The Price of Dark Traits: Strategic Exploitation and Its Limitations in Repeated Trust Games

2025-07-04

Abstract:

Trust and cooperation are fundamental to human social interaction, with personality traits significantly influencing economic decision-making. This study introduces a novel approach using adaptive computational agents to provide the first examination of how the Dark Factor of Personality (D-factor)—the unifying core of malevolent traits—affects trustworthiness and strategic exploitation in dynamic, repeated trust games. After pre-screening 1,243 participants, we selected individuals with high and low D-factor scores. Participants (N=183) played two 25-round trust games as trustees against programmed Hidden Markov Model (HMM) investors derived from human data but differing in strategic reactivity: one 'human-like' HMM exhibited a characteristic 'sticky' mid-trust state (potentially vulnerable to exploitation), while a 'responsive' HMM reacted more decisively and immediately to trustee returns. Consistent with the D-factor concept, high-D participants returned lower proportions, particularly later in the interactions. High-D individuals also demonstrated sophisticated strategic adaptation, significantly decreasing reciprocity over time only when receiving high investments from the more passively exploitable 'human-like' opponent. Crucially, this exploitative strategy proved self-limiting: high-D participants did not achieve higher total payoffs overall, as the 'responsive' HMM adaptively reduced investments. These findings uniquely demonstrate how D-factor manifests strategically in dynamic exchanges and reveal the critical role of opponent responsiveness in constraining exploitation, highlighting both the sophistication and the boundaries of dark personality expression in social interactions.

Keywords : Interpersonal functioning; Dark Factor of Personality; Strategic exploitation; Trust-based Cooperation; Hidden Markov Models

1 Introduction

Trust and cooperation are fundamental to human social interaction and economic exchange (Ostrom and Walker 2003; Yamagishi 2011). The trust game, introduced by Berg, Dickhaut, and McCabe (1995), provides a powerful and widely adopted tool for investigating the dynamics of trust and trustworthiness in controlled settings (Camerer 2003; Johnson and Mislin 2011).

While situational factors influence behavior, stable individual differences in personality traits are also significant predictors (Evans and Revelle 2008; Müller and Schwieren 2020). Among the most potent predictors of antisocial or self-interested behavior are the "dark" personality traits. Although often studied via the Dark Triad framework (Machiavellianism, Narcissism, Psychopathy; Paulhus and Williams (2002)), the considerable overlap between these traits (Zettler, Moshagen, and Hilbig 2021) spurred the development of the Dark Factor of Personality (D-factor) (Moshagen, Hilbig, and Zettler 2018). Proposed as the common core underlying various malevolent traits, D is defined 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). This unifying construct, incorporating elements like Machiavellianism and Psychopathy previously linked to reduced trustworthiness (Ibáñez et al. 2016; Gunnthorsdottir, McCabe, and Smith 2002), offers a parsimonious framework for understanding antisocial tendencies by representing the shared variance among dark traits (Benjamin E. Hilbig et al. 2021b; Zettler, Moshagen, and Hilbig 2021).

Research consistently links dark traits and D to reduced cooperation and honesty (Zhao and Smillie 2015; Thielmann and Hilbig 2019), particularly in one-shot economic games like the dictator game (Benjamin E. Hilbig et al. 2021a). However, these single-interaction studies cannot capture the dynamics crucial to real-world exchanges, such as reputation formation (Bohnet and Huck 2004) and strategic adaptation over time. Understanding how D influences behavior in repeated exchanges is critical, especially given that the strategic manipulation inherent in D (e.g., its Machiavellian component; Jones and Paulhus (2009)) may only fully manifest across multiple interactions. Yet, the specific impact of D in repeated trust games remains largely unexplored. Prior work using repeated games often involved different paradigms (e.g., sequential partners; Gong et al. (2019)) or focused on clinical samples and related constructs (Rosenberger, Tsivilis, and Müller 2019), leaving a gap in understanding how D operates strategically in sustained interactions within the general population.

The present study aims to fill this gap by examining how D-factor influences strategic trustworthiness across 25 rounds of a repeated trust game. A key challenge in studying dynamic social interactions is controlling the behavior of the interaction partner. To address this, we employ an innovative methodology using Hidden

Markov Model (HMM) agents as investors. HMMs (Rabiner 1989) are computational models ideal for capturing behavioral sequences driven by underlying states. Critically, our HMMs were not abstract models but were trained on large datasets of human behavior in prior trust games. This data-driven approach allows us to create standardized investors that exhibit realistic, human-like response patterns while enabling precise experimental manipulation of their interaction strategy – a blend of ecological validity and experimental control difficult to achieve otherwise (Macy and Willer 2002).

We leverage this methodology to directly test strategic adaptation related to D by manipulating opponent 'exploitability'. We developed two distinct HMM investor types based on typical human behavior patterns: a 'human-like' agent exhibiting a 'sticky' mid-trust state (requiring large return deviations to shift state), making it potentially vulnerable to gradual exploitation; and a 'responsive' agent specifically modified to eliminate this mid-state inertia, forcing rapid transitions based on the trustee's immediate returns. This controlled comparison between a passively exploitable versus a more reactive partner provides a novel way to assess if high-D individuals strategically adjust their behavior based on the opponent's interaction dynamics.

Therefore, this study offers several unique contributions: First, it provides the first examination of the unified D-factor within the dynamic context of a multi-round (25-round) repeated trust game. Second, it introduces and utilizes an innovative HMM-based methodology to experimentally manipulate opponent responsiveness/exploitability in a realistic manner. Finally, it investigates not only the enactment of exploitative strategies associated with D but also the potential limits of these strategies when facing adaptive partners. Understanding these dynamics has significant implications for bridging personality psychology and behavioral economics.

Based on the theoretical conceptualization of D and the dynamics of repeated games with varying opponent responsiveness, 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 individuals to show greater strategic adaptation to opponent type, with more pronounced exploitation of predictable opponents compared to responsive ones. Finally, we explore whether these behavioral patterns are accompanied by systematic differences in perception of opponents, with High-D individuals potentially showing more negative evaluations regardless of opponent behavior.

2 Methods

2.1 Participants

To form groups with distinct levels of the Dark Factor of Personality (D), 1,243 participants were pre-screened via the Prolific Academic platform (prolific.co) using the Dark Factor of Personality–16 (D16) – a concise 16-item unidimensional measure of the general D trait (range 16–80). Based on the scale developers' validation sample (N = 6,838), the D16 demonstrated strong internal consistency (Cronbach's α = .91) and excellent test–retest reliability (r = .90), with confirmatory factor analyses supporting its single-factor structure (Moshagen, Zettler, and Hilbig 2020). We selected two groups: High-D (90th percentile or higher, total score > 42, $N = 91$) and Low-D (10th percentile or lower, total score < 22, $N = 92$), yielding a final sample of 183 participants (44% female). These participants were then invited via Prolific to complete the main experiment.

A sample size of 180 participants was determined *a priori* through conducting a 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 (Human-like vs. Responsive), 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, see Thielmann, Spadaro, and Balliet (2020)), 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 at least 90% power to detect the hypothesized three-way interaction.

2.2 Design and Materials

The experiment employed a 2 (HMM Type: Human-like vs. Responsive) $\times$ 2 (D-Factor: High vs. Low) mixed design, with HMM Type as a within-subjects factor. Participants completed two phases, each involving a 25-round Repeated Trust Game (RTG), against either the Human-like or Responsive HMM investor, with the order counterbalanced. 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 [Redacted] Medical Faculty ethics commission (ID:S-708/2023) and the experiment was performed in accordance with the ethics board guidelines and regulations. All participants provided informed consent prior to their participation.

2.2.1 Repeated Trust Game (RTG)

Based on Berg, Dickhaut, and McCabe (1995), participants always acted as the trustee. In each round, the HMM investor received 20 units and decided how much (0-20) to invest. This investment was tripled, and the trustee decided how much of the tripled amount to return to the investor. Trustees knew that returning more than one-third resulted in a gain for the investor. Participants chose a representative icon; the opponent's icon changed between phases to simulate a new partner. To enhance realism, participants were not told opponents were computerised; simulated pairing delays and variable response latencies (5-10s) were included.

The mean age of participants was 33.1 years, with an 9.7-year 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. We collected data over multiple sessions between October and November 2024.

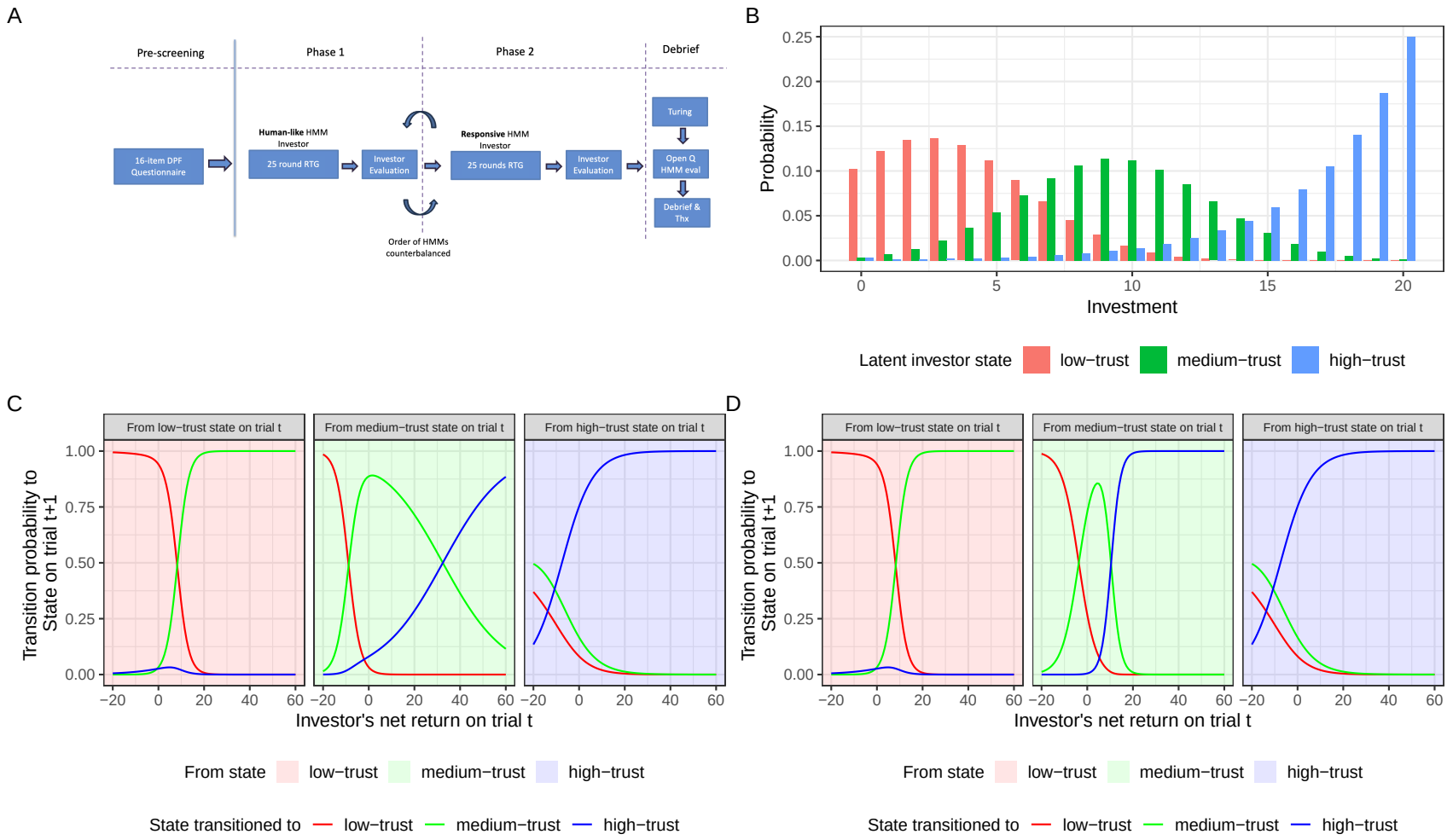

Figure 1: Panel A: A timeline of the experiment. The RTG is played in dyads, with participants always assigned the role of the trustee and the HMM agent that of the investor. The investor is endowed with 20 units at the start of each round. They need to decide how much of that endowment they want to 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 HMM 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 t+1 as a function of the net return (difference between the investment sent and the amount received in return) on trial t as shown in Panel C. Each plot in Panel C represents a different starting latent state on trial t, and each line represents the probability of transitioning to a particular state in trial t+1. The responsive HMM agent (Panel D) is much more likely to transition out of the mid-trust state. Transitions in the medium and high trust states were identical for both agents.

2.2.2 HMM Investor Construction

Computerised investors were implemented as Hidden Markov Models (HMMs) trained on an independent dataset of 388 10-round RTGs (see Supplement for details). This data-driven approach aimed to create investors mimicking human interaction patterns. The inferred HMM used three latent states ("low-trust", "medium-trust", "high-trust"), each associated with a learned probability distribution over possible investments (0-20 units; Figure 1.B). Transitions between states were probabilistic, modeled as a function of the investor's net return (amount returned - amount invested) in the previous round (Figure 1.C). For instance, higher net returns increased the probability of transitioning to higher trust states. The HMM started each 25-round game in the "medium-trust" state. Investor actions involved sampling an investment from the current state's distribution; the state for the next round was determined by sampling from the transition probabilities based on the previous round's net return. (See Supplement for full HMM specification).

2.2.3 HMM Investor Types

Two HMM types were used:

- Human-like HMM: Directly resulted from fitting the human dataset. It exhibited a tendency to remain in the "medium-trust" state unless net returns were substantially high or low (Figure 1.C).
- Responsive HMM: Created by adjusting the Human-like HMM's parameters to eliminate the 'stickiness' of the mid-trust state. Specifically, the probability of *remaining* in the medium-trust state was set near zero for non-nil net returns (Figure 1.D). This modification made the agent highly sensitive to trustee behavior, rapidly shifting to low- or high-trust states in response to exploitation or generosity, respectively. Other state parameters remained unchanged.

2.2.4 Other Measures:

After each RTG phase, participants rated the perceived cooperativeness, trustworthiness, and their willingness to play again with the investor (1-10 scales). Post-experiment measures included the Turing test (perceived humanness of opponents) and open-ended questions about strategy.

2.3 Procedure

Participants began by providing informed consent. They were then instructed that the study comprised two distinct phases, each involving interaction with a different player. The stated objective for participants was to accumulate the maximum number of points throughout all phases. Information regarding the specific number of rounds within each phase was not disclosed to them initially. During the concluding round (round

25) of both phases, participants received a visual cue in the form of a flashing message, indicating that it was the final round of that particular game.

The initial phase involved participating in a 25-round Repeated Trust Game (RTG). In this game, participants consistently played the role of the trustee, interacting with the same designated investor across all 25 rounds. The second phase replicated the setup of the first phase exactly, involving another 25-round RTG with the participant again acting as the trustee. The key difference in this phase was the introduction of a new opponent.

Upon the completion of the RTG in each separate phase, participants were prompted to provide the investor ratings. They rated the perceived cooperativeness and trustworthiness of the co-player they had just interacted with, and also indicated their willingness to engage in future games with that same player. These ratings were captured on a scale ranging from 1 to 10, where 10 signified the most positive assessment.

After finishing both game phases, participants were asked for their perception regarding the nature of the other players – specifically, whether they believed them to be human or computer agents (Turing test). They were also requested to articulate the strategy they employed during each of the two games. The process concluded with a debriefing session, during which participants were thanked for their involvement in the study.

2.4 Statistical Analysis

To analyse participants' behavior in the RTG, we employed several complementary statistical approaches. Our primary analysis used linear mixed effects modelling to examine how percentage returns (proportion of tripled investment returned to investor) varied as a function of experimental factors. The model included Opponent type (Human-like vs. Responsive HMM), opponent presentation order (Order: Responsive First vs. Human-like First), Investment amount, round number, and D-factor group (High vs. Low) as fixed effects, along with all their interactions. Random effects included participant-wise intercepts and slopes for Investment. This approach allowed us to account for the nested structure of the data while examining how D-factor influenced behavior across conditions.

The model was estimated using the `afex` package (Singmann et al. 2022) in R with Kenward-Roger approximation for degrees of freedom. We Z-transformed the Investment variable to facilitate interpretation of main effects in the presence of interactions. For significant interactions, we conducted planned contrasts using the `emmeans` package with Sidak corrections to control familywise error rates. Model complexity was constrained by reliable estimation considerations, resulting in an optimal random effects structure (Matuschek

et al. 2017). Similar approaches were used for analysing HMM agent investments and participant ratings of co-players.

Participants' perceptions (ratings of cooperativeness, trustworthiness, willingness to play again) were analysed with LME models including D-factor level, HMM type, game order, and interactions as fixed effects, with participant-wise random intercepts.

To examine temporal dynamics within the game, the main interaction period (rounds 1-24) was divided into three equal, consecutive periods: early (rounds 1-8), middle (rounds 9-16), and late (rounds 17-24), representing the beginning, middle, and end of the game, respectively. Analyses compared key variables between groups across these periods. For the final round of each game, where no future interactions were anticipated (resembling a one-shot dictator game), we compared absolute returns and percentage returns between High-D and Low-D groups using both Welch's t-tests and non-parametric Wilcoxon rank-sum tests. The latter was included given the smaller sample size for this single-round comparison and to ensure robustness of our findings regardless of distributional assumptions. Finally, we analysed total payoffs earned by participants to determine whether behavioral differences resulted in differing economic outcomes. For debrief questions, we calculated the percentage of participants who believed they played against human opponents or were uncertain about their opponent's nature, providing insight into the ecological validity of our HMM agents. Full model specifications and additional analytical details are provided in the supplement.

2.5 Data, Materials, and Code Availability

The data and analysis code that support the findings of this study are available at a public GitHub repository. The repository contains all materials needed to reproduce the analyses, including anonymized data, code (in R), and a README file. To maintain anonymity during the peer-review process, the link will be provided upon acceptance or can be shared upon request to the editorial office.

3 Results

3.1 Mean investment and return per round

On average, investments and returns 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-D versus Low-D participants across all rounds, we observed several behavioral differences.

High-D 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 participants returning approximately 2-4 percentage points less of the tripled investment.

To examine how trust behavior evolved over time, we divided the 25-round game into three approximately equal periods—early (rounds 1–8), mid (rounds 9–16), and late (rounds 17–24). This tripartite split allows us to track temporal dynamics while ensuring each period contains a similar number of rounds. The final round (round 25) was excluded due to endgame effects that could distort typical behavior. 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.

As shown in Figure 2, 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.

3.2 Last Round Analysis

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.

3.3 Total Payoff Analysis

Finally, we analysed 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).

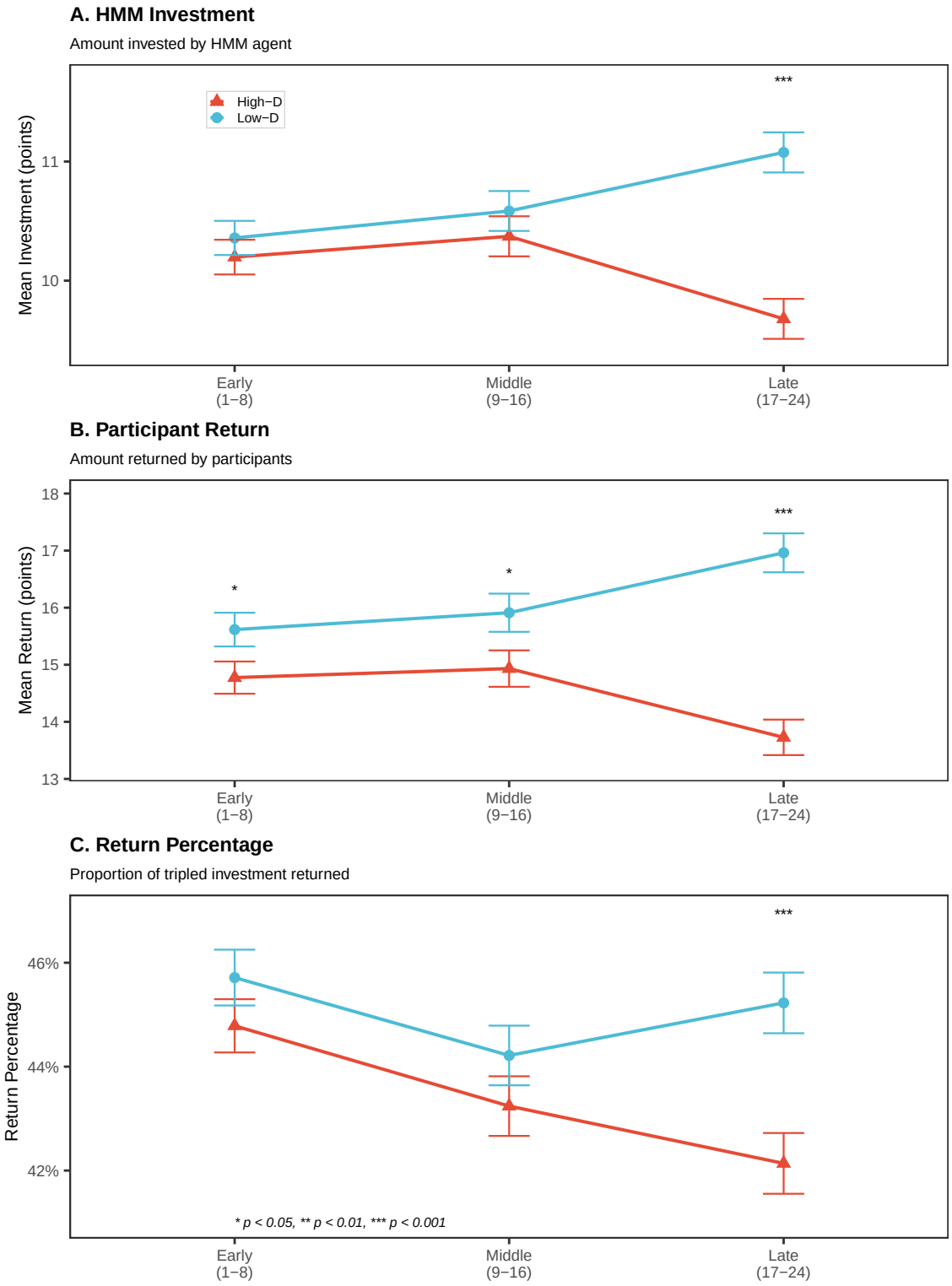

Figure 2: Investment and return patterns across both HMM types and over the 3 game periods by D-factor level. (A) Mean investments from HMM agents decrease for High-D participants in late game periods. (B) Absolute returns similarly decrease for High-D participants. (C) Return percentages show High-D participants consistently return a smaller proportion of received investments, with the difference becoming significant in later periods. Error bars represent standard errors of the mean. This pattern demonstrates how High-D participants' decreasing reciprocity triggers defensive responses from HMM agents.

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.

3.4 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.

3.4.1 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.

3.4.2 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 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 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).

3.4.3 D-Factor, Opponent Type and Investment Interaction

Low-D 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 Responsive HMM (3.1%, $p = 0.000$).

In contrast, High-D participants did *not* show significant reciprocity with either the Human-like HMM (slope = 0.008, $p = 0.314$) or the Responsive HMM (slope = 0.005, $p = 0.509$), indicating their returns were less

influenced by investment amount.

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

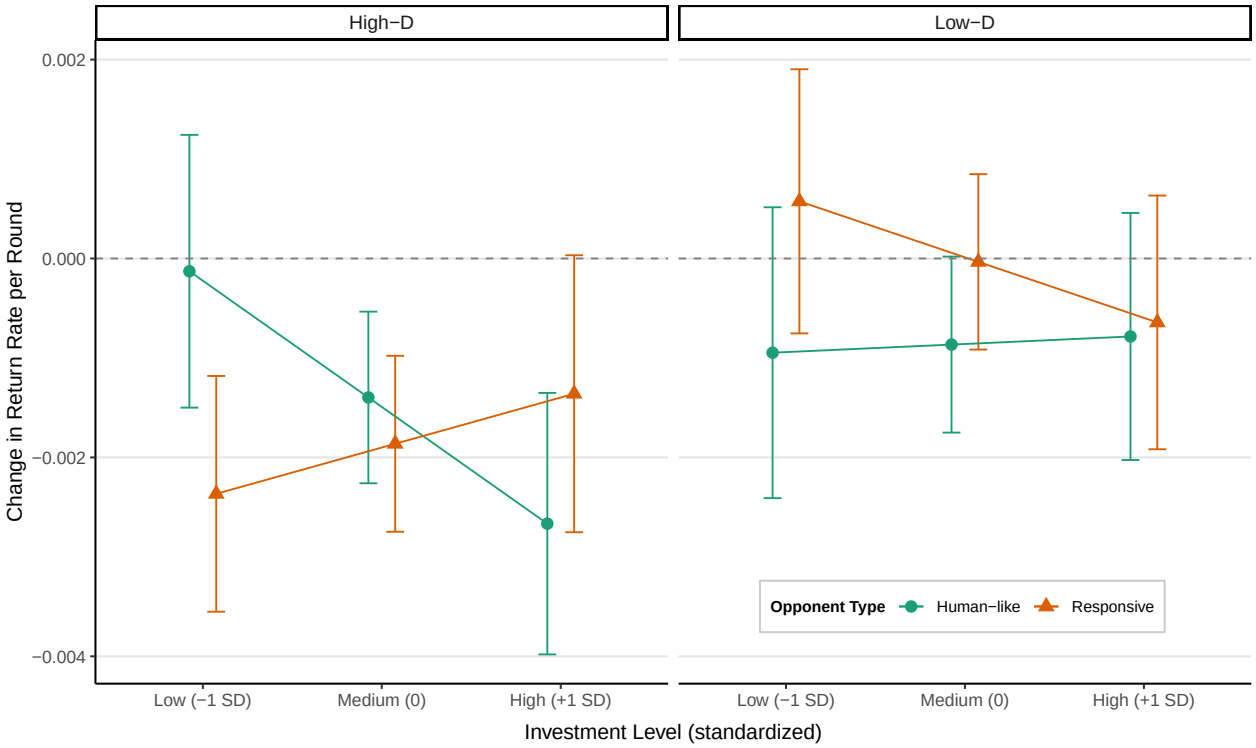

Figure 3: Changes in return rate over successive rounds by investment level, opponent type, and D-factor group. Bars represent the estimated slope coefficients (change in return percentage per round) with error bars showing 95 percent confidence intervals. Negative values indicate decreasing reciprocity over time. High-D participants show more pronounced decreases in return rates, particularly with the human-like opponent at high investment levels, suggesting they learned strategic exploitation of predictable, trusting partners.

Analysis of the significant four-way interaction ($F(1, 8215.95) = 5.92$, $p = .015$), visualised in Figure 3, revealed that only High-D 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 participants nor High-D participants facing the responsive opponent showed this strategic pattern. This suggests High-D participants specifically exploit a weakness in Human-like opponent strategies (the propensity to remain in a mid-trust state even after receiving lower returns) by systematically reducing reciprocity over time on the most lucrative (high-investment) trials.

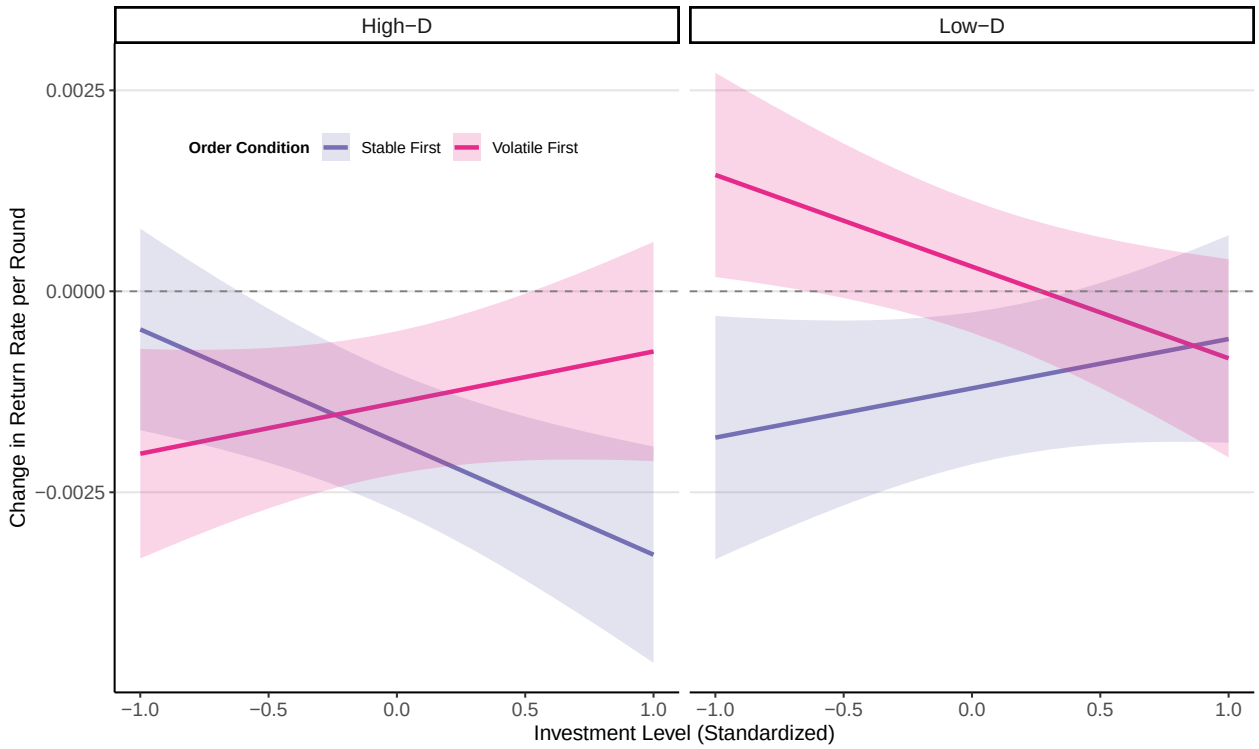

Figure 4: Effect of game order on changes in return rate across rounds at different investment levels. Lines represent the estimated slope coefficients (change in return rate per round) with shaded areas showing 95 percent confidence intervals. Negative values indicate decreasing reciprocity over time. High-D participants who played against the stable opponent first showed strategic exploitation primarily at high investment levels, while those who faced the volatile opponent first showed different strategic patterns. Low D-factor participants maintained more consistent return rates regardless of order condition.

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

The significant four-way interaction involving investment amount, order of opponent presentation (responsive 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. As seen in Figure 4, 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 an exploitable partner. In contrast, High-D participants who faced the *responsive 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.

3.5 Analysis of opponent ratings

Beyond behavioral measures, we also analysed participants' explicit evaluations of their opponents using linear mixed-effects models. These models assessed ratings of cooperativeness, willingness to play again, and trust based on participants' D-level, the opponent type faced, and the order of games (see Figure 5 for a summary).

3.5.1 Cooperative Ratings

Linear mixed-effects analysis revealed a significant main effect of D-level ($F(1, 179) = 7.35$, $p = .007$), with Low-D participants rating their opponents as more cooperative than High-D 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 Responsive opponents ($t(179) = 2.31$, $p = .022$).

3.5.2 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 $\times$ Game Order interaction ($F(1, 179) = 4.05$, $p = .046$). Post-hoc analyses revealed that in the first game, High-D participants were significantly less willing to play again with their opponents compared to Low-D 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 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 participants maintained consistent ratings across games ($t(179) = 1.21$, $p = .229$).

3.5.3 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 participants rated Human-like opponents as significantly less trusting than Low-D participants in the first game ($t(314.81) = -2.96$, $p = .003$). In contrast, High-D participants rated Responsive opponents as significantly less trusting than Low-D participants in the second game ($t(314.81) = -2.70$, $p = .007$). Low-D 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 participants showed a significant *decrease* in trust ratings for Responsive opponents from the first to the second game ($t(314.81) = 2.47$, $p = .014$). Additionally, High-D participants in the second game differentiated between opponent types, rating Human-like opponents as significantly more trusting than Responsive 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 participants showed greater sensitivity to game order, with more pronounced decreases in ratings from first to second game, while High-D participants demonstrated greater discrimination between opponent types in their trust ratings.

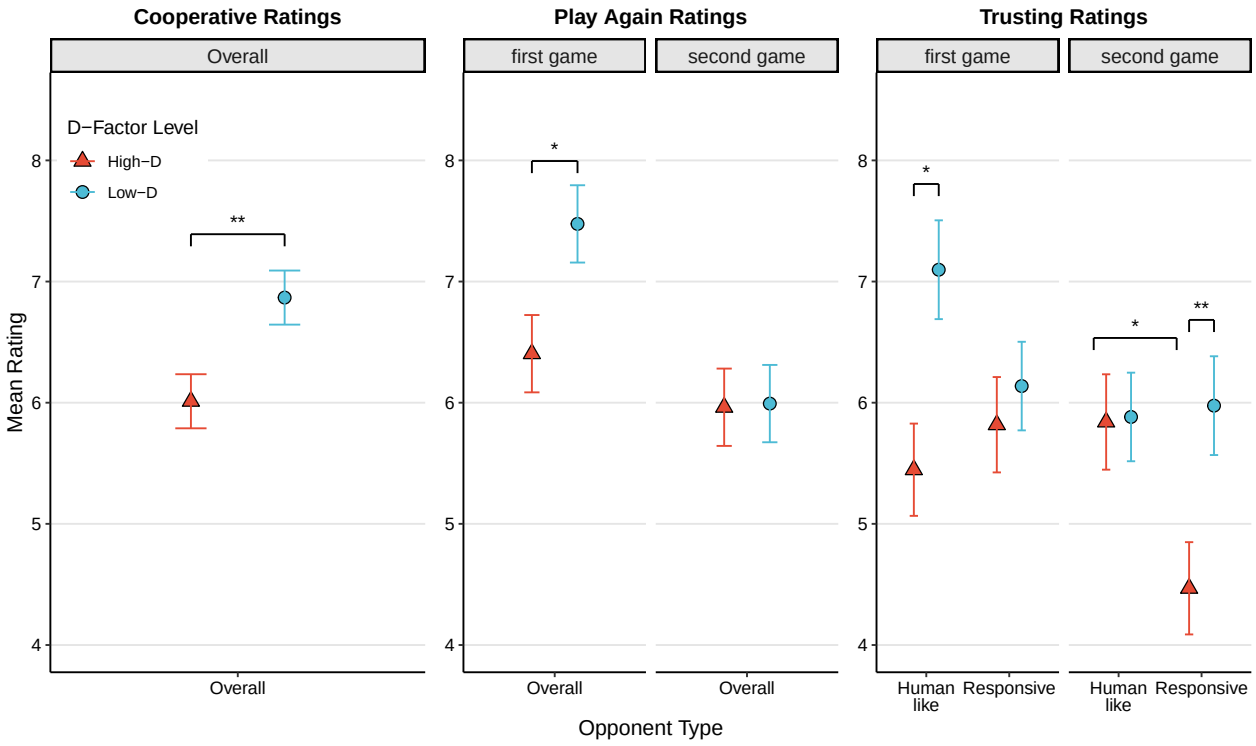

Figure 5: Averages and standard errors of the participants ratings of the opponent (y-axis) by each game and D-factor group for each opponent (x-axis). The left panel represents participants' perception of HMM cooperativeness, the right one indicates perceived HMM trust rating, and the middle panel shows the participants' willingness to play again with the same HMM. Cooperation, trust perception, and willingness to play again ratings were generally lower for the high DFP group.

3.6 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. There was no significant difference between the proportion of correct answers between High-D and Low-D groups.

4 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 scores. High-D participants consistently returned lower proportions of investment compared to their Low-D 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 and as participants learned the co-player's contingencies. 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 individuals. When facing the predictable Human-like HMM opponent, High-D 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 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 Responsive opponent. This distinction suggests that High-D individuals may be particularly adept at identifying and exploiting predictable social dynamics, while showing more caution in Responsive or unpredictable social environments. This contextual sensitivity adds important nuance to our understanding of how dark personality traits manifest in economic decisions. An alternative, though not mutually exclusive, explanation could be that High-D individuals do not necessarily

possess more sophisticated social intelligence, but rather a different learning model. For instance, they may be faster to abandon pro-social norms when they detect environmental predictability, a possibility that could be tested in future work by modeling individual learning rates. This distinction is important, as it reframes the behavior from one of pure strategic superiority to one of differential adaptation to social cues.

The significant four-way interaction involving investment, order of opponent presentation, D-factor, and round number further shows the adaptive nature of exploitation strategies. High-D 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 Responsive opponent reduced returns for low and medium investments while maintaining returns for high investments. This pattern suggests that High-D 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 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 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 individuals update their social perceptions based on experience. While Low-D participants showed increased trust in human-like opponents from first to second game, High-D participants showed decreased trust in Responsive opponents. This differential updating may reflect differences in attribution processes: Low-D individuals may attribute positive interactions to stable traits of their partner, while High-D 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 participants returned significantly lower amounts than Low-D 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 (Benjamin E. Hilbig et al. 2021a). 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 participants did not achieve significantly higher total payoffs compared to Low-D 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 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.

4.1 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 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 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 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.

4.2 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 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, and potentially explore whether we can replicate these results when including participants with the whole range of D-scores. 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. Furthermore, our participant sample, while diverse in country of origin, was recruited through an online platform and consisted primarily of individuals from Western countries. Future research should aim to replicate these findings in different cultural and socio-economic contexts to establish the broader generalizability of the D-factor's influence on strategic social behavior

5 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 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.

Evans, Anthony M., and William Revelle. 2008. "Survey and Behavioral Measurements of Interpersonal Trust." *Journal of Research in Personality* 42 (6): 1583–99. https://doi.org/10.1016/j.jrp.2008.07.011.

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.

Hilbig, Benjamin E, Isabel Thielmann, Sina A Klein, Morten Moshagen, and Ingo Zettler. 2021a. "The Dark Core of Personality and Socially Aversive Psychopathology." *Journal of Personality* 89 (2): 216–27. https://doi.org/10.1111/jopy.12577.

Hilbig, Benjamin E., Ingo Zettler, Morten Moshagen, and Isabel Thielmann. 2021b. "Theoretical and Empirical Dissociations Between the Dark Factor of Personality and Low Honesty-Humility." *Journal of Research in Personality* 95: 104154. https://doi.org/10.1016/j.jrp.2021.104154.

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.

Johnson, Noel D., and Alexandra A. Mislin. 2011. "Trust Games: A Meta-Analysis." *Journal of Economic Psychology* 32 (5): 865–89. https://doi.org/10.1016/j.joep.2011.05.007.

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. New York, NY: The Guilford Press.

Macy, Michael W., and Robert Willer. 2002. "From Factors to Actors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology* 28: 143–66. https://doi.org/10.1146/annurev.soc.28.110601.141117.

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.

Moshagen, Morten, Ingo Zettler, and Benjamin E. Hilbig. 2020. "Measuring the Dark Core of Personality." *Psychological Assessment* 32 (2): 182–96. https://doi.org/10.1037/pas0000778.

Müller, Julia, and Christiane Schwieren. 2020. "Big Five Personality Factors in the Trust Game." *Journal of Business Economics* 90 (1): 37–55. https://doi.org/10.1007/s11573-019-00928-3.

Ostrom, Elinor, and James Walker, eds. 2003. *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*. Russell Sage Foundation Series on Trust. New York: Russell Sage Foundation.

Paulhus, Delroy L., and Kevin M. Williams. 2002. "The Dark Triad of Personality: Narcissism, Machiavellianism, and Psychopathy." *Journal of Research in Personality* 36 (6): 556–63. https://doi.org/10.1016/S0092-6566(02)00505-6.

Rabiner, Lawrence R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77 (2): 257–86. https://doi.org/10.1109/5.18626.

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.

Thielmann, Isabel, Giuliana Spadaro, and Daniel Balliet. 2020. "Personality and Prosocial Behavior: A Theoretical Framework and Meta-Analysis." *Psychological Bulletin* 146 (1): 30–90. https://doi.org/10.1037/bul0000217.

Yamagishi, Toshio. 2011. *Trust: The Evolutionary Game of Mind and Society*. Evolutionary Psychology. New York: Springer. https://doi.org/10.1007/978-4-431-54005-9.

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.