Poggi, A., Ricciardelli, P., Fertuck, E. A., & Preti, E. (2018). The emotional components of rejection sensitivity as a mediator between Borderline Personality Disorder and biased appraisal of trust in faces. *Clinical Neuropsychiatry: Journal of Treatment Evaluation*, *15*(4), 200–205.

Romero-Canyas, R., Downey, G., Berenson, K. R., Ayduk, O., & Kang, N. J. (2010). Rejection sensitivity and the rejection-hostility link in romantic relationships. *Journal of Personality*, *78*(1), 119–148. https://doi.org/10.1111/

j.1467-6494.2009.00611.x

Schuster, F., Hoerz-Sagstetter, S., Seidl, E., Mauer, C., Zeiss, M., Reinhard, M. A., Padberg, F., & Jobst, A. (2021). Ambiguous social rejection from a close other affects neural and behavioral responses in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *12*(6), 583–594. https://doi.org/10.1037/per0000454

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2022). *Afex: Analysis of Factorial Experiments*.

Staebler, K., Renneberg, B., Stopsack, M., Fiedler, P., Weiler, M., & Roepke, S. (2011). Facial emotional expression in reaction to social exclusion in borderline personality disorder. *Psychological Medicine*, *41*(9), 1929–1938. https://doi.org/10.1017/S0033291711000080

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. https://doi.org/10.1037/bul0000217

Weekers, L. C., Hutsebaut, J., & Kamphuis, J. H. (2019). The Level of Personality Functioning Scale-Brief Form 2.0: Update of a brief instrument for assessing level of personality functioning. *Personality and Mental Health*, *13*(1), 3–14. https://doi.org/10.1002/pmh.1434

Zentall, T. 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