DFP RTG

Ismail Guennouni

2025-03-24

Introduction

Trust and cooperation are fundamental to human social interaction and economic exchange (Berg et al., 1995). The trust game, particularly in its repeated form, has emerged as a powerful tool for investigating the dynamics of trust and reciprocity in controlled settings (Camerer, 2003). While numerous studies have explored various personality traits as predictors of behavior in trust games, recent developments in personality psychology offer new avenues for understanding the underlying factors that influence trustworthiness.

The Dark Factor of Personality (D-factor), proposed by Moshagen et al. (2018), represents a unified construct encompassing various malevolent personality traits. Defined as the general tendency to maximize one's utility at the expense of others, accompanied by beliefs that serve as justifications, the D-factor offers a comprehensive framework for understanding antisocial tendencies. This construct incorporates elements of Machiavellianism, Narcissism, and Psychopathy - traits previously linked to reduced trustworthiness in economic games (Ibáñez et al., 2016; Gunnthorsdottir et al., 2002).

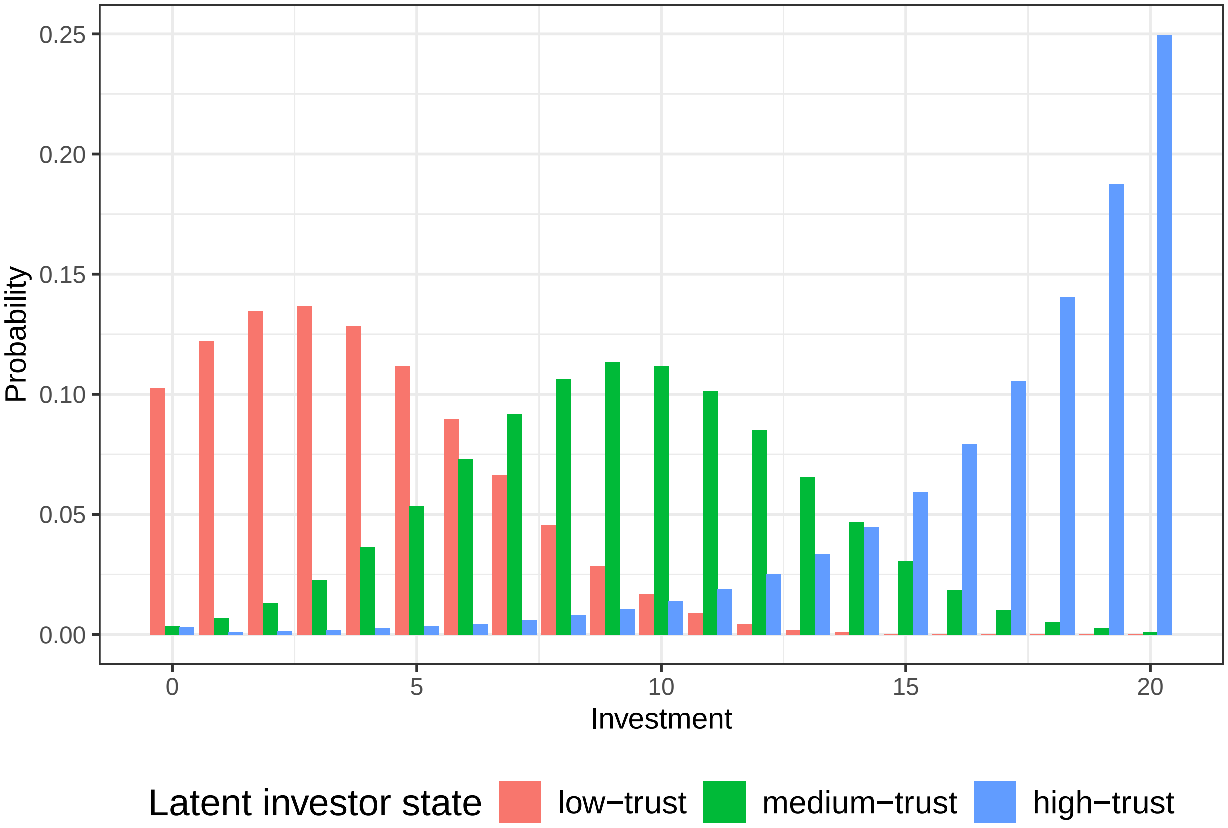
Research has consistently demonstrated negative correlations between dark personality traits and cooperative behavior in various economic games. A meta-analysis by Zhao and Smillie (2015) found that dark triad traits negatively predict cooperation across different economic paradigms. Similarly, Thielmann and Hilbig (2019) observed that dark personality traits predict dishonest behavior in economic interactions. These findings suggest that the D-factor, as a unifying construct, may serve as a potent predictor of untrustworthy behavior in trust games.

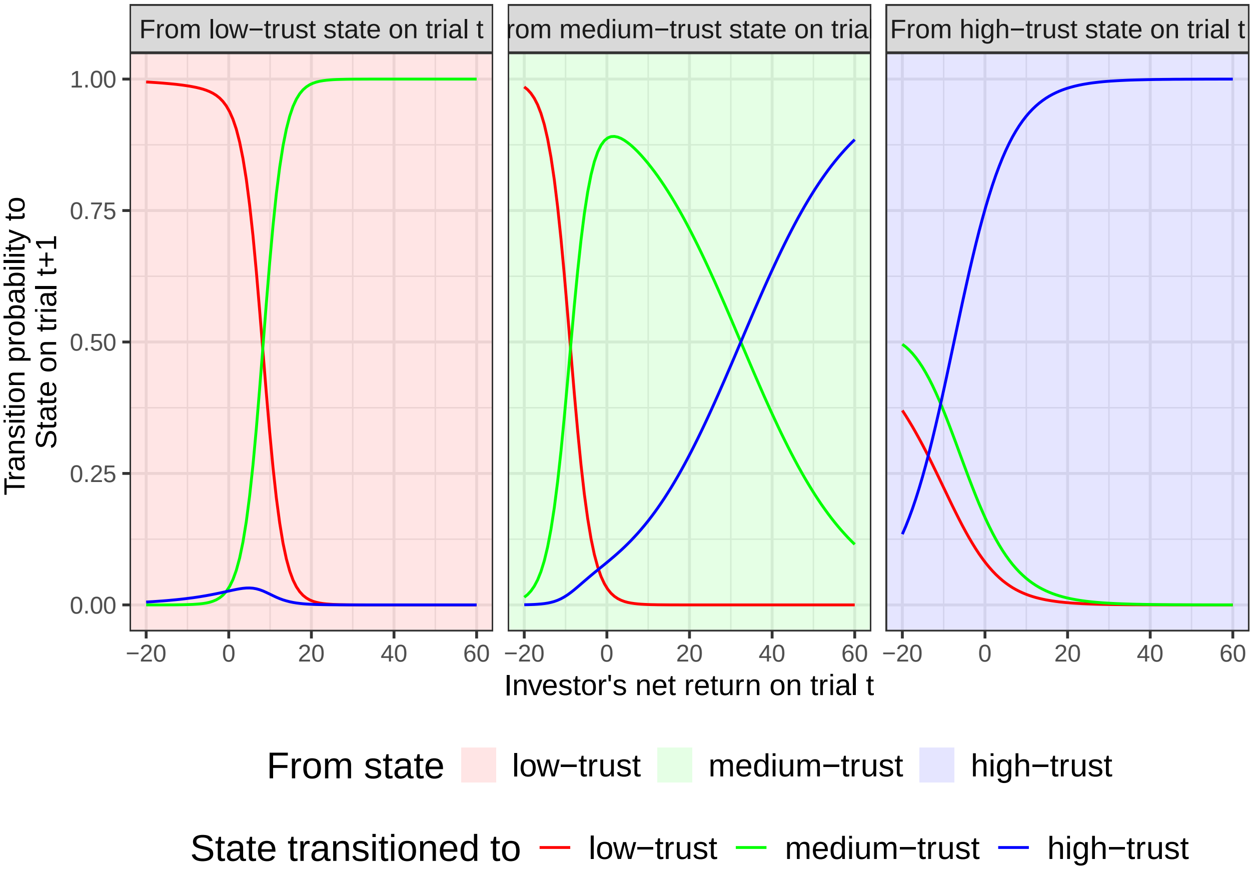
While the D-factor has shown associations with selfish behavior in dictator games (Moshagen et al., 2020) and lower levels of honesty-humility (Zettler et al., 2021), its specific impact on trustworthiness in repeated trust games remains unexplored. This gap is particularly notable given the unique features of the repeated trust game, which allows for the development of reputation and the potential for strategic behavior over multiple interactions (Bohnet & Huck, 2004). The repeated nature of the game introduces complexity not present in one-shot interactions. Individuals with high D-factor scores may exhibit different patterns of behavior over repeated rounds, potentially engaging in strategic trust-building before exploitation. This dynamic aligns with the Machiavellian aspect of the D-factor, which involves a strategic, long-term orientation to personal gain (Jones & Paulhus, 2009).

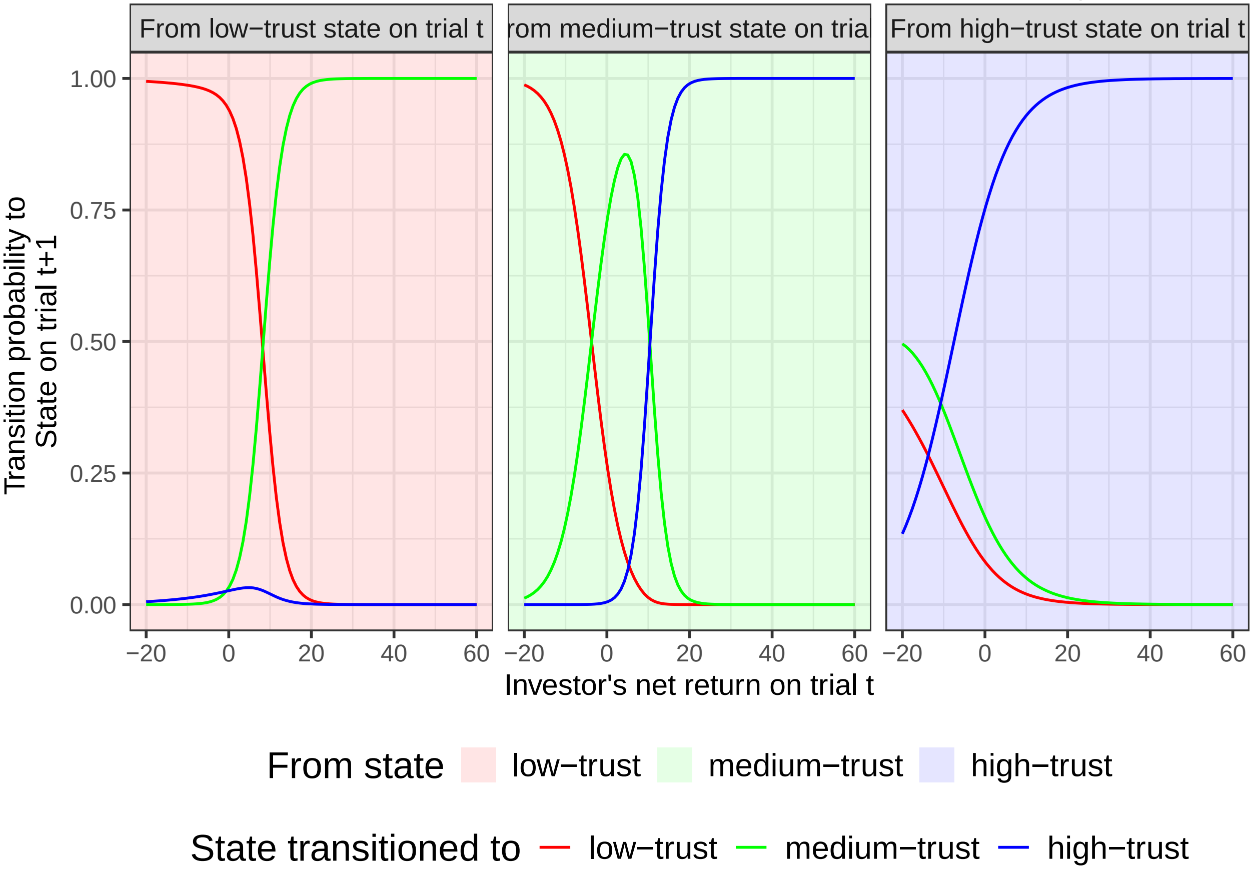
Understanding the relationship between the D-factor and trustworthiness in repeated interactions has significant implications. From an academic perspective, it would bridge the gap between personality psychology and behavioral economics, offering insights into the stability of dark personality influences across repeated social interactions. Practically, such understanding could inform strategies and interventions to promote more cooperative outcomes for people with high scores on the Dark Factor of Personality.

The present study aims to investigate the predictive power of the Dark Factor of Personality on trustworthiness in the repeated trust game. We hypothesize that individuals scoring higher on the D-factor will exhibit less trustworthy behavior as trustees, particularly in later rounds of the game. Additionally, we expect to observe more volatile patterns of reciprocation among high D-factor individuals, potentially reflecting strategic manipulation of trust. By examining these relationships, we seek to contribute to the growing body of literature on the intersection of personality and economic behavior, offering new insights into the psychological underpinnings of trust and reciprocity in repeated social interactions

Methods




From low−trust state on trial t
rom medium−trust state on tria
From high−trust state on trial t
Transition probability to
State on trial t+1
1.00
0.75
0.50
0.25
0.00
−20
0
20
40
60
Investor's net return on trial t
From state
low−trust
medium−trust
high−trust
State transitioned to
low−trust
medium−trust
high−trust


From low−trust state on trial t
rom medium−trust state on tria
From high−trust state on trial t
Transition probability to
State on trial t+1
1.00
0.75
0.50
0.25
0.00
−20
0
20
40
60
Investor's net return on trial t
From state
low−trust
medium−trust
high−trust
State transitioned to
low−trust
medium−trust
high−trust

## Participants

To have participants with large differences in the D factor of personality, a total of 1243 participants were pre-screened on the Prolific Academic platform (prolific.co) using the 16 item Dark Factor of Personality Questionnaire (D16) to finally select two similarly sized groups: One with high D factor scores (90th percentile or higher, D score > 42, N=91) and the other with low D factor scores (10th percentile, score < 22, N=92) totalling 183 participants (44% female). These were then invited through prolific to take part in the main experiment.

To determine the appropriate sample size, we conducted an a priori power analysis using Monte Carlo simulations with the *simr* package in R. The analysis specifically targeted the three-way interaction between d-score, opponent type (stable vs. volatile), and investment level. Parameters for the simulation were based on previous studies, with an expected effect size of -0.1 (correlation between d-score and returns), alpha level of 0.05, and desired power of 0.90. Starting with 50 participants, we iteratively generated synthetic data for a task with 25 rounds per condition and fitted linear mixed-effects models with random intercepts for participants. The simulations incorporated realistic parameter estimates and fixed effects derived from previous research using the same paradigm. This analysis indicated that a sample of 180 participants would provide sufficient power (more than 90) to detect the hypothesized three-way interaction.

The mean age of participants was 33.1 years, with an 9.7 years standard deviation. The majority of participants identified ethnically as White (57%). The online cohort registered 38 unique countries of birth with the most frequent being South Africa (24%), the U.K (20%) followed by Poland (5%) and Greece (4%). Participants were paid a fixed fee of £4 plus a bonus payment dependent on their performance that averaged £0.5. Data was collected over multiple sessions between October and November 2024.

Design and Procedure

The experiment employed a 2 (HMM Type: Human-like or Volatile) × 2 (D-Factor: High or Low) mixed design, with repeated measures on the HMM Type factor. Participants were pre-screened using the 16-item Dark Personality Factor Questionnaire, with individuals classified as either High D or Low D. Participants completed two phases of a Repeated Trust Game (RTG), playing 25 rounds in each phase against different Hidden Markov Model (HMM) investors: a "Human-like" HMM and a more "Responsive" Volatile HMM, with the order counterbalanced across participants. After each RTG phase, participants completed investor evaluations. The experiment concluded with a Turing test to assess perceived humanness of the AI partners, open-ended questions about the interaction, and a final debrief. The experimental interface was designed and implemented online using Empirica v1 [@almaatouq_empirica_2021], with an estimated completion time of 30 minutes per participant. The study 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.

Tasks and Measures

Repeated Trust Game and HMM Investor

Participants played a 25-round RTG [@joyce_trust_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 @ref(fig:HMMPanels).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 @ref(fig:HMMPanels).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.

On all 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).

Investor types

In addition to the human-like HMM resulting from fitting to existing datasets of dyadic play 1 , we created a more volatile HMM. This was achieved by adjusting the parameters of the human-like HMM to alter the state transition probabilities. Specifically, the transition probability for remaining in the "medium-trust" state was set to zero when net returns were singificantly non-nil. The resulting transition function is illustrated in Figure @ref(fig:HMMPanels).D. The state-conditional policies and the transition function in the other latent states remained unchanged.

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 face the Human-like of Volatile HMM first. The timeline of the experiment is shown in Figure @ref(fig:HMMPanels).A. Game one consisted of a 25 round RTG in which participants took the role of trustee, facing the same investor over all 25 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. Game 2 consisted of the exact same set up as in game 1, except for the opponent faced.

At the beginning of each game 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 trusting 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 two games, 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 asked to describe their strategy for both games and finally debriefed and thanked for their participation.

Statistical Analysis

To test whether participants behaved differently in the RTG depending on their D-factor group and opponent faced, we model the percentage return (percentage of tripled investment returned to investor) using a linear mixed effects model to participants returns, with Opponent (Human-like vs. Volatile HMM), The order of opponents (Volatile first = True or false), Investment, round number and D-factor (High vs Low D-factor score) as well as their interactions as fixed effects, and player-wise random intercepts and slopes for the Investment variable. The full specification of the statistical model can be found in the supplement.

The model was estimated using the `afex` package [@singmann_afex_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_balancing_2017]. A similar process was used to establish the random effects structures of linear mixed-effects models used to analyse the HMM agent investments as well as the participants' ratings of the co-players. There is no agreed upon way to calculate effect sizes for mixed effects models. Instead, we will report on testing differences in marginal means. 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 (subtract the overall investment mean and divide by overall standard deviation) as centering is beneficial to interpreting the main effects more easily in the presence of interactions. To probe significant interactions, we conducted planned contrasts using the `emmeans` package in R. Given that we were testing multiple pre-planned comparisons, we applied the "Sidak" correction to control for familywise error rate while maintaining reasonable statistical power. This approach allowed us to investigate specific hypotheses about the differential effects of our manipulation across phases and RS groups, while protecting against inflated Type I error rates.

Results

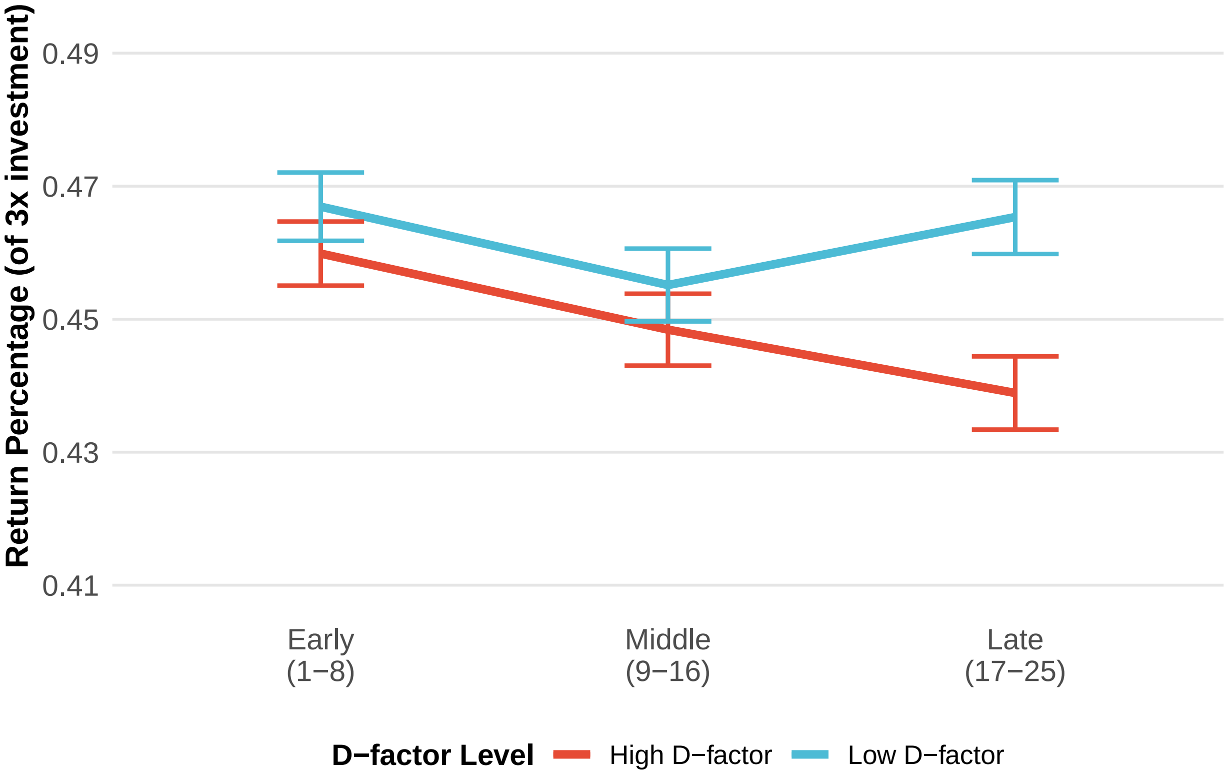
Mean investment and return per round

On average, investments and returns, as shown in Figure @ref(fig:XXXX), 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_investment_2008; @fiedler_social_2011].

Comparing high versus low D-factor participants across all rounds, we observed several behavioral differences. High D-factor participants received lower investments ($t(9147.28) = -4.74$, $p < .001$) and consistently returned less money to investors ($t(9116.25) = -6.90$, $p < .001$). The difference in return percentage was statistically significant ($t(9147.91) = -4.03$, $p < .001$), with high D-factor participants returning approximately 2-4 percentage points less of the tripled investment.

We then examined differences in trust game behavior between participants with high and low Dark Factor of Personality (D-factor) scores across three game periods: early (rounds 1-8), mid (rounds 9-16), and late (rounds 17+ excluding the alst round). We compared HMM investments, absolute returns, and percentage returns between the two groups using Welch's t-tests. We excluded the last round and analysed that data separately as participants were told it was the last interaction in that round.

Whilst there were no significant difference in investment received and percentage returns sent by the participants between high_D and low_D groups during early and mid periods, significant differences emerged for all three measures during the late period. The HMM invested significantly *less* in high_D participants than low_D participants (t(2925.95) = -5.88, p = 4.631e-09). Furthermore, high_D participants sent back significantly lower absolute returns (t(2904.36) = -7.01, p = 3.050e-12) and lower percentage returns (t(2925.87) = -3.74, p = 1.894e-04) compared to low_D participants.



Last Round Analysis

We conducted a separate analysis focusing solely on the final round of each game, where participants knew there would be no further interactions. This allows us to examine behavior in a context resembling a dictator game. We compared absolute returns and percentage returns between high_D and low_D groups. We used both Welch's t-tests and Wilcoxon rank-sum tests (also known as Mann-Whitney U tests). The Wilcoxon test is a non-parametric test that does not assume normality, making it a more robust choice if the data are not normally distributed, which is often the case with economic game data, especially in smaller samples or with outliers.

In the last round, high_D participants sent back significantly lower absolute returns than low_D participants (t(358.42) = -2.31, p = 0.021); Wilcoxon W = 14504, p = 0.025). Similarly, high_D participants sent back a significantly lower percentage of the tripled investment (t(363.98) = -2.18, p = 0.030); Wilcoxon W = 14445, p = 0.021). Both parametric (t-test) and non parametric tests (Wilcoxon) show significant differences.

Total Payoff Analysis

Finally, we analyzed the total payoffs earned by participants across both games, comparing high_D and low_D individuals. This analysis aimed to determine whether differences in strategy observed during the game (particularly in the later periods) translated into overall differences in earnings. We used a Welch's t-test and a Wilcoxon rank-sum test.

The results showed no significant difference in total payoffs between high_D and low_D participants (t(180.99) = -0.17, p = 0.862; Wilcoxon W = 4253, p = 0.853).

Although high-D participants sent back lower returns in the late period of the trust game, their total accumulated payoff across all rounds was not significantly different from that of low-D participants. This seemingly paradoxical result can be explained by the adaptive behavior of the HMM opponent. While high-D individuals adopted a less cooperative strategy in later rounds, keeping a larger portion of the returns for themselves, the HMM responded by reducing its investments in these individuals. Therefore, the higher proportion kept by high-D participants was offset by a reduction in the amount they received, leading to similar overall earnings compared to the more cooperative low-D participants.

Round by round analysis

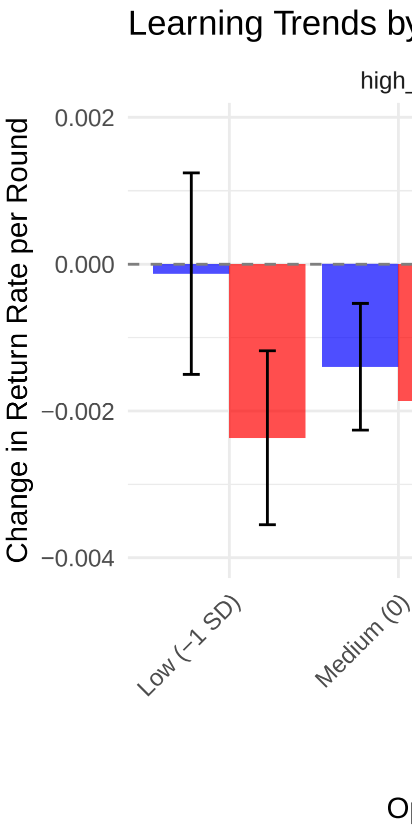
Key Findings:

Main effects Our analysis revealed a significant main effect of investment amount ($F(1, 344.27) = 10.38$, $p = .001$), with participants returning higher percentages when they received larger investments, demonstrating positive reciprocity. We also found a significant main effect of round number ($F(1, 8147.38) = 21.47$, $p < .001$), showing that return percentages generally decreased over time as the game progressed.

D-Factor by Round Number interaction We found a significant interaction between D-factor and round number ($F(1, 8147.38) = 6.91$, $p = .009$). Participants with high D-factor scores demonstrated a significant negative slope in their return proportions as the game progressed, indicating a systematic decrease in reciprocity over time (slope = -0.0016, 95% CI [-0.0023, -0.0010]). In contrast, participants with low D-factor scores maintained relatively stable return rates across rounds, with a slope not significantly different from zero. The difference between these slopes was statistically significant (z = -2.64, p = 0.008).

Opponent Type, Investment, and D-factor Interaction Low D-factor participants showed significant positive reciprocity with both opponents: a one-unit increase in investment led to a significant increase in return percentage for both the human-like HMM (2.1%*,* p* = 0.011) and the volatile HMM (3.1%***,* p* = 0.000).

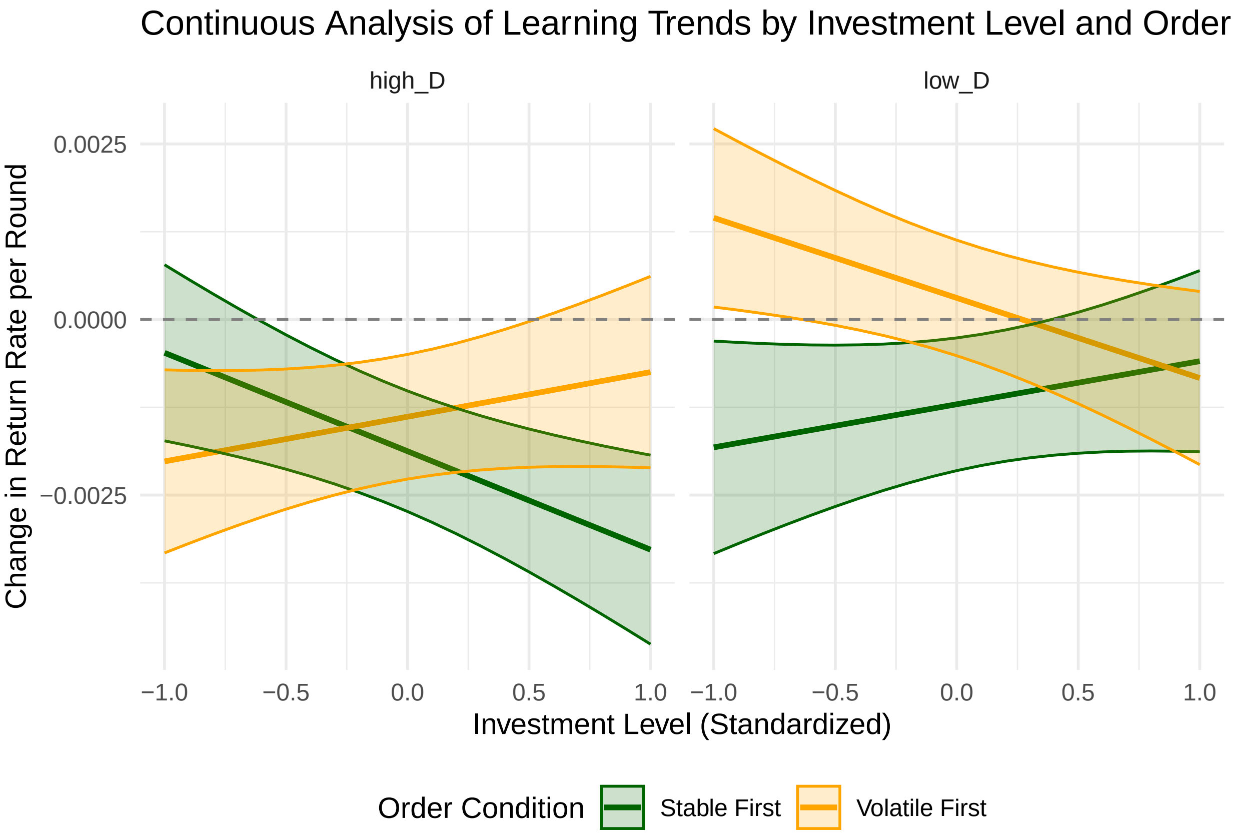
In contrast, high D-factor participants did *not* show significant reciprocity with either the human-like HMM (slope = 0.008, $p = 0.314$) or the volatile HMM (slope = 0.005, $p = 0.509$), indicating their returns were less influenced by investment amount.



Four-Way Interaction: Opponent, Investment, D-factor, and Round Number

Analysis of the significant four-way interaction ($F(1, 8215.95) = 5.92$, $p = .015$) revealed that only high D-factor participants facing the human-like opponent showed investment-dependent changes in behavior across rounds (p = 0.042). For these participants, returns decreased significantly across rounds with high investments (slope = -0.00267, 95% CI [-0.00398, -0.00135]), but remained stable with low investments (slope = -0.00013, 95% CI [-0.00150, 0.00124]), a significant difference in slopes (p = 0.042).

Neither low D-factor participants nor high D-factor participants facing the volatile opponent showed this strategic pattern. This suggests high D-factor participants specifically exploit predictable opponents by systematically reducing reciprocity over time on high-investment trials.



Four-Way Interaction: Investment, Order, D-factor, and Round Number The significant four-way interaction involving investment amount, order of opponent presentation (volatile first or stable/human-like first), D-factor, and round number ($F(1, 8209.77) = 14.07$, $p < .001$) reveals a complex interplay of factors influencing return behavior. The key finding is that the *order* in which participants faced the opponents, combined with their D-factor level, influenced how their returns changed over time *depending on the investment level*.

To disentangle this interaction, we used `emtrends` to examine the simple slopes of return percentage over rounds, for each combination of D-factor, order condition, and investment level. We then used `test()` to determine if these slopes were significantly different from zero.

High-D participants who faced the *stable* (human-like) *opponent first* showed a strategic pattern: they significantly decreased their returns over rounds for *high* (slope = -0.00328, p < .001) and *medium* investments (slope = -0.00188, p < .001), but not for low investments (slope = -0.00047, p = 0.841). This reinforces the idea that high-D individuals are more likely to reduce cooperation when they perceive an opportunity for greater gain (higher investments) and a predictable partner. In contrast, high-D participants who faced the *volatile opponent first* showed the *opposite* pattern: decreasing returns for *low* (slope = -0.00202, p = 0.007) and *medium* investments (slope = -0.00138, p = 0.007), but not for high investments (slope = -0.00075, p = 0.629).

Low-D participants, regardless of the order in which they faced the opponents, did *not* show significant changes in their returns across rounds for any investment level.

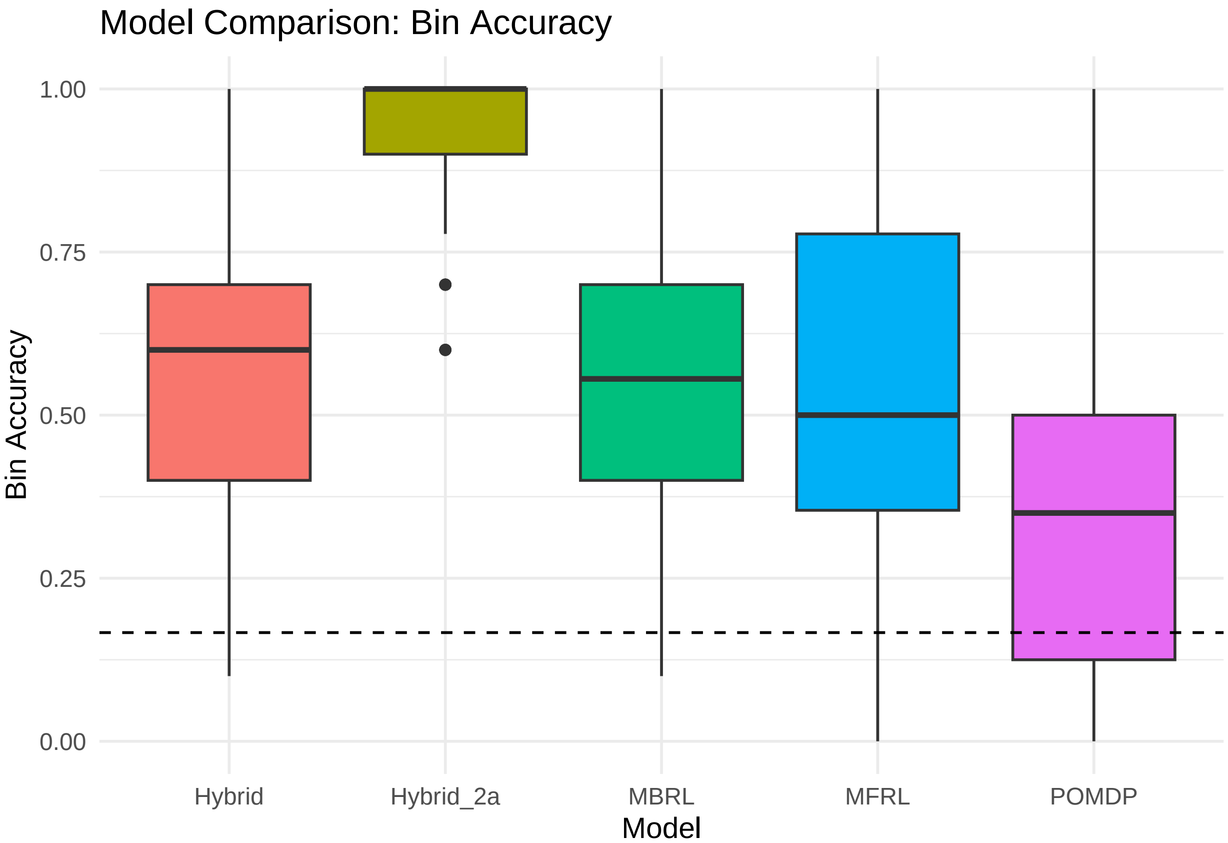
The analysis revealed different patterns of responses to investor states between high and low D-factor participants. Low D-factor participants showed significant increases in return rates from unhappy to neutral states (b = 0.033, SE = 0.0075, z = 4.39, p < .001) and from unhappy to happy states (b = 0.057, SE = 0.0081, z = 7.04, p < .001), as well as from neutral to happy states (b = 0.024, SE = 0.0063, z = 3.87, p < .001). In contrast, high D-factor participants showed no significant differences in returns between unhappy and neutral states (b = -0.00005, SE = 0.0072, z = -0.007, p = .994) or unhappy and happy states (b =

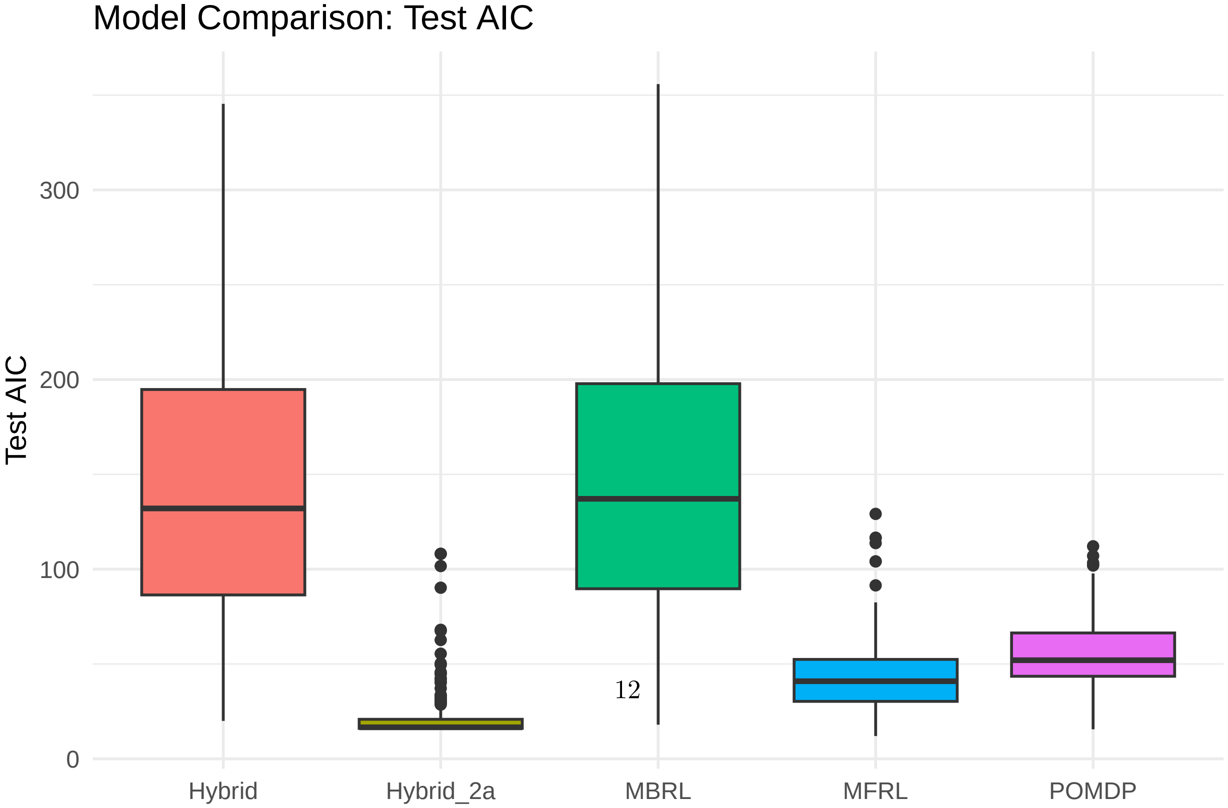
0.013, SE = 0.0082, z = 1.53, p = .127), though they showed a marginally significant increase from neutral to happy states (b = 0.013, SE = 0.0065, z = 1.92, p = .054). Direct comparisons between D-factor groups revealed that low D-factor participants showed significantly stronger positive responses than high D-factor participants between unhappy and neutral states (b = 0.033, SE = 0.0104, z = 3.18, p = .002) and between unhappy and happy states (b = 0.045, SE = 0.0116, z = 3.87, p < .001). However, the groups did not differ significantly in their response to the transition from neutral to happy states (b = 0.012, SE = 0.0091, z = 1.29, p = .196). These results suggest that low D-factor participants were more responsive to improvements in investor state, particularly when recovering from an unhappy state, while high D-factor participants showed more stable returns across investor states.

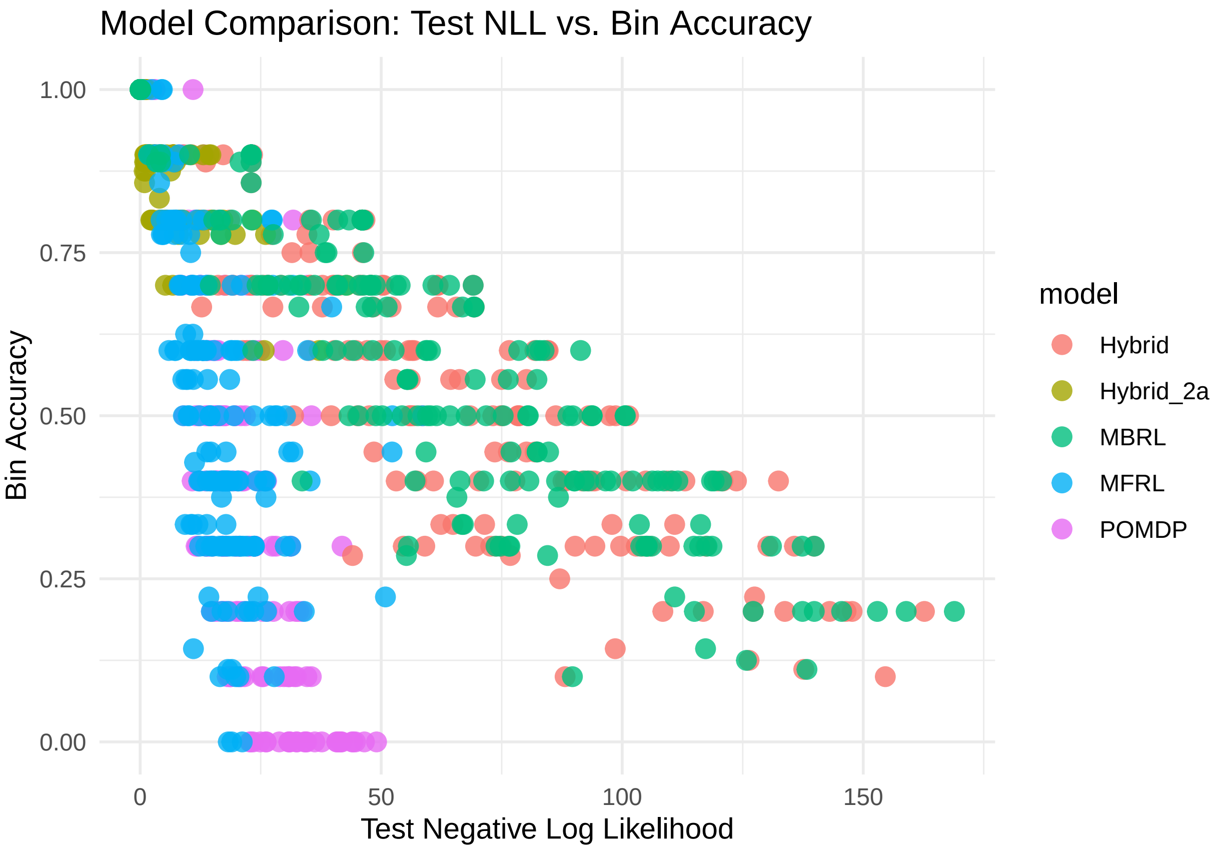
Computational Modelling

Model comparison (Simple RL, MBRL, hybrid with planning, POMDP)

Model comparison, OOF testing







checking model assumptions

Discussion

Contrary to our expectations, high D-factor participants did not demonstrate strategic exploitation across different investor states. The literature suggests that individuals high in dark personality traits, particularly the Machiavellian aspect of the D-factor, should show strategic adaptation to maximize personal gain, potentially through initial trust-building followed by exploitation. However, our results reveal an opposing pattern: high D-factor participants showed relatively stable returns across investor states, particularly with the volatile investor, suggesting a form of behavioral rigidity rather than strategic flexibility.

This behavioral inflexibility was especially evident in the volatile HMM condition, where high D-factor participants maintained consistent return rates regardless of the investor's state. In contrast, low D-factor participants showed marked sensitivity to investor states, adjusting their returns upward as the investor's state improved from unhappy to happy. This pattern held across both human-like and volatile HMM conditions, though it was more pronounced with the volatile investor. These findings suggest that contrary to the Machiavellian tendency for strategic manipulation, high D-factor individuals might actually be less adept at reading and responding to social cues in economic interactions.

One possible explanation for this unexpected pattern lies in the fundamental nature of the D-factor as "the tendency to maximize one's individual utility at the expense of others with self-justifying beliefs." The behavioral rigidity we observed might represent a form of defensive strategy - by maintaining stable (and relatively lower) returns regardless of the investor's state, high D-factor participants could be prioritizing consistent personal gain over reciprocity. This interpretation aligns with recent work suggesting that dark personality traits might manifest not just as active exploitation, but as a general insensitivity to social cues that would typically motivate cooperative behavior.

The finding that low D-factor participants showed greater behavioral flexibility and responsiveness to investor states suggests that the ability to maintain cooperative relationships might require active engagement with partner behavior rather than strategic manipulation. This has important implications for understanding how personality traits influence economic decision-making and challenges the traditional view of dark personality traits as primarily manifesting through strategic exploitation.

Conclusion