Supplementary Information

A Snapshot of the Repeated Trust Game as Seen by Participants

Figure S1 shows a screenshot of the repeated Trust Game at the moment the participant is required to make a decision of how much to send back to the Investor.

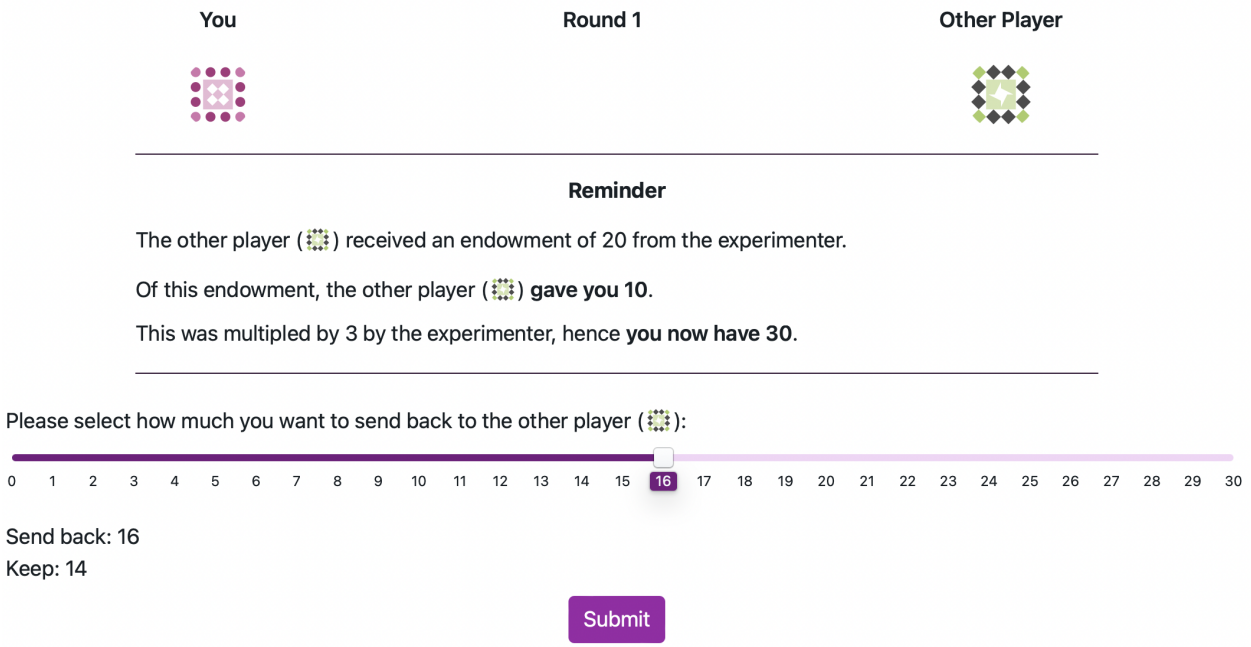

Figure S1: Screenshot of the RTG as seen by participants at the decision phase.

Hidden Markov Model Used to Simulate the Investor's Actions

The HMM assumes that the probability of each investment $I_t = 0, \ldots, 20$, at each trial t, conditional on the current state of the investor S_t, is dependent on an underlying normal distribution with mean μ_s and standard deviation σ_s. The probability of each discrete investment was determined from the cumulative normal distribution Φ, computing the probability of a Normal variate falling between the midway points of the response options. As responses were bounded at 0 and 20, we normalized these probabilities further by taking the endpoints into account. For instance, the probability of an investment $I_t = 2$ is defined as:

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

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

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

where $X_t = R_t - I_t$ is the net return to the investor with R_t the amount returned by the trustee and I_t is the Investment sent.

The advantages of this approach is that it does not require any a priori assumptions about the model features. The number of states, the policy conditional on the state, and the transition function between states can all determined in a purely data-driven way. These HMMs can in turn be used to simulate a human-like agent playing the trust game. This agent may transition to a new state depending on the other player's actions and adopt a policy reflecting its state, thus simulating changes in emotional dispositions of human players during a repeated game. When the investor gains from the interaction, they become more likely to transition to a state where their policy is more "trusting" with generally higher investments. However, faced with losses, the investor is more likely to transition to a more cautious policy with generally lower investments. The policies and the transitions between states are sufficient to build an agent that reflects this type of adaptive behavior and reacts to the trustee's action choices in a way that mimics a human player.

We estimated a three-state model for investor's behaviour, using maximum likelihood estimation via the Expectation-Maximisation algorithm as implemented in the depmixS4 package for R (Visser and Speekenbrink 2021). The model was estimated using investments from existing datasets of human dyads playing 10 rounds of the RTG with the same trustee. The dataset consisted of a total of 381 games from two data sources: First, a total of 93 repeated trust games with healthy investors and a mix of healthy trustees and trustees diagnosed with Borderline Personality Disorder (BPD) (King-Casas et al. 2008). The second source was from data collected as part of a project investigating social exchanges in BPD and antisocial personality disorder reported on elsewhere (Euler et al. 2021; Huang et al. 2020; Rifkin-Zybutz et al. 2021) and consists of 288 games. In both datasets, the investor on which we modelled the HMM's strategy was always selected from a healthy population and the trustees were a mix of healthy participants and those with personality disorders allowing for a diversified interaction behavior.

Mixed-effects Models for Participant Returns

We fit a linear mixed-effects model to participant returns as a proportion of the multiplied investment received. The model specification is described below, and the results are presented in Table **??**.

$$\begin{aligned}\text{ReturnPercentage}_{ij} = &\beta_0 + \beta_1 \text{Opponent}_i + \beta_2 \text{Investment}_i + \beta_3 \text{D-level}_i + \beta_4 \text{Order}_i + \beta_5 \text{Round}_i + \\ &[\text{All interaction terms}] + \\ &b_{0j} + b_{1j} \text{Opponent}_i + b_{2j} \text{Investment}_i + \epsilon_{ij}\end{aligned}$$

where:

- $\text{ReturnPercentage}_{ij}$: The percentage of the tripled investment returned by participant j on trial i.
- The **Fixed Effects** are:
  - β_0: The overall intercept.
  - β_1 to β_5: The main effects for Opponent Type (Opponent_i), scaled Investment (Investment_i), D-Factor level (D-level_i), presentation order (Order_i), and round number (Round_i).
  - The model includes all possible two-way, three-way, four-way, and five-way interaction terms among the five fixed effects, represented by [All interaction terms].
- The **Random Effects** account for by-participant variability:
  - b_{0j}: A random intercept for each participant j.
  - b_{1j}: A random slope for the effect of Opponent Type for each participant j.
  - b_{2j}: A random slope for the effect of Investment for each participant j.
- ϵ_{ij}: The residual error term for participant j on trial i.

Table 1: Full Results of the Linear Mixed-Effects Model of Participant Percentage Returns.

Term	Estimate	Std. Error	df	t-value	p-value	
(Intercept)	0.460	0.015	199.901	30.771	0.000	***
Opponent	-0.001	0.006	669.512	-0.135	0.892	
Investment	0.021	0.009	341.738	2.341	0.020	*
D-Level	-0.010	0.015	199.901	-0.653	0.514	
Order	0.000	0.021	199.889	0.006	0.996	
Round	-0.002	0.000	8140.419	-4.733	0.000	***
Opponent x Investment	-0.002	0.006	8173.610	-0.363	0.716	
Opponent x D-Level	-0.004	0.006	669.512	-0.672	0.502	
Investment x D-Level	0.002	0.009	341.738	0.169	0.866	
Opponent x Order	0.003	0.008	667.589	0.370	0.712	
Investment x Order	-0.002	0.013	344.063	-0.143	0.886	
D-Level x Order	0.013	0.021	199.889	0.648	0.518	
Opponent x Round	0.000	0.000	8154.545	-0.663	0.507	
Investment x Round	0.000	0.000	8189.035	-1.082	0.279	
D-Level x Round	0.000	0.000	8140.419	-1.027	0.304	
Order x Round	0.001	0.000	8135.490	2.234	0.026	*
Opponent x Investment x D-Level	0.012	0.006	8173.610	2.182	0.029	*
Opponent x Investment x Order	0.007	0.008	8134.185	0.870	0.384	
Opponent x D-Level x Order	-0.005	0.008	667.589	-0.663	0.508	
Investment x D-Level x Order	-0.021	0.013	344.063	-1.650	0.100	
Opponent x Investment x Round	0.000	0.000	8207.016	0.402	0.688	
Opponent x D-Level x Round	0.000	0.000	8154.545	0.548	0.584	
Investment x D-Level x Round	-0.001	0.000	8189.035	-2.766	0.006	**
Opponent x Order x Round	0.000	0.000	8144.127	0.553	0.580	
Investment x Order x Round	0.000	0.001	8201.183	0.280	0.779	
D-Level x Order x Round	-0.001	0.000	8135.490	-1.140	0.254	
Opponent x Investment x D-Level x Order	-0.002	0.008	8134.185	-0.273	0.785	
Opponent x Investment x D-Level x Round	-0.001	0.000	8207.016	-1.638	0.102	
Opponent x Investment x Order x Round	-0.001	0.001	8206.907	-1.651	0.099	
Opponent x D-Level x Order x Round	0.000	0.000	8144.127	0.654	0.513	
Investment x D-Level x Order x Round	0.002	0.001	8201.183	3.753	0.000	***
Opponent x Investment x D-Level x Order x Round	0.000	0.001	8206.907	-0.069	0.945	

Significance codes: * p < 0.05, ** p < 0.01, *** p < 0.001

1 Mixed effects models for opponent ratings

We fit linear mixed-effects models to participants' ratings of their opponents (cooperativeness, willingness to play again, and trustworthiness). The model specification for the cooperative ratings is described below as an example; identical model structures were used for the other ratings. The results of the models are presented in Section 3.5 of the main text.

$$\begin{aligned}\text{CooperativeRating}_{ij} = \beta_0 + \beta_1 \text{D-level}_i + \beta_2 \text{Order}_i + \beta_3 \text{Opponent}_i + \\ [\text{Interaction terms}] + \\ b_{0j} + \epsilon_{ij}\end{aligned}$$

where:

- $\text{CooperativeRating}_{ij}$: The cooperativeness rating given by participant j for opponent i.
- The **Fixed Effects** are:

  - β_0: The overall intercept.
  - β_1 to β_3: The main effects for D-Factor level (D-level$_i$), game order (Order$_i$), and Opponent Type (Opponent$_i$).
  - The model includes all two-way and three-way interaction terms among the three fixed effects, represented by [Interaction terms].
- The **Random Effects** account for by-participant variability:
  - b_{0j}: A random intercept for each participant j.
- ϵ_{ij}: The residual error term for participant j's rating of opponent i.

References

Euler, Sebastian, Tobias Nolte, Matthew Constantinou, Julia Griem, P. Read Montague, Peter Fonagy, and Personality and Mood Disorders Research Network. 2021. "Interpersonal Problems in Borderline Personality Disorder: Associations With Mentalizing, Emotion Regulation, and Impulsiveness." *Journal of Personality Disorders* 35 (2): 177–93. https://doi.org/10.1521/pedi_2019_33_427.

Huang, Yu Lien, Peter Fonagy, Janet Feigenbaum, P. Read Montague, Tobias Nolte, and and London Personality and Mood Disorder Research Consortium. 2020. "Multidirectional Pathways Between Attachment, Mentalizing, and Posttraumatic Stress Symptomatology in the Context of Childhood Trauma." *Psychopathology* 53 (1): 48–58. https://doi.org/10.1159/000506406.

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

Rifkin-Zybutz, R. P., P. Moran, T. Nolte, Janet Feigenbaum, Brooks King-Casas, P. Fonagy, and R. P. Montague. 2021. "Impaired Mentalizing in Depression and the Effects of Borderline Personality Disorder on This Relationship." *Borderline Personality Disorder and Emotion Dysregulation* 8 (1): 15. https://doi.org/10.1186/s40479-021-00153-x.

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