DFP RTG

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

2025-03-27

[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*

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

Abstract:

Introduction

Trust and cooperation are fundamental to human social interaction and economic exchange (Berg, Dickhaut, and McCabe 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, Hilbig, and Zettler (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, McCabe, and Smith 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, Hilbig, and Zettler 2020) and lower levels of honesty-humility (Zettler, Moshagen, and Hilbig 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 and 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 and Paulhus 2009).

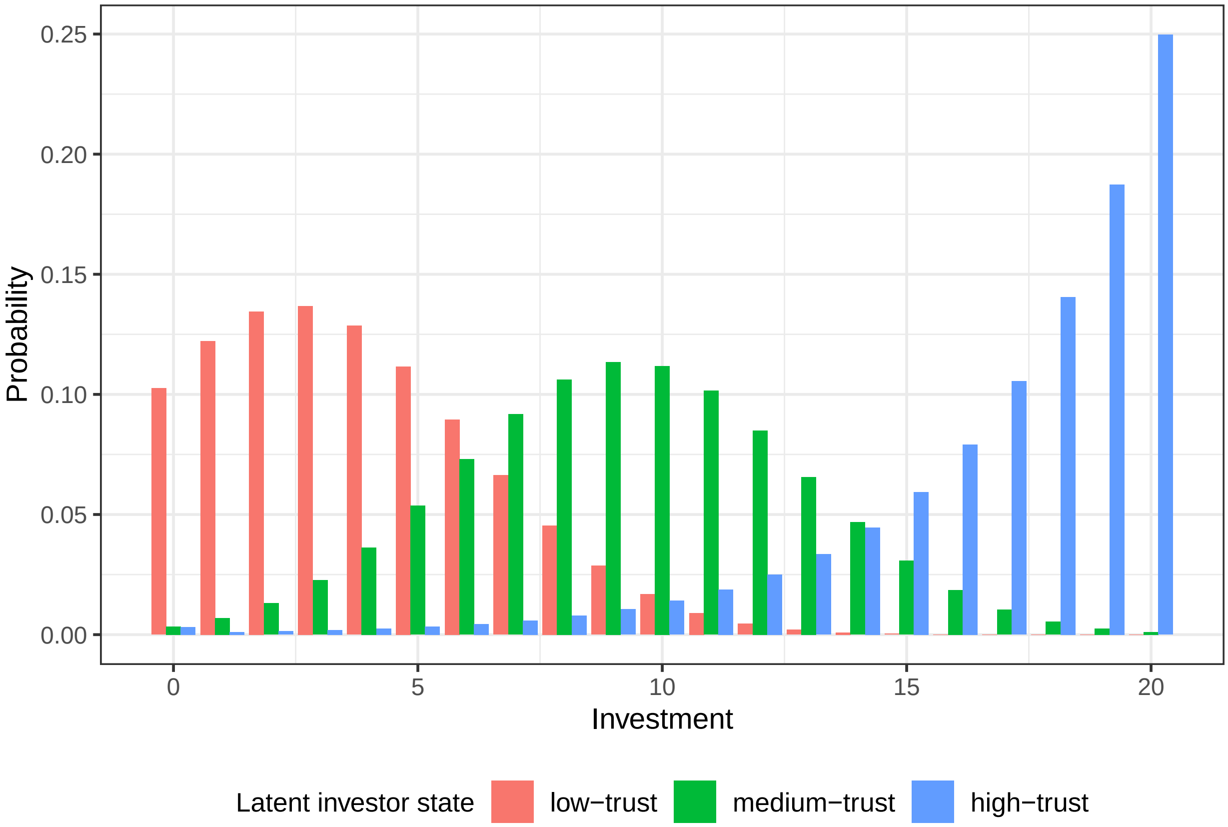
A notable exception is the study by Gong et al. (2019), which examined the relationship between psychopathic traits and behavior in a modified trust game. However, their design involved multiple different partners rather than repeated interactions with the same opponent. Similarly, Rosenberger, Tsivilis, and Müller (2019) examined fairness norm violations in violent offenders during a repeated trust game, finding that antisocial traits (Factor 2 of psychopathy) rather than interpersonal/affective traits (Factor 1) were associated with reduced reciprocity. However, this study focused on clinical psychopathy in offenders rather than the D-factor in the general population.

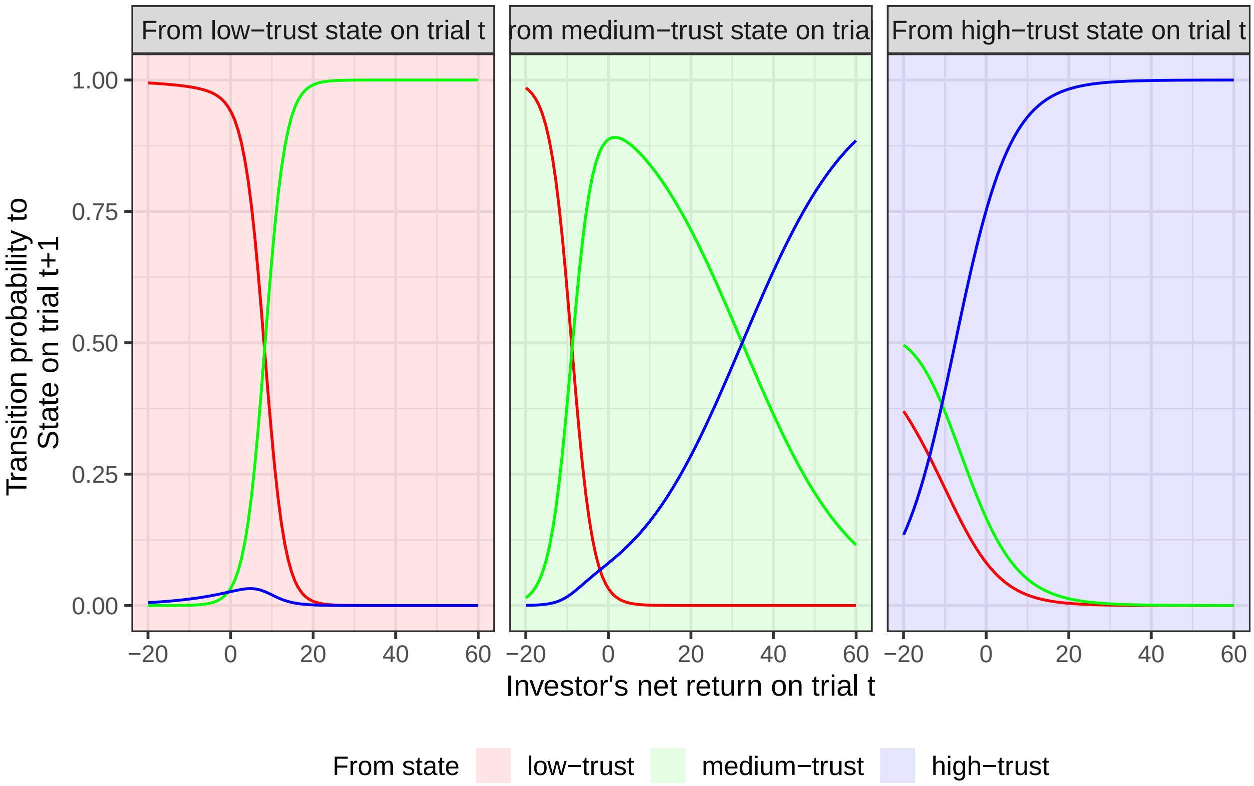
The present study addresses several crucial gaps in our understanding of how dark personality traits influence economic behavior. First, by using a repeated trust game paradigm with 25 rounds, we can examine whether high D-factor individuals demonstrate strategic patterns that evolve over time, such as initial trust-building followed by exploitation. Second, by manipulating opponent predictability through the use of different Hidden Markov Model (HMM) agents—one "human-like" and one more "volatile"—we can assess whether the behavioral manifestation of dark personality traits varies based on environmental stability. This approach allows us to test whether high D-factor individuals demonstrate sophisticated social intelligence by adapting their strategies to different counterparts. The use of HMM agents as opponents provides significant methodological advantages for studying dark personality traits. These computational models, trained on real human behavior data, allow us to create standardized yet responsive interaction partners whose strategies can be precisely characterized and manipulated. This approach combines ecological validity (as the agents mimic actual human behavior) with experimental control, allowing us to isolate the effects of personality while maintaining the dynamic nature of trust interactions.

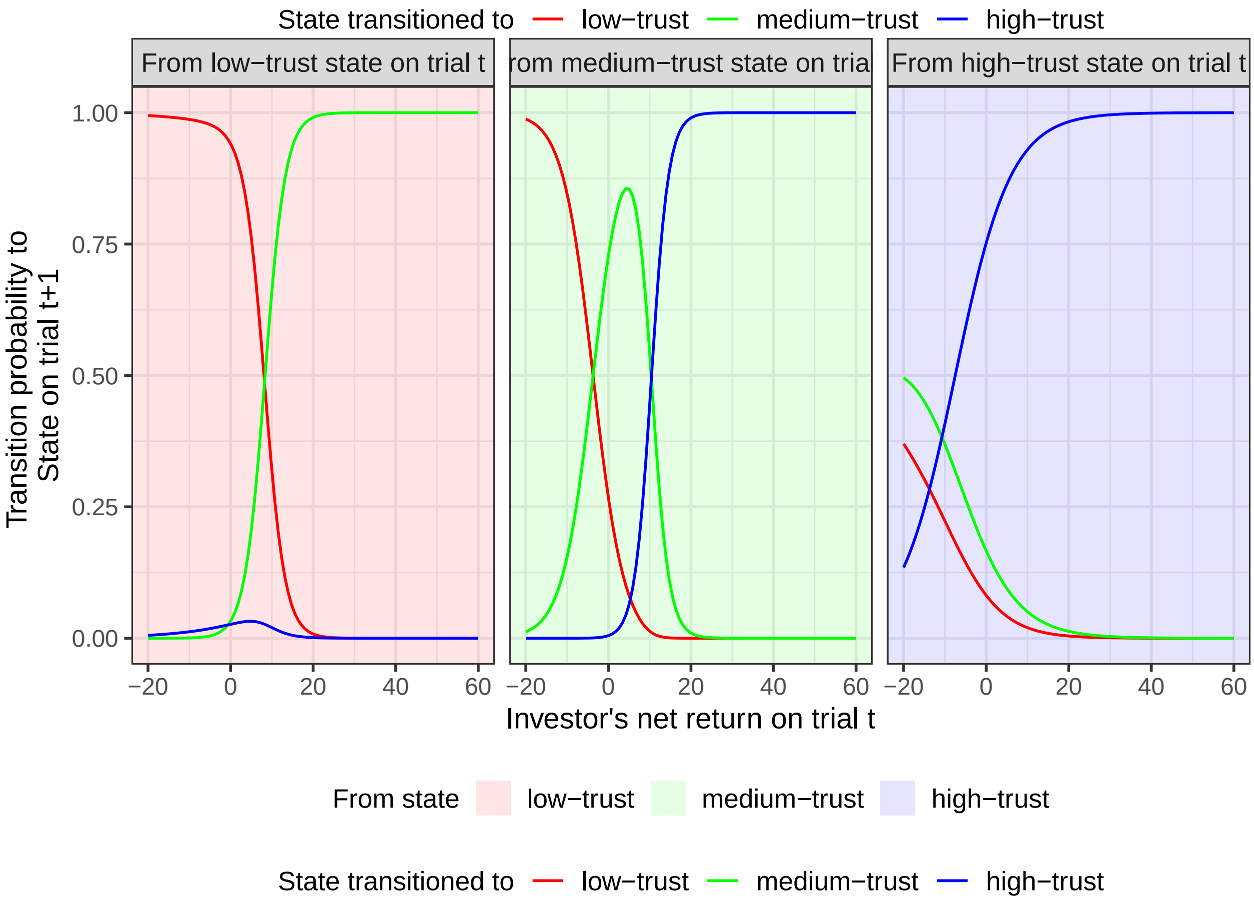
Understanding the relationship between the D-factor and trustworthiness in repeated interactions has significant implications. From an academic perspective, it would bridge the gap between personality psychology and behavioral economics, offering insights into the stability of dark personality influences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

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 high D-factor individuals to show greater strategic adaptation to opponent type, with more pronounced exploitation of predictable opponents compared to volatile ones. Finally, we explore whether these behavioral patterns are accompanied by systematic differences in perception of opponents, with high D-factor individuals potentially showing more negative evaluations regardless of opponent behavior.

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.

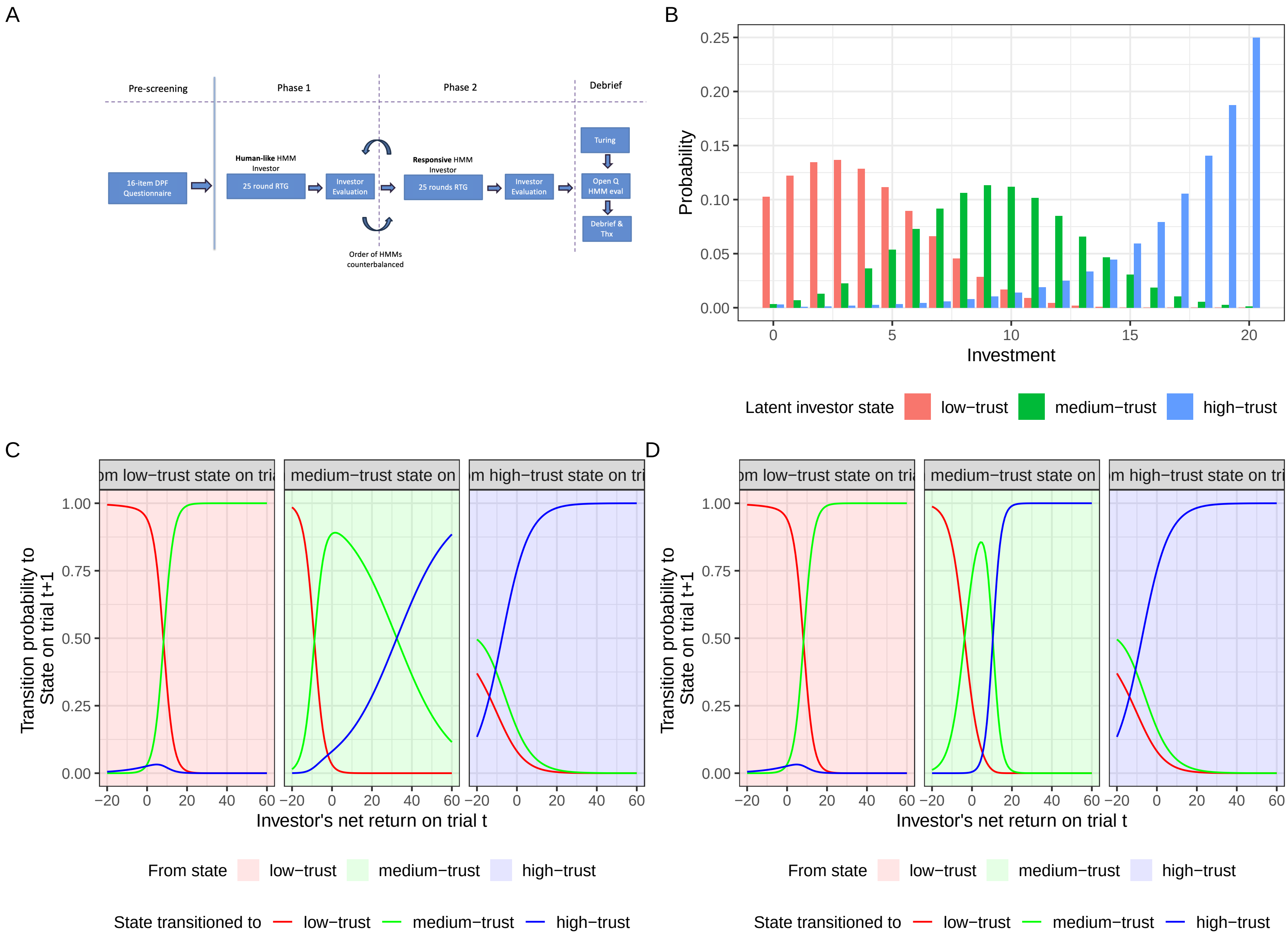


Figure 1: Panel A: A timeline of the experiment. The RTG is played in dyads, with participants always assigned the role of the trustee and the HMM agent that of the investor. The investor is endowed with 20 units at the start of each round. They need to decide how much of that endowment they want to invest with the trustee. The investment is then multiplied by a factor of 3 and sent to the trustee who needs to decide how much of the multiplied investment they want to send back to the investor. The difference between phases is the type of agents participants are facing. Panels B - C: We construct the artificial investor agent by fitting a three-state hidden Markov model to data of human investors engaged in the 10 round RTG. From the fitted HMM, we get the distribution of investments by the human-like agent, conditional on its latent state as shown in Panel B. The fitted HMM also yields the transition probability of the agent to a state on trial

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 et al. 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 (Berg, Dickhaut, and McCabe 1995) in the trustee role against a computer-programmed investor. On each round the investor is endowed with 20 units and decides how much of that endowment to invest. This investment is tripled and the trustee then decides how to split this tripled amount between them and the investor. If the trustee returns more than one third of the amount, the investor makes a gain. Each player was represented with an icon with the participant always on the left of the screen and the co-player on the right. 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 group) 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 et al. 2022) in R. More complex models with additional random effects could not be estimated reliably, and as such the estimated model can be considered to include the optimal random effects structure (Matuschek et al. 2017). A similar process was used to establish the random effects structures of 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 DFP groups, while protecting against inflated Type I error rates.

Results

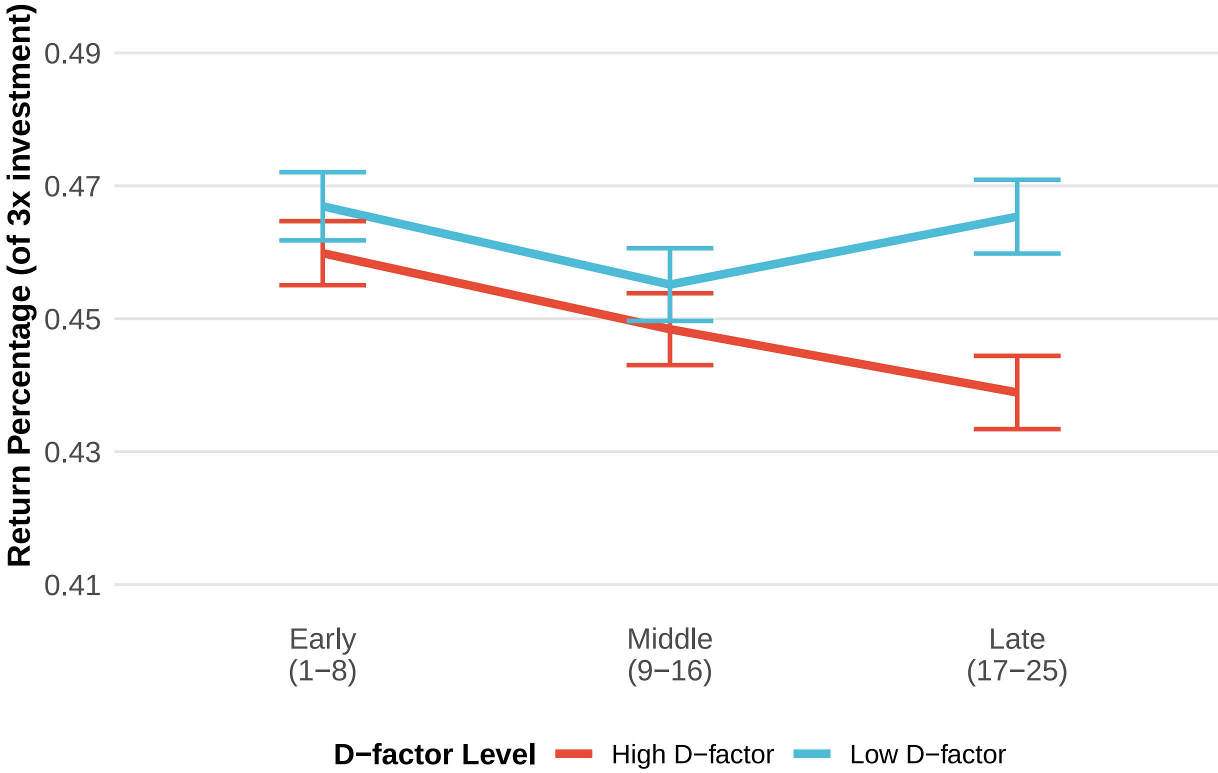
Mean investment and return per round

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, Cobo-Reyes, and Jiménez 2008; Fiedler, Haruvy, and Li 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

To analyse participants behavior on a round by round basis, we look at the fit results from the linear mixed effects model of participant percentage returns detailed in the methods section.

Main effects We found 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.

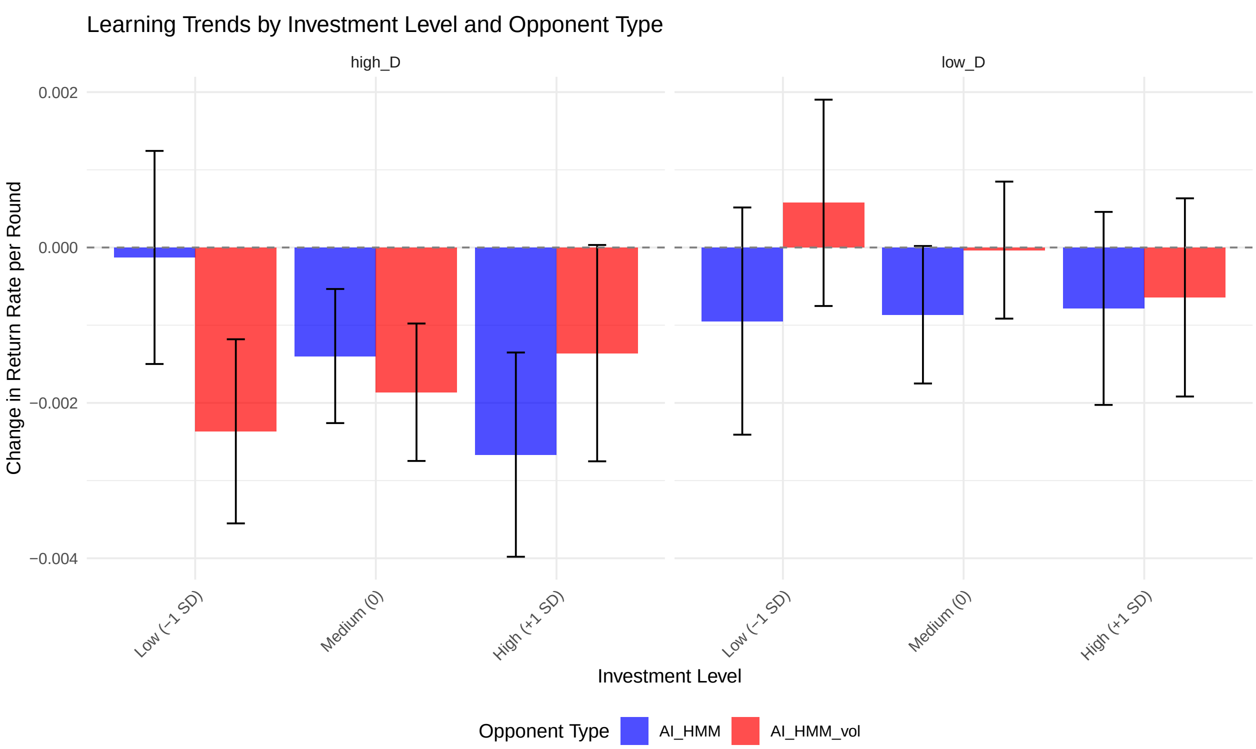
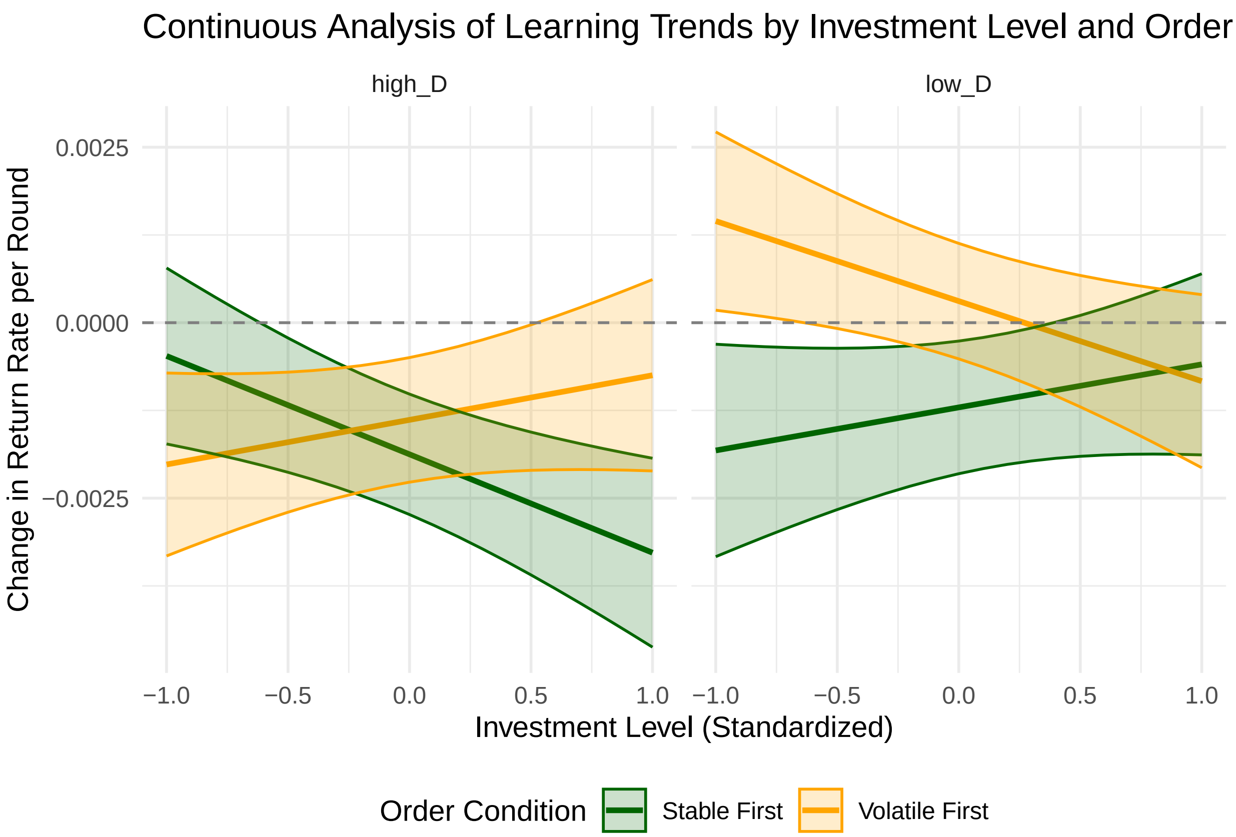


Figure 2: Averages and standard errors of thange in returns per round of...

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.*

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.

Analysis of opponent ratings

Cooperative Ratings

Linear mixed-effects analysis revealed a significant main effect of D-level ($F(1, 179) = 7.35$, $p = .007$), with low D-factor participants rating their opponents as more cooperative than high D-factor participants ($t(179) = -2.71$, $p = .007$).

A significant main effect of game order ($F(1,179) = 10.67$, $p = .001$) indicated participants rated opponents in their first game as more cooperative than those in their second game ($t(179) = 3.27$, $p = .001$). Additionally, there was a significant main effect of opponent type ($F(1,179) = 5.31$, $p = .022$), with Human-like HMM opponents receiving higher cooperative ratings than Volatile opponents ($t(179) = 2.31$, $p = .022$).

Play Again Ratings

Analysis of participants' willingness to play with the same opponent again revealed a significant main effect of game order ($F(1,179) = 13.83$, $p < .001$), with participants generally more willing to play again with opponents from their first game ($t(179) = 3.72$, $p < .001$). This main effect was qualified by a significant D-level × Game Order interaction ($F(1,179) = 4.05$, $p = .046$). Post-hoc analyses revealed that in the first game, high D-factor participants were significantly less willing to play again with their opponents compared to low D-factor participants ($t(320.47) = -2.37$, $p = .018$), while no such difference existed in the second game ($t(320.47) = -0.07$, $p = .947$). Examining changes across games, low D-factor participants showed a significant decrease in willingness to play again from the first to the second game ($t(179) = 4.05$, $p < .001$), while high D-factor participants maintained consistent ratings across games ($t(179) = 1.21$, $p = .229$).

Trusting Ratings

For trust ratings, significant main effects were observed for D-level ($F(1,179) = 7.53$, $p = .007$), game order ($F(1,179) = 7.21$, $p = .008$), and opponent type ($F(1,179) = 4.62$, $p = .033$). These effects were qualified by a significant three-way interaction between D-level, game order, and opponent type ($F(1,179) = 4.76$, $p = .030$).

Post-hoc analyses revealed a complex pattern of trust perceptions. High D-factor participants rated Human-like opponents as significantly less trusting than low D-factor participants in the first game ($t(314.81) = -2.96$, $p = .003$). In contrast, high D-factor participants rated Volatile opponents as significantly less trusting than low D-factor participants in the second game ($t(314.81) = -2.70$, $p = .007$). Low D-factor participants showed a significant *increase* in trust ratings for Human-like opponents from the first to the second game ($t(314.81) = 2.22$, $p = .027$), while high D-factor participants showed a significant *decrease* in trust ratings for Volatile opponents from the first to the second game ($t(314.81) = 2.47$, $p = .014$). Additionally, high D-factor participants in the second game differentiated between opponent types, rating Human-like opponents as significantly more trusting than Volatile opponents ($t(314.81) = 2.51$, $p = .013$).

In summary, participants with higher Dark Factor scores demonstrated consistently more negative perceptions of their opponents, particularly regarding cooperation and trust. The pattern of results indicates that individual differences in Dark Factor traits influence not only the overall level of opponent ratings but also how these ratings change across repeated interactions and between different opponent types. Notably, low D-factor participants showed greater sensitivity to game order, with more pronounced decreases in ratings from first to second game, while high D-factor participants demonstrated greater discrimination between opponent types in their trust ratings.

```
## NOTE: Results may be misleading due to involvement in interactions
## NOTE: Results may be misleading due to involvement in interactions
```

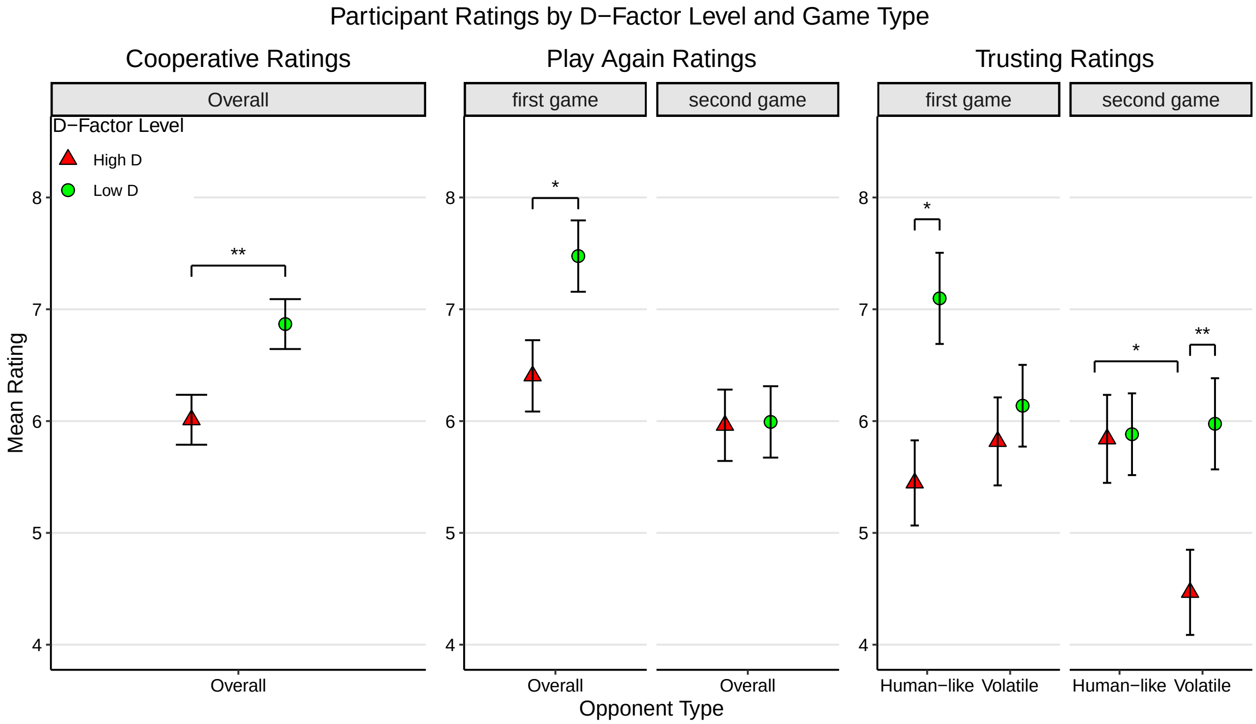


Figure 3: Averages and standard errors of the participants ratings of the opponent (y-axis) by each game and d-factor group for each opponent (x-axis). The left panel represents participants' perception of co-player cooperativeness, the right one one indicates perceived co-player trut rating, and the middel panel shows the participants' willingness to play again with the same co-player. Cooperation, trust perception, and willingness to play again ratings were generally lower for the high DFP group

Debrief questions

Around 57% of participants either thought that they played against a human opponent or were not sure whether the investor was a human or a machine.

Computational Modelling

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

Model comparison, OOF testing

checking model assumptions

Discussion

–>

The present study investigated how the Dark Factor of Personality (D-factor) influences behavior in repeated trust games, focusing on trustworthiness patterns, strategic adaptation, and perception of counterparts. Our findings reveal nuanced relationships between dark personality traits and economic decision-making that both confirm and challenge existing theoretical frameworks.

Our results demonstrated significant behavioral differences between individuals with high and low D-factor scores.

High D-factor participants consistently returned lower proportions of investment compared to their low D-factor counterparts, with this difference becoming particularly pronounced in later rounds of the game. This pattern aligns with the fundamental definition of the D-factor as “the general tendency to maximize one’s utility at the expense of others, accompanied by beliefs that serve as justifications” (Moshagen, Hilbig, and Zettler 2018). The lower return percentages directly translate to greater self-benefit at the investor’s expense, supporting the construct validity of the D-factor in predicting economic behavior. Interestingly, the timing of these differences suggests a strategic component to this behavior. The absence of significant differences in early rounds, followed by emerging disparities in later stages of interaction, indicates a possible exploitation pattern that develops over time. This temporal dimension of trustworthiness aligns with prior research suggesting that dark personality traits may manifest most strongly after establishing a baseline relationship (Jones and Paulhus 2009). This finding extends previous work on dark traits in one-shot economic games (Zhao and Smillie 2015; Thielmann and Hilbig 2019) by demonstrating how these tendencies unfold over repeated interactions.

The significant interaction between D-factor, investment amount, round number, and opponent type reveals sophisticated strategic differences between high and low D-factor individuals. When facing the predictable human-like HMM opponent, high D-factor participants demonstrated a distinct pattern: they significantly decreased their returns over time for high-investment trials while maintaining relatively stable returns for low-investment trials. This selective exploitation strategy suggests a calculated approach to maximize gains while minimizing the risk of triggering retaliation from the investor. This pattern is particularly significant because it represents a form of Machiavellian exploitation that targets situations of high trust (indicated by higher investments) rather than indiscriminate exploitation across all conditions. By selectively reducing reciprocity in high-stake interactions, high D-factor individuals effectively exploit the trust placed in them when the potential gains are greatest. This finding aligns with the conceptualization of Machiavellianism as involving strategic, long-term orientation to personal gain (Jones and Paulhus 2009) and supports previous research indicating that dark personality traits are associated with strategic rather than impulsive exploitation (Gunnthorsdottir, McCabe, and Smith 2002). Notably, this strategic exploitation pattern was only observed with the more predictable human-like HMM opponent, not with the volatile opponent. This distinction suggests that high D-factor individuals may be particularly adept at identifying and exploiting predictable social dynamics, while showing more caution in volatile or unpredictable social environments. This contextual sensitivity adds important nuance to our understanding of how dark personality traits manifest in economic decisions.

The significant four-way interaction involving investment, order of opponent presentation, D-factor, and round number further illuminates the adaptive nature of exploitation strategies. High D-factor participants who first encountered the stable (human-like) opponent showed decreasing returns over rounds for high and medium investments but maintained stable returns for low investments. In contrast, those who initially faced the volatile opponent reduced returns for low and medium investments while maintaining returns for high investments. This pattern suggests that high D-factor individuals rapidly adapt their exploitation strategies based on initial experiences. When first exposed to a predictable environment, they learn to exploit high-trust situations. Conversely, when first exposed to volatility, they adopt a more conservative strategy that maintains cooperation in high-stake interactions while reducing reciprocity in lower-risk situations. This adaptive learning demonstrates sophisticated social intelligence that may underlie the effectiveness of dark personality traits in navigating complex social environments. Low D-factor participants, regardless of the order in which they faced opponents, maintained relatively stable return rates across rounds and investment levels, suggesting a more consistent approach to reciprocity that is less influenced by strategic considerations or learning effects. This stability in cooperative behavior may reflect stronger adherence to fairness norms and less susceptibility to exploitative tendencies.

The analysis of opponent ratings revealed that D-factor scores significantly influenced how participants perceived their counterparts. High D-factor participants consistently rated their opponents lower on cooperativeness, trustworthiness, and desirability for future interaction, regardless of the opponent’s actual behavior. This negative perceptual bias is consistent with research suggesting that individuals with dark personality traits may have distorted social perceptions that justify exploitation (Moshagen, Hilbig, and Zettler 2018; Zettler, Moshagen, and Hilbig 2021). The interaction between D-factor, game order, and opponent type for trust ratings suggests complex differences in how high and low D-factor individuals update their social perceptions based on experience. While low D-factor participants showed increased trust in human-like opponents from first to second game, high D-factor participants showed decreased trust in volatile opponents. This differential updating may reflect differences in attribution processes: low D-factor individuals may attribute positive interactions to stable traits of their partner, while high D-factor individuals may be more sensitive to negative interactions and use them to justify subsequent exploitation. These findings extend beyond economic behavior to suggest that the D-factor influences the entire process of social perception and decision-making. The negative bias in opponent evaluation may serve as a cognitive mechanism that facilitates exploitation by reducing empathic concern and moral constraints associated with harming a positively regarded other.

The analysis of final-round behavior, where participants knew there would be no further interactions, provides insight into purely self-interested tendencies without the strategic considerations of reputation building. In these last rounds, high D-factor participants returned significantly lower amounts than low D-factor participants, both in absolute terms and as a percentage of investment. This finding represents a clear manifestation of the maximizing self-interest component of the D-factor when strategic constraints are removed. The last-round effect essentially transforms the trust game into a dictator game, where participants can freely decide how much to return without fear of future consequences. The significant D-factor difference in this context aligns with previous research showing associations between the D-factor and selfish behavior in dictator games (Moshagen, Hilbig, and Zettler 2020). The consistency across economic paradigms strengthens the conclusion that the D-factor represents a stable tendency toward self-maximization when social constraints are minimal.

A particularly interesting finding was that despite returning lower percentages, high D-factor participants did not achieve significantly higher total payoffs compared to low D-factor participants. This seemingly paradoxical result can be explained by the adaptive nature of the HMM opponent, which reduced investments in response to lower returns from high D-factor participants. This dynamic illustrates how exploitative strategies may fail to maximize long-term gains in environments with responsive counterparts, as initial exploitation triggers defensive reactions that ultimately limit future opportunities for gain. This outcome has important implications for understanding the evolutionary stability of dark personality traits. While the D-factor may confer advantages in certain one-shot interactions or where reputation effects are minimal, its effectiveness as a long-term strategy in repeated interactions with responsive partners appears limited. This aligns with theoretical accounts suggesting that dark personality traits may represent frequency-dependent strategies that are most beneficial when rare in a population (Mealey 1995), as widespread exploitation would trigger universal defensive responses that limit its effectiveness.

Theoretical and practical implications

Our findings have several important implications for personality psychology and behavioral economics. First, they demonstrate that the D-factor, as a unifying construct of dark personality traits, provides meaningful predictive power for understanding trustworthiness in economic exchanges. The convergent patterns of exploitation, negative social perception, and self-maximization across different measures support the conceptual coherence of the D-factor construct. Second, our results highlight the importance of considering the temporal dimension of trust and reciprocity. The emerging differences between high and low D-factor participants over repeated interactions suggest that single-round economic games may underestimate the influence of personality traits on economic behavior. Future research should continue to examine how personality influences behavioral trajectories rather than just static decision points. Third, the interaction between D-factor and opponent volatility provides insight into the contextual sensitivity of dark personality traits. The finding that high D-factor individuals modulate their exploitation strategies based on opponent predictability suggests sophisticated social intelligence rather than rigid antisocial tendencies. This nuance is important for developing more accurate models of how personality influences social decision-making across different environments. Finally, our findings have practical implications for promoting cooperation in economic exchanges. The fact that high D-factor participants received lower investments over time indicates that exploitative strategies trigger defensive responses that ultimately limit opportunities for mutual gain. Interventions that highlight these long-term consequences might help redirect self-interested motivations toward more sustainable cooperative strategies.

Limitations and future directions

Several limitations of the current study suggest directions for future research. First, while we observed clear behavioral differences between high and low D-factor individuals, our design cannot determine which specific aspects of the D-factor (e.g., Machiavellianism, psychopathy, or narcissism) drive these effects. Future studies could include measures of these specific traits alongside the D-factor to examine their relative contributions. Second, our use of HMM opponents provided excellent experimental control but may limit ecological validity. Future research could examine D-factor influences in fully human interactions to capture the richer social dynamics of real-world trust building. Finally, while we found significant differences in behavior and perception, we did not explore the underlying affective or cognitive processes that mediate these effects. Future studies could incorporate measures of empathy, moral disengagement, or social value orientation to understand how dark personality traits influence the subjective experience of economic exchanges.

Conclusion

This study provides novel insights into how the Dark Factor of Personality influences behavior in repeated trust games. Our findings demonstrate that individuals with high D-factor scores exhibit systematic patterns of lower reciprocity that emerge most strongly in later rounds of interaction, particularly when facing predictable opponents and receiving high investments. These behavioral differences are accompanied by more negative perceptions of interaction partners, suggesting a comprehensive influence of dark personality traits on both social cognition and economic decision-making. The sophistication of exploitation strategies—adapting to opponent type, investment level, and interaction history—indicates that dark personality traits may involve complex social intelligence rather than simple antisocial tendencies. However, the failure of these exploitative strategies to yield higher total payoffs highlights the self-limiting nature of exploitation in responsive social environments. These findings bridge the gap between personality psychology and behavioral economics, demonstrating how stable personality traits manifest in dynamic economic exchanges. They extend previous research on dark personality traits by revealing how exploitation unfolds over time and varies across contexts. Future research should continue to explore the cognitive and affective mechanisms underlying these behavioral patterns and examine how interventions might promote cooperation even among individuals with stronger exploitative tendencies. Understanding the relationship between the D-factor and trustworthiness has significant implications for promoting cooperative outcomes in economic and social interactions. By recognizing how dark personality traits influence trust dynamics, we can develop more effective strategies for fostering cooperation and limiting the social costs of exploitation.

References

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

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1): 122–42. https://doi.org/10.1006/game.1995.1027.

Bohnet, Iris, and Steffen Huck. 2004. "Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change." *American Economic Review* 94 (2): 362–66. https://doi.org/10.1257/0002828041301506.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Charness, Gary, Ramon Cobo-Reyes, and Natalia Jiménez. 2008. "An Investment Game with Third-Party Intervention." *Journal of Economic Behavior & Organization* 68 (1): 18–28. https://doi.org/10.1016/j.jebo.2008.02.006.

Fiedler, Martin, Ernan Haruvy, and Sherry Xin Li. 2011. "Social Distance in a Virtual World Experiment." *Games and Economic Behavior* 72 (2): 400–426. https://doi.org/10.1016/j.geb.2010.09.004.

Gong, Xiaoxiao, Inti A. Brazil, Luke J. Chang, and Alan G. Sanfey. 2019. "Psychopathic Traits Are Related to Diminished Guilt Aversion and Reduced Trustworthiness During Social Decision-Making." *Scientific Reports* 9 (1): 7307. https://doi.org/10.1038/s41598-019-43727-0.

Gunnthorsdottir, Anna, Kevin McCabe, and Vernon Smith. 2002. "Using the Machiavellianism Instrument to Predict Trustworthiness in a Bargaining Game." *Journal of Economic Psychology* 23 (1): 49–66. https://doi.org/10.1016/S0167-4870(01)00067-8.

Ibáñez, María I., Gerardo Sabater-Grande, Ivan Barreda-Tarrazona, Laura Mezquita, Sara López-Ovejero, Héctor Villa, Pandelis Perakakis, Generos Ortet, Aurora García-Gallego, and Nikolaos Georgantzís. 2016. "Take the Money and Run: Psychopathic Behavior in the Trust Game." *Frontiers in Psychology* 7: 1866. https://doi.org/10.3389/fpsyg.2016.01866.

Jones, Daniel N., and Delroy L. Paulhus. 2009. "Machiavellianism." In *Handbook of Individual Differences in Social Behavior*, edited by Mark R. Leary and Rick H. Hoyle, 93–108. The Guilford Press.

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

Mealey, Linda. 1995. "The Sociobiology of Sociopathy: An Integrated Evolutionary Model." *Behavioral and Brain Sciences* 18 (3): 523–41. https://doi.org/10.1017/S0140525X00039595.

Moshagen, Morten, Benjamin E. Hilbig, and Ingo Zettler. 2018. "The Dark Core of Personality." *Psychological Review* 125 (5): 656–88. https://doi.org/10.1037/rev0000111.

———. 2020. "The Dark Core of Personality and Socially Aversive Psychopathology." *Journal of Personality* 88 (6): 1253–70. https://doi.org/10.1111/jopy.12577.

Rosenberger, Lisa A., Dimitris Tsivilis, and Florian Müller. 2019. "Fairness Norm Violations in Antisocial Offenders During Trust Games: A Repeated Trust Game Investigation." *Personality and Individual Differences* 136: 113–21. https://doi.org/10.1016/j.paid.2018.01.007.

Singmann, Henrik, Ben Bolker, Jake Westfall, Frederik Aust, Mattan S. Ben-Shachar, Søren Højsgaard, John Fox, et al. 2022. "Afex: Analysis of Factorial Experiments." https://CRAN.R-project.org/package=afex.

Thielmann, Isabel, and Benjamin E. Hilbig. 2019. "No Gain Without Pain: The Psychological Costs of Dishonesty." *Journal of Economic Psychology* 71: 126–37. https://doi.org/10.1016/j.joep.2018.06.001.

Zettler, Ingo, Morten Moshagen, and Benjamin E. Hilbig. 2021. "Stability and Change: The Dark Factor of

Personality Shapes Dark Traits." *Social Psychological and Personality Science* 12 (7): 974–83. https://doi.org/10.1177/1948550620953288.

Zhao, Kaiming, and Luke D. Smillie. 2015. "The Role of Interpersonal Traits in Social Decision Making: Exploring Sources of Behavioral Heterogeneity in Economic Games." *Personality and Social Psychology Review* 19 (3): 277–302. https://doi.org/10.1177/1088868314553709.