Can Exposure To Forgiving AI Foster Cooperative Play?

Supplementary Information

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. Each player was represented with an icon with the participant always on the left of the screen and the opponent on the right. The participants were able to choose the icon that represents them at the start of the experiment. The icon representing the opponent changed at the start of each new game, to simulate a new interaction partner. Participants were not told they were facing computerised opponents. 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 random time lapse (randomly chosen between 5 and 10 seconds).

2 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.

2.1 Mixed-effects model results for participant returns

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

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

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

Table 1: Summary of Mixed-Effects Model of participant returns over all rounds

Term	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.43	0.01	200.97	49.06	0.00
Phase	0.01	0.00	199.27	2.41	0.02
Condition	0.00	0.01	200.97	-0.36	0.72
Investment	0.04	0.00	5,955.45	18.05	0.00
high_RS	0.01	0.01	200.97	1.44	0.15
Phase:Condition	-0.01	0.00	199.27	-2.09	0.04
Phase:Investment	0.00	0.00	5,864.71	-0.78	0.44
Condition:Investment	0.01	0.00	5,955.45	3.73	0.00
Phase:high_RS	0.01	0.00	199.27	1.54	0.12
Condition:high_RS	0.01	0.01	200.97	0.66	0.51
Investment:high_RS	0.00	0.00	5,955.45	0.37	0.71
Phase:Condition:Investment	0.00	0.00	5,864.71	-0.38	0.71
Phase:Condition:high_RS	0.00	0.00	199.27	-0.56	0.57
Phase:Investment:high_RS	-0.01	0.00	5,864.71	-2.80	0.01
Condition:Investment:high_RS	0.00	0.00	5,955.45	-0.83	0.41
Phase:Condition:Investment:high_RS	0.01	0.00	5,864.71	3.04	0.00

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.

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

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

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."