The Power of Perception in Human-AI Interaction:
Investigating Psychological Factors and Cognitive Biases that Shape User Belief and Behavior

by

Eunhae Lee

Bachelor of Arts, Bryn Mawr College (2015)

Submitted to the Integrated Design and Management Program and
the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2024

Authored by:	Eunhae Lee Integrated Design and Management and Department of Electrical Engineering and Computer Science August 16, 2024
Certified by:	Manish Raghavan Assistant Professor of EECS, Thesis Supervisor
Certified by:	Pattie Maes Professor of the Media Lab, Thesis Supervisor
Accepted by:	Joan Rubin Executive Director, System Design and Management
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

The Power of Perception in Human-AI Interaction: Investigating Psychological Factors and Cognitive Biases that Shape User Belief and Behavior

by

Eunhae Lee

Submitted to the Integrated Design and Management Program and
the Department of Electrical Engineering and Computer Science
on August 16, 2024 in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

This thesis investigates the psychological factors that influence belief in AI predictions, comparing them to belief in astrology- and personality-based predictions, and examines the "personal validation effect" in the context of AI, particularly with Large Language Models (LLMs). Through two interconnected studies involving 238 participants, the first study explores how cognitive style, paranormal beliefs, AI attitudes, and personality traits impact perceptions of the validity, reliability, usefulness, and personalization of predictions from different sources. The study finds a positive correlation between belief in AI predictions and belief in astrology- and personality-based predictions, highlighting a "rational superstition" phenomenon where belief is more influenced by mental heuristics and intuition than by critical evaluation. Interestingly, cognitive style did not significantly affect belief in predictions, while paranormal beliefs, positive AI attitudes, and conscientiousness played significant roles. The second study reveals that positive predictions are perceived as significantly more valid, personalized, reliable, and useful than negative ones, emphasizing the strong influence of prediction valence on user perceptions. This underscores the need for AI systems to manage user expectations and foster balanced trust. The thesis concludes with a proposal for future research on how belief in AI predictions influences actual user behavior, exploring it through the lens of self-fulfilling prophecy. Overall, this thesis enhances understanding of human-AI interaction and provides insights for developing AI systems across various applications.

Thesis supervisor: Manish Raghavan
Title: Assistant Professor of EECS

Thesis supervisor: Pattie Maes
Title: Professor of the Media Lab

Acknowledgments

First, I would like to express my sincere gratitude to my advisors, Professor Pattie Maes and Manish Raghavan, for their support, guidance, and encouragement throughout my research. Their expertise and support have greatly contributed to the completion of this thesis.

I extend my sincere thanks to my collaborator, Pat Pataranutaporn, for his invaluable assistance, stimulating discussions, and boundless energy throughout this process. His support made this journey not only intellectually rewarding but also deeply enjoyable. I also want to thank Judith Amores at Microsoft Research for her valuable feedback and enthusiastic contribution.

I am grateful to the Fluid Interfaces Group at the MIT Media Lab for providing the essential resources needed to conduct my research and for offering constructive feedback along the way. Additionally, I would like to express my appreciation to Jinjie Liu at the Institute for Quantitative Social Science at Harvard University, for sharing her insightful perspective on the analysis, which greatly enriched this work.

I also wish to acknowledge the academic advisors and administrators in the IDM/SDM program and the EECS department, whose guidance has been invaluable as I navigated the winding path of pursuing a dual Master's degree.

My deepest thanks go to my family for their endless love and support. To my parents, Jong In Lee and Ji Young Kim, your encouragement and belief in me have been the driving force behind my success. I am also profoundly grateful to my partner, Bryan Ibarra Wong, for his patience, understanding, and unwavering support during the most challenging moments of this journey.

Finally, I would like to thank everyone who contributed to the successful completion of this thesis, including the advisors and colleagues at the MIT AgeLab who provided early guidance, my friends in the IDM program, and the external colleagues and friends who offered fresh perspectives that enriched my work along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid advancement of artificial intelligence (AI) has sparked a complex dialogue that intertwines technological progress with elements of quasi-religious narratives and psychological manipulation. As AI systems, particularly generative AI and large language models (LLMs), become more integrated into daily life, understanding the psychological dynamics that influence human-AI interaction is increasingly crucial. This thesis builds on two interconnected studies that explore these dynamics through different but complementary lenses: the concept of "rational superstition" in AI and the "personal validation effect" in human-AI interactions.

The first study investigates how psychological factors influence belief in AI predictions, drawing unexpected parallels between AI and astrology. Despite AI's foundation in scientific and rational methodologies, public perception often mirrors belief systems associated with pseudoscience, such as astrology. This phenomenon, termed "rational superstition" [1], suggests that people may trust AI predictions not because of critical evaluation but due to mental heuristics and intuitive thinking. By comparing responses to AI predictions to those derived from astrology and personality psychology, this study examines how cognitive style, paranormal beliefs, AI attitudes, and personality traits influence perceptions of AI's validity, reliability, usefulness, and personalization.

Building on the insights from the first study, the second study delves into the "personal validation effect" in the context of LLM use. Here, we explore how the valence of AI-generated predictions—whether positive or negative—affects users' perceptions of these predictions' validity, reliability, usefulness, and personalization. This study uncovers the significant impact of positive AI feedback, revealing that users are more likely to perceive favorable predictions as more accurate and personalized, even when these predictions are demonstrably false. This effect highlights the potential for AI systems to manipulate user perceptions through positive reinforcement.

Together, these studies offer a comprehensive examination of how belief in AI predictions is shaped by a combination of cognitive biases, psychological factors, and the framing of AI outputs. The first study provides insights on the psychological underpinnings of trust in AI, while the

second study extends this understanding by exploring specific mechanisms by which AI systems can manipulate these beliefs. By integrating these findings, this thesis aims to contribute to the broader discourse on AI ethics, transparency, and the responsible design of AI systems that foster appropriate trust and skepticism.

Finally, this thesis outlines ongoing and future work aimed at extending our understanding of how people perceive and believe in AI predictions, focusing on the underlying psychological factors. I discuss a proposed study that would explore how these perceptions influence actual behavior, drawing on the well-known phenomenon of the "self-fulfilling prophecy." Specifically, it would investigate how AI-generated predictions about individuals may shape their actions in ways that ultimately fulfill those predictions.

The flow of the research presented in this thesis reflects a logical progression from understanding the broader psychological factors influencing belief in AI predictions to exploring how AI predictions may impact actual behavior and outcomes. This integrated approach underscores the importance of considering both the cognitive and affective dimensions of human-AI interaction, ultimately informing strategies for designing AI systems that are both effective and ethically sound.

Chapter 2

Super-intelligence or Superstition? Exploring Psychological Factors Underlying Unwarranted Belief in AI Predictions

2.1   Introduction and Related Works

Technology is often associated with scientific advancement and frequently viewed in opposition to superstition. However, the recent rapid advancements in artificial intelligence (AI) have given rise to quasi-religious perspectives among some leaders in the tech industry [2]. In fact, as generative AI and large language models (LLMs) continue to evolve, a contemporary critique posits that the narrative surrounding their development bears striking resemblances to religious discourse [3]. Promoting concepts such as superintelligence and sentience, prominent figures in the tech industry have been observed making prophetic claims regarding AI's potential to either salvage or obliterate humanity [4, 5]. These claims are often based on personal speculation and subjective views rather than rigorous scientific research. Nonetheless, empirical research has demonstrated that these popular narratives are mirrored in public perceptions of AI, frequently manifesting as exaggerated utopian or dystopian visions [6, 7]. This intersection of technological advancement and quasi-religious perspective presents a complex landscape for scholars to navigate, as it blurs the boundaries between scientific progress and speculation.

While AI has been developed through predictive algorithms and statistical pattern recognition, exemplifying advancements in mathematical and rational thinking, public perception of AI may draw unexpected parallels with astrology [8–10]. Despite astrology representing an opposing episte-

mology based on pseudoscientific methodologies antithetical to empirical reasoning, both AI and astrology function as prediction-generating mechanisms helping people make sense of the uncertainty in the world [8, 9]. Anthropologist Christophe Lazaro suggests that AI, powered by data, has become a new mode of speculative thinking, with algorithms taking on the role of oracles in contemporary societies, framing AI as "artificial divination" [9, p. 128].

On the other hand, AI, unlike astrology, presents an aura of scientific legitimacy, backed by a vast body of scientific research in computer science, mathematics, cognitive science, and more. This foundation contributes to the belief that AI systems are rational and objective. However, studies have shown that algorithms could produce unpredictable outcomes, biased results, and inherit implicit societal biases [11–13]. Moreover, despite ongoing efforts to increase transparency, much of the inner workings of AI systems remain "black boxes" [14–17]. Such misconceptions regarding AI may engender a form of naive technological optimism and irrational trust in the system, which Wilson refers to as "rational superstition" [1]. Scholars have investigated the concept of algorithmic appreciation—the propensity to rely on algorithms for decision-making processes—which may lead to an overreliance on AI in critical decision-making contexts [18, 19].

Popular narratives that over-inflate the promises of AI and its capabilities is particularly alarming when considering the impact it has on mental models and perception of AI systems [7]. According to the dual process theory in psychology, people engage in two types of thinking—fast and intuitive (System 1) and slow and reflective (System 2) [20, 21]. Incomplete or incorrect mental models shaped by speculative claims about AI rather than scientific evidence can become a mental heuristic that triggers people's System 1 thinking rather than engaging their System 2 thinking, blurring the lines between speculation and informed reasoning [22]. This, in turn, can lead to miscalibration of trust and over-reliance on AI, which drives ineffective use of AI [18, 19, 23].

For instance, consider an AI system that claims to give you personalized product recommendations based on your social media usage and purchase history. A user with an incorrect mental model of the system's capabilities and limitations may hastily rely on its recommendations without considering other options. On the flip side, a user who has developed an overly skeptical mental model of AI may reject helpful recommendations and miss critical opportunities. Therefore, without ongoing effort to establish appropriate mental models, users can become more prone to manipulation and potential negative outcomes.

While belief in AI supremacy is growing despite its demonstrated limitations and biases, our study presents the first empirical investigation into the phenomenon of "rational superstition" [1] in AI[1]. To do so, we examine how people's perception and trust in AI predictions may correlate with other irrational behaviors, such as belief in astrology. This research question is particularly timely

---

[1]This chapter is adapted from a preprint on arXiv:2408.06602, co-authored with Pat Pataranutaporn, Judith Amores, and Pattie Maes, 2024 [24].

and relevant given the recent surge in popularity of astrology, especially among Millennials and Gen Z [25, 26], despite the lack of empirical evidence supporting its validity [27, 28]. A recent consumer survey of over 2000 people in the US revealed that about a third of Millennials and a quarter of Gen Z have made financial decisions based on their horoscope [29].

To address this research question, we conducted an experiment involving 238 participants who were presented with fictitious predictions from three sources (AI, astrology, and personality psychology) and rated their perceived validity, reliability, usefulness, and personalization. To create a reasonable basis for the predictions, participants answered questionnaires about astrology and personality and engaged in a simulated investment game that appeared to analyze their interactions and predict their future investment behavior. Participants were randomly assigned to either positive (N=119) or negative (N=119) prediction groups, receiving forecasts about their future investment behavior and outcomes accordingly.

Our study also explores how cognitive style, paranormal beliefs, gullibility, trust in AI, and personality traits may influence belief in predictions based on AI, astrology, and personality psychology. This investigation builds upon previous work examining how user characteristics such as personality traits influence the perception of AI-generated advice [30] and trust in AI [31]. Additionally, we consider the relationship between cognitive style and paranormal beliefs, including astrology [32–36], as well as how cognitive reflection affects social media usage and the sharing of (mis)information and conspiracy theories [37, 38].

Furthermore, we explore how demographic factors such as age, gender, and education level, as well as prior experience in AI, astrology, and personality psychology, and the level of interest in the topic influence belief in the predictions. By examining these various factors, our research aims to provide a comprehensive understanding of the "rational superstition" phenomenon in AI and its implications for decision-making in an increasingly AI-driven world.

Based on our research questions, we tested the following hypotheses:

- H1: Individuals who are more likely to believe in predictions based on astrology and personality will be more likely to believe in AI predictions.

- H2: Individuals with higher cognitive reflection and need for cognition will find AI predictions more credible and reliable than astrology- or personality-based predictions.

- H3: Demographic factors and personal attributes (cognitive reflection, trust in AI, paranormal beliefs, gullibility, personality, level of interest, and prior experience) will moderate the perceived validity, reliability, usefulness, and personalization of predictions from different sources.

The study's findings contribute valuable insights to the ongoing discourse on AI ethics, transparency, and responsible implementation. By shedding light on the factors that influence trust and

belief in AI predictions, we aim to inform policymakers, AI developers, and educators about potential pitfalls and areas for improvement in AI literacy and public engagement with AI technologies.

2.2   Methods

2.2.1   Participants

Data were initially collected from a survey administered to 300 participants recruited through Prolific. After excluding 62 participants due to incomplete responses or failed attention checks, the final analysis was conducted with data from the remaining 238 participants. These individuals were adults aged 18 and above from diverse socioeconomic backgrounds (see Table 2.1 for an overview of participants' demographic information).

Table 2.1: Summary of participants' demographic information

Number per condition	Positive	119
	Negative	119
Age	Mean	41.04
	SD	12.38
Gender	Female	53%
	Male	43%
	Other	3%
Education	High school	12%
	Some college	21%
	Bachelor	43%
	Master	14%
	Other	9%

2.2.2   Experiment Protocol

Assessments and simulated investment game

A visualization of the study procedure is shown in Figure 2.1. First, participants were asked to complete a short zodiac-related questionnaire (date, hour, and location of birth), a short personality test (adaptation of Myers-Briggs Type Indicator (MBTI) [39], and a simulated investment game. In the instructions, participants were told that they will receive personalized predictions about their future investment behavior generated from three distinct sources (astrology, personality, and AI) based on their game interactions and the information they provided. What they were *not* told in the
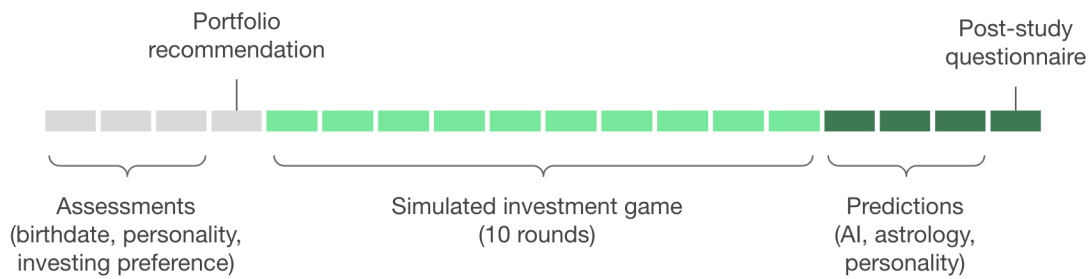
Figure 2.1: Diagram of the study procedure

beginning was that these predictions were generic, pre-determined statements that were not based on their responses, which was revealed to them after completion.

The simulated investment game was designed to elicit interactions comparable to that of modern robo-advising platforms [40], but in a simplified way. Participants were provided with virtual currency of $10,000 to invest across three investment categories (high risk/return, medium risk/return, low risk/return) over ten rounds (representing years). At each round, they could allocate up to 100% of their assets across the three categories, with the goal of maximizing their total portfolio value.

A few design choices were made to increase the sense of realism for the investing game and to increase engagement:

- **Recommended portfolio allocation:** Participants filled out a short questionnaire about their risk preferences for investing before starting the game and received a "customized" portfolio recommendation that they could view throughout the game. In reality, all participants received the same portfolio recommendation.

- **Realistic market scenario:** The game was designed so that participants could experience the hypothetical ups and downs of the market, through a pre-determined series of alternating "bull market" (more gains) and "bear market" (more losses) scenarios that influenced the probability of the possible investment outcomes at each round.

- **Market forecasts:** To further enhance the emotional engagement, each round started off with either a positive or negative market forecast that were aligned with the behind-the-scenes market scenario. A detailed market scenario that was used in the game and examples of the market forecast can be found in Appendix D.

- **Financial incentives:** To motivate engagement, participants were told their investment performance would increase their total compensation for the study, inspired by previous studies that used real-life incentives to increase engagement [41, 42].

15
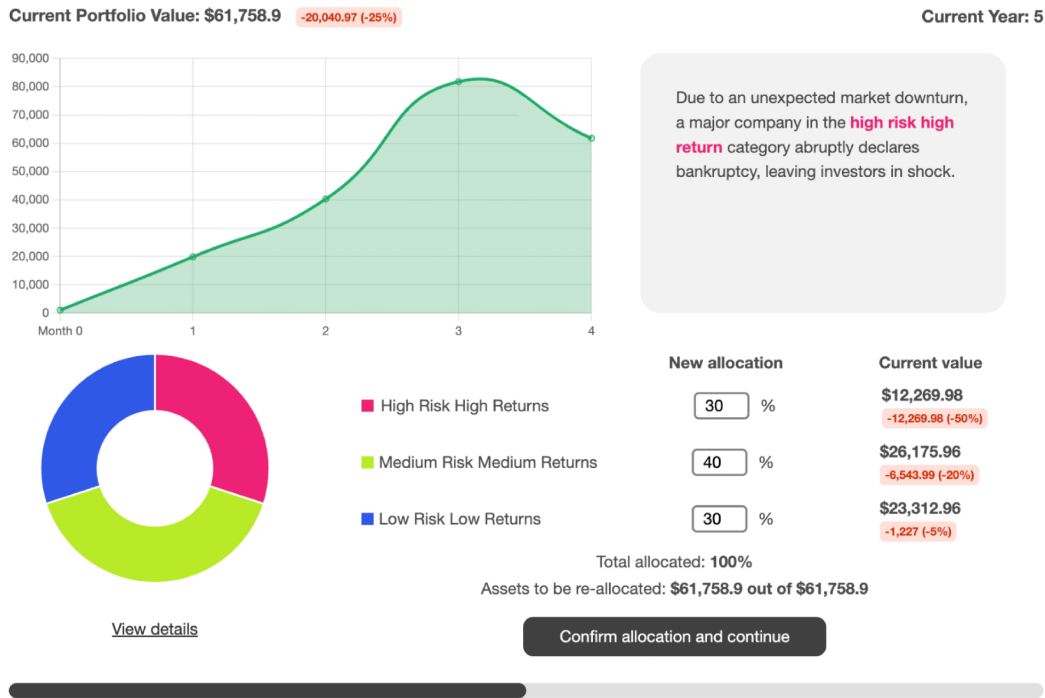
Figure 2.2: Dashboard of the simulated investment game

A snapshot of the simulated investment game can be seen in Figure 2.2. The game is publicly accessible at https://prophecy-study-1.replit.app and the source code is available on GitHub (https://github.com/mitmedialab/ai-prophecy).

Predictions

We employed a placebo approach that was proven effective in previous studies [43, 44] where participants were informed that predictions were AI-generated, but were in fact pre-determined (either positive or negative), and randomly assigned. After the game, participants were presented with these pre-determined predictions from three different sources (AI, astrology, personality) in randomized order. Participants were randomly assigned to either the "Positive" prediction group (rational investor, higher returns) or "Negative" prediction group (impulsive investor, lower returns). The initial predictions were created by the authors, with variations generated using ChatGPT-4. Examples of predictions received are shown in Table 2.2. For each prediction, they were asked to respond to evaluate its perceived validity, personalization, reliability, and usefulness. The evaluation questionnaire is provided in Section 2.2.3.

Table 2.2: Example of predictions received by Positive and Negative prediction groups

Source	Positive prediction group	Negative prediction group
Astrology	As the Moon's celestial path brings it into close proximity with the stoic Saturn, this cosmic alignment sheds light on the inherent human struggle between emotion and reason when making financial decisions. Based on your astrological sign, you are more likely to make rational investment decisions based on reason rather than emotion. You have a good grasp of your emotions especially when it comes to decisions about money. Therefore, we predict that your portfolio will have higher-than-average performance in the long run.	As the Moon's celestial path brings it into close proximity with the stoic Saturn, this cosmic alignment sheds light on the inherent human struggle between emotion and reason when making financial decisions. Based on your astrological sign, your investment decisions are often impulsive, swayed by emotions rather than thoughtful analysis. Your emotions play a significant role in your investment decisions. As a result, it's expected that your portfolio's performance will be under the average in the long run.
Personality	Whether you're an analytical thinker or led by intuition, your Myers-Briggs profile offers valuable insights into your behavior when it comes to making investment decisions. Based on your personality analysis, you are more likely to make rational investment decisions based on reason rather than emotion. You have a good grasp of your emotions especially when it comes to decisions about money. Therefore, we predict that your portfolio will have higher-than-average performance in the long run.	Whether you're an analytical thinker or led by intuition, your Myers-Briggs profile offers valuable insights into your behavior when it comes to making investment decisions. Based on your personality analysis, your investment decisions are often impulsive, swayed by emotions rather than thoughtful analysis. Your emotions play a significant role in your investment decisions. As a result, it's expected that your portfolio's performance will be under the average in the long run.
AI	Within the digital echo of your actions, AI sifts through the noise to forecast your next move. It transforms your clicks, likes, and digital whispers into a map of future investment choices. Based on our AI model analysis, you are more likely to make rational investment decisions based on reason rather than emotion. You have a good grasp of your emotions especially when it comes to decisions about money. Therefore, we predict that your portfolio will have higher-than-average performance in the long run.	Within the digital echo of your actions, AI sifts through the noise to forecast your next move. It transforms your clicks, likes, and digital whispers into a map of future investment choices. Based on our AI model analysis, your investment decisions are often impulsive, swayed by emotions rather than thoughtful analysis. Your emotions play a significant role in your investment decisions. As a result, it's expected that your portfolio's performance will be under the average in the long run.

Table 2.3: Summary of the measures for the user characteristics

Variable name	Measure
Cognitive Style	A composite measure of cognitive style that combines cognitive reflection test (CRT-2) [45] and need for cognition (NCS-6) [46]. Higher score indicates more analytic cognitive style.
Paranormal Beliefs	A measure of participants' belief in paranormal phenomena. A shortened version of R-PBS [47] was used. Higher score indicates stronger belief in paranormal phenomena.
AI Attitude Score	A score indicating participants' attitudes towards AI predictions. AIAS-4 [48], a 4-item scale was used. Higher score indicates more positive attitudes.
Gullibility	A self-reported measure of participants' tendency to be gullible, developed by [49]. Higher score indicates greater gullibility.
Big Five Personality Traits	Includes extraversion, openness, agreeableness, conscientiousness, and emotional stability. Used 10-item Big Five Personality Inventory (TIPI) [50].
Interest in behavior	A measure of participants' interest in the topic of prediction (personal investment behavior). "How would you rate your level of interest in understanding your future investment behavior?" (5-point Likert scale, 1=Not interested, 5=Extremely interested)
Familiarity	A measure of participants' familiarity with the prediction sources (AI, Astrology, Personality). "How would you rate your level of familiarity with the following: 1) Astrology and horoscopes, 2) Personality psychology, 3) AI prediction systems" (5-point Likert scale, 1=Not at all, 5=Extremely familiar)

Post-study questionnaire

Following the experiment, participants were asked to respond to a series of questionnaires to measure key user characteristics (cognitive style, paranormal beliefs, gullibility, trust in AI/attitudes towards AI, personality, etc.) and demographic information (age, gender, education level). A summary of the scales used in the questionnaires is provided in Table 2.3. Detailed questionnaires can be found in Section 2.2.3.

2.2.3 Measures

This section includes an overview of the questionnaires that were used to measure the variables. For detailed questionnaires, see Appendix A.

Perceived validity, personalization, reliability, and usefulness

To measure the believability of the predictions, we presented a short questionnaire that assess perceived validity, reliability, usefulness, and personalization of given predictions. These factors were chosen based on existing literature on the different qualities of a statement that contribute to persuasion, user adoption of an information technology, and positive attitudes.

Perceived validity/accuracy of the content and reliability of the source was added according to the Elaboration Likelihood Model (ELM) in sociology, which argues that the logical validity of an argument is considered central to its believability, but people can also engage in peripheral ways of being persuaded, such as the reliability of the source [51]. Moreover, perceived usefulness was included as a metric that contributes to believability of a prediction, inspired by the Technology Acceptance Model (TAM)[52]. According to this model, perceived usefulness, defined as the degree to which a person believes that using a particular system would enhance their job performance, significantly influences users' attitudes and intentions, which in turn affect actual technology usage. Lastly, perceived personalization has been found to increase attention and positive attitudes toward a statement [53].

We asked participants to rate eight statements on 7-point Likert scales (1=Strongly disagree, 7=Strongly agree), which include two statements per subscale (perceived validity, personalization, reliability, and usefulness).

Cognitive Reflection (CRT-2) and Need for Cognition (NCS-6)

Cognitive reflection, or the tendency to suppress an intuitive but incorrect response in favor of a more reflective and correct one, was chosen as a moderating factor for belief in predictions to explore how it influences the acceptance and trust in predictions about personal behavior. This approach aims to determine if higher cognitive abilities correlate with more critical evaluation and discernment of AI predictions, as well as predictions based on astrology and personality.

The Cognitive Reflection Test (CRT) was designed to predict performance in normative decision-making [54]. In our study, we used CRT-2 [45], an alternate form of the original 6-item CRT, in order to address issues of familiarity and learning effects that could potentially skew results. The 4-item CRT-2 maintains the same objective of measuring cognitive reflection but includes different questions to ensure a fresh and unbiased assessment, with less focus on numeric abilities [45].

The Need for Cognition (NFC) test is a self-report measure of an individual's tendency to engage in and enjoy effortful thinking [55]. We use the 6-item NFC assessment (NCS-6) developed by [46].

While the CRT measures objective performance, the NFC measures subjective preference, which have been found to be correlated [56]. To account for the conceptual similarity of these two measures, they were combined to create a composite measure under the assumption that they

have equal contribution to the final score. To do so, we summed the z-scores of the results from the two tests to account for the differences in range and variability [57].

Paranormal beliefs (R-PBS)

Paranormal beliefs refer to the conviction in phenomena beyond scientific explanation, such as supernatural events, astrology, and mystical experiences. Paranormal beliefs were chosen as a moderating factor for belief in AI, astrology, and personality predictions to investigate how these unconventional beliefs influence the acceptance of various predictive statements. To quantify the degree of paranormal belief, we adapted the Revised Paranormal Belief Scale (R-PBS) [47] (7-point Likert scale, 1=Strongly disagree, 7=Strongly agree) to focus on the more relevant factors for our study, including traditional religious beliefs (4), spiritualism (4), precognition (4), superstition (3).

AI Attitude/Trust in AI (AIAS-4)

A 4-item, 1-factor AI attitude scale (AIAS-4) [48] (10-point Likert scale, 1=Not at all, 10=Completely agree) was used to measure participants' attitude towards AI.

Gullibility

A shortened version of the originally 12-item, 2-factor scale by Teunisse et al. (2020) [49] (7-point Likert scale, 1=Strongly disagree, 7=Strongly agree) was used to measure self-reported gullibility.

Big Five Personality

Based on the Five Factor personality model, the 10-item Big Five Personality Inventory (TIPI) [50] (7-point Likert scale, 1=Strongly disagree, 7=Strongly agree) was used to measure participants' personality.

Familiarity/Level of expertise

How would you rate your level of familiarity with the following (5-point Likert scale, 1=Not at all, 5=Extremely familiar):

- Astrology and horoscopes
- Personality psychology
- AI prediction systems

Interest in topic of prediction

How would you rate your level of interest in understanding your future investment behavior? (5-point Likert scale, 1=Not interested, 5=Extremely interested)

2.2.4 Approvals

This research was reviewed and approved by the MIT Committee on the Use of Humans as Experimental Subjects, protocol number E-5768. The study was pre-registered on AsPredicted.org (see https://aspredicted.org/QDF_8KX).

2.2.5 Analysis

Variables

The dependent variable (Y) was the subscale scores for perceived validity, personalization, reliability, and usefulness (7-point Likert scale, 1=Strongly disagree, 7=Strongly agree).

The analysis included a range of predictor variables (X) to examine their effects on the subscale scores. The predictor and control variables were:

- *prophecy_source*: A categorical variable indicating the source of the prophecy (Astrology, Personality).

- *prophecy_group*: A categorical variable indicating the valence of prophecy (Positive, Negative).

- *composite_score*: A composite measure of cognitive style.

- *paranormal_score*: A measure of participants' belief in paranormal phenomena.

- *aias_score*: A score indicating participants' attitudes/trust towards AI predictions.

- *gullibility_score*: A self-reported measure of participants' tendency to be gullible.

- Big Five Personality Traits: A measure including extraversion, openness, agreeableness, conscientiousness, and emotional stability.

- *interest_behavior*: A measure of participants' interest in the topic of prediction (personal investment behavior).

- *familiarity*: A measure of participants' familiarity with the prediction sources (AI, Astrology, Personality).

- *age*

- *gender*: A categorical variable (Female, Male, Other)

- *education*: A categorical variable (Bachelor, Master, Doctorate, Professional degree, Associate, Some college, High school, Less than high school)

Multiple linear regression

To test the hypothesis that individuals who are more likely to believe in astrology and personality-based predictions are also more likely to believe in AI predictions (H1), a multiple linear regression model was first applied to the data in wide format. The dependent variable was *ai_overall_score*, and the main predictors were *zodiac_overall_score* and *personality_overall_score*, with the rest of the variables mentioned in Section 2.2.5 included as control variables. The model was fitted using the lm function in R. The full results and diagnostics are in Appendix B.

$$ai\_overall\_score \sim zodiac\_overall\_score + personality\_overall\_score + control\ variables \quad (2.1)$$

The following model diagnostics were performed to ensure the validity of the model:

- **Linearity**: Residual plots for the linear regression model indicated no clear patterns, suggesting linearity.

- **Homoscedasticity**: The residual vs. fitted values plot for the linear regression model showed some heteroscedasticity, which was addressed in the mixed-effects model.

- **Normality of Residuals**: The Q-Q plot indicated that the residuals were approximately normally distributed.

- **Multicollinearity**: Variance Inflation Factors (VIFs) were calculated for the linear regression model, indicating that multicollinearity was not a concern.

- **Independence of Residuals**: The Durbin-Watson test for the linear regression model showed no significant autocorrelation.

Mixed effects model

Mixed-effects models are well-suited for hierarchical data, incorporating both fixed and random effects to test multiple hypotheses simultaneously. They use partial pooling, which shifts group estimates toward the overall mean, effectively reducing Type I errors without the severe power loss of traditional multiple comparisons corrections. This approach accounts for data dependencies and correlations, minimizing the need for explicit adjustments for multiple comparisons [58].

To examine the main effects and moderating effects of psychological and cognitive factors (H2, H3), the data was transformed to a long format, creating a three-level hierarchical structure. Each subject engaged in three types of conditions (*prophecy_source*) and evaluated across four subscales (*subscale*). The outcome variable was the *subscale_score*.

The mixed effects model was specified as follows. The fixed effects were defined as:

$$
\begin{aligned}
\text{subscale\_score} \sim \text{subscale} * (& \text{prophecy\_source} * \text{prophecy\_group} \\
& + \text{prophecy\_source} * \text{composite\_score} + \text{prophecy\_source} * \text{paranormal\_score} \\
& + \text{prophecy\_source} * \text{aias\_score} + \text{prophecy\_source} * \text{gullibility\_score} \\
& + \text{big5\_extraversion} + \text{big5\_openness} + \text{big5\_agreeableness} + \text{big5\_conscientiousness} \\
& + \text{big5\_emotional\_stability} + \text{interest\_behavior} + \text{familiarity} \\
& + \text{prophecy\_source} * \text{Age} + \text{education} + \text{prophecy\_source} * \text{gender})
\end{aligned}
$$

(2.2)

For the random effects, random intercepts were included to account for variability in baseline levels between subjects (*qualtrics_code*), and random slopes were included to account for variability in how subjects respond to different conditions (*prophecy_source*):

$$
\text{random} = \ \sim \text{prophecy\_source} \mid \text{qualtrics\_code}
$$

(2.3)

Additionally, the model incorporated a correlation structure and variance weights to further account for the hierarchical structure and heteroscedasticity in the data:

$$
\begin{aligned}
& \text{correlation} = \text{corSymm(form} = \ 1 \mid \text{qualtrics\_code/prophecy\_source),} \\
& \text{weights} = \text{varIdent(form} = \ 1 \mid \text{subscale * prophecy\_source * prophecy\_group)}
\end{aligned}
$$

(2.4)

We centered the continuous predictor variables to reduce multicollinearity and improve the stability and interpretability of the coefficient estimates. Centering involved subtracting the mean of the variable from each individual value. This transforms the variable to have a mean of zero while preserving its variance and distribution. After centering the predictors, we observed that four previously non-significant main and interaction terms became significant, including the main effect of Personalization subscale, the main effect Astrology prediction source, and interaction terms between subscale (Personalization, Reliability) and Astrology prophecy source. This change is attributed to the reduction in multicollinearity, resulting in more precise coefficient estimates.

The model was implemented using the lme function from the nlme package in R. The analysis controlled for missing values through listwise deletion. Optimization settings were adjusted to

ensure convergence, with maximum iterations and evaluations set to 1000.

Several model diagnostics were performed to ensure the validity of the model:

- **Linearity**: A residuals vs fitted values plot using standardized residuals indicated no clear patterns, suggesting linearity.

- **Homoscedasticity**: A residuals vs fitted values plot showed a random spread around the horizontal line at zero, confirming that heteroscedasticity was successfully addressed using the variance function (varIdent). One limitation to this diagnostics was the less continuous nature of the outcome variable (*subscale_score*), which was multilevel with 13 levels (value of 1 to 7 with an increment of 0.5), which caused the residual plot to exhibit a slight diagonal pattern.

- **Normality of Residuals**: Normality of residuals was assessed using a Q-Q plot, using standardized residuals to account for both fixed and random effects. While there were slight deviations at the extremes, they were not severe and thus considered as normal deviations from real-world data.

- **Multicollinearity**: To address potential structural multicollinearity, we centered the continuous predictor variables. This process reduced the Variance Inflation Factor (VIF) values for all fixed effects to below 5, indicating that multicollinearity was reduced to an acceptable level. The model diagnostics, including standard errors and confidence intervals, confirmed the stability and reliability of the coefficient estimates.

- **Independence of Residuals**: We applied the Durbin-Watson test at the subscale level to check for autocorrelation, finding significant positive autocorrelation in each group. Initially, we used an autoregressive correlation structure (corAR1), which resolved the issue for the "Validity" subscale and slightly reduced autocorrelation in others. However, model fit statistics suggested that an unstructured symmetric correlation structure (corSymm) was a better fit, though it did not fully address the positive autocorrelation. Comparison of the p-values, standard errors, and estimates between models with corAR1 and corSymm structures revealed no notable differences. We also attempted to model the order effect of the prophecies and subscales to mitigate this autocorrelation, but this approach did not lead to notable improvements. This unresolved autocorrelation may limit the validity of our results, and future studies should explore alternative approaches or additional data collection to address this issue.

- **Normality of Random Effects**: Q-Q plots of random effects were used to assess normality. The points closely followed the reference line, indicating overall normality, with slight devia-

Table 2.4: Random Effects Variance, Standard Deviation, and Correlations

Random Effect	Variance	SD	Corr w Intercept	Corr w Astrology
Intercept	1.560	1.249	-	-
prophecy_sourceAstrology	1.211	1.100	-0.260	-
prophecy_sourcePersonality	1.002	1.001	-0.221	0.293
Residual	0.121	0.348	-	-

tions at the tails suggesting minor departures. However, these deviations were not significant enough to challenge the model assumptions.

- **Random Effects Evaluation:** The variance components and Intraclass Correlation Coefficient (ICC) were analyzed to evaluate the random effects structure. The ICC values (Adjusted ICC: 0.939, Unadjusted ICC: 0.598) indicated that a substantial portion of the variance was due to random effects. To further assess their contribution to overall model variability, we examined the variance components in the mixed effects model, which included random intercepts and slopes for *prophecy_source* at the *qualtrics_code* level. The estimated variance components are presented in Table 2.4.

The results of the model diagnostics and the full results table of the mixed effects model are provided in Appendix B. The code for analysis is in Appendix E, and the full code for implementation is available on GitHub (https://github.com/mitmedialab/ai-prophecy).

2.3 Results

People who are more likely to believe in astrology and personality-based predictions are more likely to believe in AI predictions.

To test the first hypothesis (H1), a multiple linear regression model was applied to the data in wide format, with believability score for AI predictions as the outcome variable and believability scores for astrology- and personality-based predictions as predictor variables, along with control variables (cognitive style, paranormal beliefs, gullibility, AI attitude/trust in AI, big five personality, familiarity with sources, interest in topic, age, gender, education level). The scores across the four subscales (perceived validity, reliability, usefulness, and personalization) were averaged as overall "believability" scores to provide a straightforward test of the hypothesis.

The multiple linear regression analysis explained a significant proportion of the variance in the AI overall score ($R^2 = 0.7606$, Adjusted $R^2 = 0.7337$, $F(24, 213) = 28.2$, $p < 0.001$). Significant predictors included the zodiac overall score (Estimate = 0.3119, $p < 0.001$) and the personality overall score (Estimate = 0.4585, $p < 0.001$), supporting the hypothesis that belief in astrology and personality-based predictions is positively associated with belief in AI predictions. This association can be visually observed in the scatterplots in Figure 2.3.
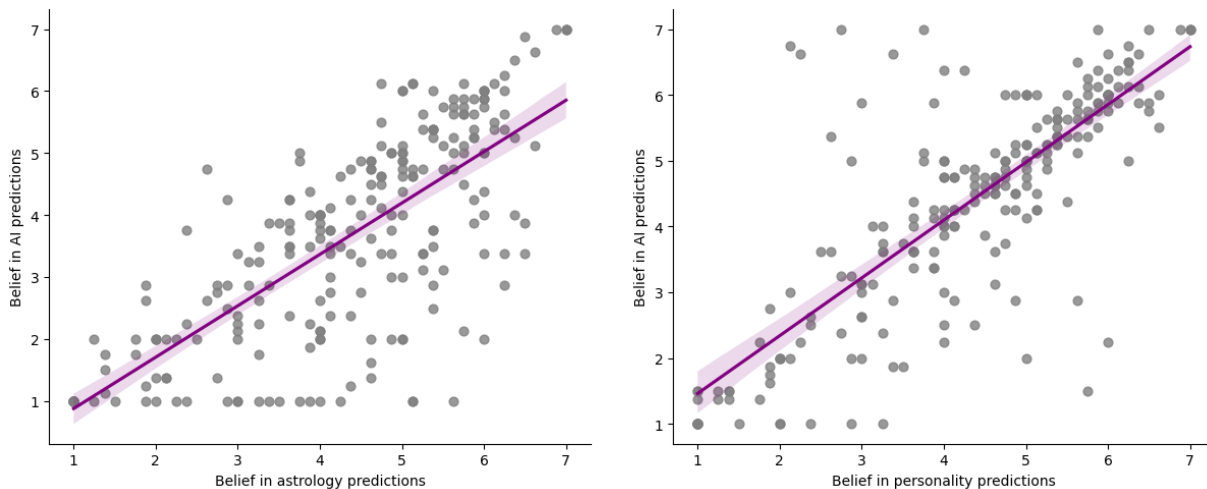
Figure 2.3: Scatterplots showing the relationship between belief in AI predictions and astrology predictions (left) and belief in AI predictions and personality predictions (right)

The scatterplots show individual scores, with a linear regression line indicating a positive correlation. The shaded area represents the 95% confidence interval for the regression line.

People generally find fictitious AI predictions about their personal behavior convincing.

To examine the influence of moderating factors including cognitive style (H2), paranormal beliefs, trust in AI, personality traits, demographic factors, etc. (H3), a mixed-effects model was fitted to the data in a nested long format, with prediction source (AI, astrology, personality) and believability subscale (perceived validity, reliability, usefulness, and personalization) as categorical predictors, and subscale score as the outcome variable. This approach allowed examination of the main effects and moderating effects of psychological and cognitive factors on belief in AI predictions.

The model included fixed effects for subscale, prophecy source, prophecy group, cognitive style, paranormal beliefs, gullibility, AI attitude/trust in AI, big five personality, familiarity with sources, interest in topic, age, gender, education level. Interaction terms were included to assess the combined effect of prophecy source and select moderating factors on subscale scores. The model also included random effects to account for individual variability, with the random intercepts capturing the baseline differences among subjects and the random slopes capturing the variability within each

participant across different conditions. The inclusion of random intercepts and slopes improved the model fit. For more on our methodology, see Section **??**.

Based on the mixed effects model analysis, the baseline subscale score, associated with "AI" prophecy source, "Validity" subscale, "Positive" prophecy group, "Female" gender, and "Bachelor" education level, was 5.10 (p < 0.001) on a 7-point Likert scale. The main effects of prophecy source on believability of AI predictions showed that on average, perceived validity of astrology-based predictions was 0.66 points lower compared to AI predictions (95% CI [-0.93, -0.39], p < 0.001). The difference was not statistically significant for personality-based predictions (0.07, 95% CI [-0.18, 0.32], p = 0.593). These results suggest that people generally found fictitious predictions from AI, astrology, and personality convincing.

There were significant main effects of subscales on the perception of AI predictions. On average, perceived personalization was rated 0.23 point higher than perceived validity (95% CI [0.08, 0.38], p = 0.002). On the other hand, perceived reliability was rated 0.91 points lower than perceived validity (95% CI [-1.13, -0.68], p < 0.001), and perceived usefulness was rated 0.63 points lower than perceived validity (95% CI [-0.83, -0.44], p < 0.001). This relationship between the subscales by prediction source can be visualized in Figure 2.4. The interaction between subscales and prophecy source showed that the main effects of subscales were largely consistent across AI, astrology, and personality, with some subscales being more pronounced for astrology. Perceived personalization was further increased by 0.23 points (95% CI [0.02, 0.44], p = 0.035) and perceived reliability further decreased by 0.31 (95% CI [-0.62, -0.00], p = 0.048) for astrology-based predictions, compared to the AI baseline. Other interaction terms were not significant, suggesting that the effects are not significantly different from the main effects.

Furthermore, we found that compared to the baseline "Positive" prediction group, the "Negative" prediction group was associated with a decrease in perceived validity by 1.19 points (95% CI [-1.53, -0.85], p < 0.001). We found no significant interactions between prophecy group and prophecy source, suggesting that this effect was consistent across AI, astrology, and personality. This suggests that while people generally found fictitious predictions believable, it was not the case for those who received negative predictions.

There is no evidence of correlation between belief in predictions and cognitive style.

Based on our hypothesis (H2), we expected to see either a significant positive association between cognitive style and belief in AI predictions, significant negative interactions between cognitive style and other prediction sources (astrology, personality), or both. However, our results showed that the composite cognitive score, as measured as the composite of the performance-based Cognitive Reflection Test (CRT-2) [45] and the preference-based Need for Cognition (NFC-6) [46], did not significantly increase the perceived validity of AI predictions. The estimated increase in perceived
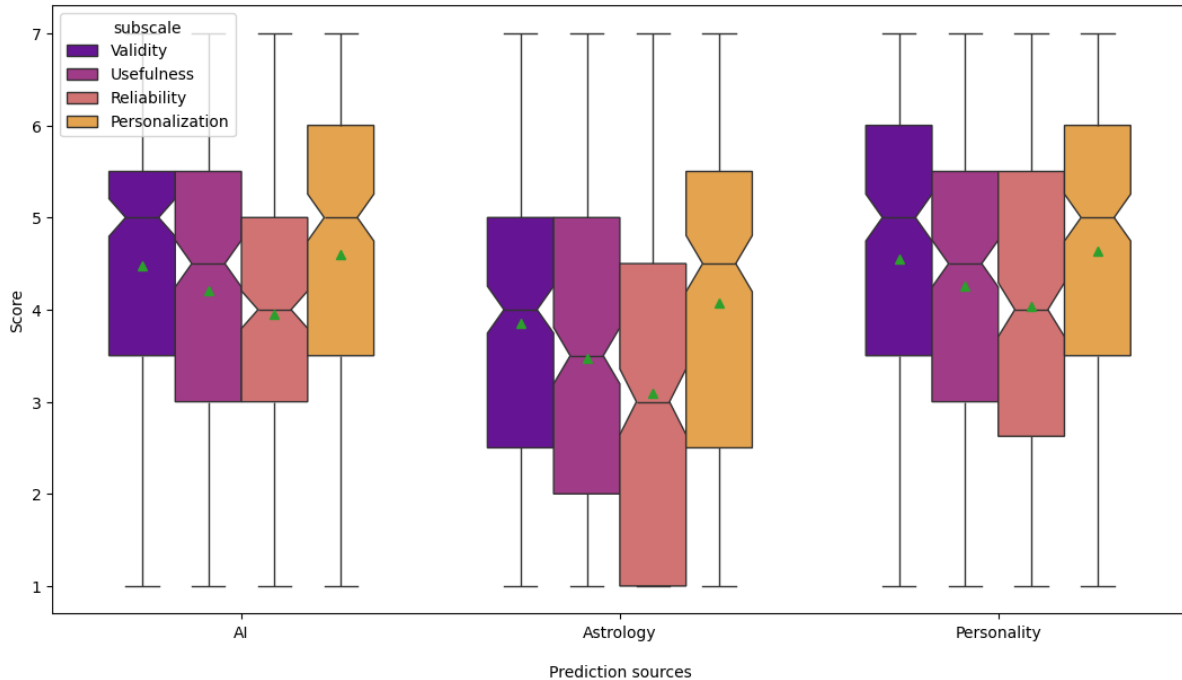
Figure 2.4: Boxplot showing the distribution of subscale scores by prophecy source and subscale

The boxes represent the interquartile range, with the means indicated in green triangles and notches representing the 95% confidence intervals for the medians.

validity was 0.13 points with one point increase in the composite cognitive score, but this effect was not statistically significant (95% CI [-0.01, 0.26], p = 0.065). The composite cognitive score ranged from -4.29 to 2.61 (Mean = 0.0, SD = 1.44).

However, we found some significant negative interactions between cognitive score and the subscales. Compared to perceived validity, higher composite cognitive score was associated with a decrease in perceived reliability (-0.11, 95% CI [-0.20, -0.02], p = 0.021) and perceived usefulness (-0.12, 95% CI [-0.20, -0.04], p = 0.004), making the main effect less positive. The interaction with perceived personalization was not statistically significant (-0.05, 95% CI [-0.11, 0.01], p = 0.092).

The interactions between cognitive style and prediction sources showed that while there was some negative interaction for astrology (-0.10, 95% CI [-0.21, 0.01], p = 0.079) and personality (-0.07, 95% CI [-0.18, 0.03], p = 0.170), the effects were not statistically significant. This suggests a lack of evidence that cognitive style is an influential factor in how people perceive predictions based on AI, astrology, and personality. Three-way interactions between subscales, prophecy source, and composite cognitive score did not lead to any significant results.

Higher paranormal beliefs increases perceived validity, reliability, usefulness, and personalization of AI predictions.

One of the most interesting findings was that having paranormal beliefs was positively associated with belief in AI predictions. The results from the mixed effects model showed that the paranormal beliefs score, as measured by a shortened version of the Revised Paranormal Belief Scale (R-PBS) [47], significantly increased perceived validity of AI predictions. With each point increase in the paranormal beliefs scale, perceived validity of AI predictions increased by an average of 0.02 points (95% CI [0.01, 0.03], p = 0.001) on a 7-point scale. The paranormal score, which was centered for the analysis, originally ranged from 15 to 95 (Mean = 45.6, SD = 20.08).

The interaction between paranormal beliefs and subscales showed that the effect was stronger for perceived reliability (0.01, 95% CI [0.00, 0.01], p = 0.020) and perceived usefulness (0.01, 95% CI [0.01, 0.02], p < 0.001), both being statistically significant. There was no significant interaction effect for perceived personalization (0.00, 95% CI [-0.00, 0.01], p = 0.427), suggesting that the effect did not differ significantly from perceived validity.

Paranormal beliefs score was an even stronger predictor for belief in astrology-based predictions, perhaps less surprisingly given that the scale includes questions around astrology (see Section 2.2.3). Each standard deviation increase in paranormal beliefs led to an additional 0.01 point increase in perceived validity (95% CI [0.01, 0.02], p = 0.001) compared to the AI baseline. Differences for personality-based predictions were not significant compared to the AI prediction baseline (-0.00, 95% CI [-0.01, 0.00], p = 0.274). No significant interactions were observed between the subscales, prophecy source, and paranormal score.

Positive AI attitudes increases belief in AI predictions, especially perceived reliability.

Individuals with more positive attitudes towards AI found predictions based on AI more believable. One point increase in AIAS score [48] led to an increase of perceived validity of AI predictions by 0.04 points (95% CI [0.01, 0.06], p = 0.001) on a 7-point scale. The AIAS score ranged from 4 to 40 (Mean = 25.79, SD = 8.68).

The main effect was augmented by an additional 0.03 points for perceived reliability (95% CI [0.02, 0.05], p < 0.001), while the interactions were not significant for perceived personalization (-0.00, 95% CI [-0.01, 0.01], p = 0.941) and usefulness (0.01, 95% CI [-0.00, 0.03], p = 0.071). This suggests that positive attitudes toward AI/trust in AI is more closely related to perceived reliability among all subscales, which is aligned with the findings of studies that explore the relationship between trust/attitude and reliance in AI [59–63].

The interaction between AI attitude (AIAS) and prediction source showed that the positive impact of positive AI attitudes on perceived validity was mostly reversed for astrology-based pre-

dictions (-0.24, 95% CI [-0.41, -0.08], p = 0.004), while the effect did not differ significantly for personality-based predictions (0.04, 95% CI [-0.11, 0.19], p = 0.590). The interaction effects between subscales, prophecy source, and AI attitude score were not statistically significant.

People with high conscientiousness are less likely to believe in predictions about personal behavior.

Out of the big five personality traits (Extraversion, Openness, Agreeableness, Conscientiousness, Emotional stability), conscientiousness was negatively associated with perceived validity of predictions across all sources. With each point increase in the conscientiousness score (Mean = 5.29, SD = 1.32, Range = [1.0, 7.0], perceived validity was estimated to decrease by an average of 0.15 points (95% CI [-0.30, -0.01], p = 0.032) on a 7-point scale. There were no significant interaction effects with subscales observed, suggesting the main effect was consistent across the four subscales.

While the other domains of the five-factor model were not found to have significant influence, there were some variation in interaction terms. Extraversion was associated with a 0.06 point increase in perceived usefulness (95% CI [0.02, 0.10], p = 0.004), while Openness was associated with a 0.06 decrease in perceived personalization (95% CI [-0.10, -0.01], p = 0.012).

Level of interest in the topic of prediction increases perceived validity, reliability, usefulness, and personalization.

While often overlooked in the context of studying trust in AI predictions, individuals' interest in the topic of prediction (in our case, personal investing behavior) was an influential factor in perceived validity, personalization, reliabiliity, and usefulness of predictions. With each point increase in the level of interest in the topic, perceived validity significantly increased on average by 0.27 points (95% CI [0.11, 0.43], p = 0.001) on a 7-point scale. The interest in behavior scale ranged from 1 to 5 (Mean = 3.18, SD = 1.09).

The positive association was observed across other subscales. Compared to perceived validity, perceived personalization increased slightly less, reduced by 0.06 points (95% CI [-0.11, -0.01], p = 0.018), while interaction was not statistically significant for perceived reliabilty (-0.07, 95% CI [-0.14, 0.01], p = 0.068) and perceived usefulness (0.02, 95% CI [-0.04, 0.09], p = 0.475), suggesting that the effects were comparable to perceived validity. The results suggest that when all else is held constant, the more people are interested in the topic, the more likely they will perceive fictitious predictions to be valid, reliable, useful, and personalized.

There is no evidence that the level of familiarity with the prediction sources influences belief in predictions.

On the other hand, familiarity with the prediction sources (AI, astrology, personality psychology), which measures the level of self-reported familiarity/prior knowledge in each source, did not have a significant effect on perceived validity (-0.02, 95% CI [-0.11, 0.07], p = 0.650) across all sources. The familiarity scale ranged from 1 to 5 (Mean = 2.96, SD = 1.05).

The interaction between familiarity and subscales shows that perceived personalization increases slightly but significantly by 0.05 points (95% CI [0.00, 0.11], p = 0.035), while differences were not statistically significant for perceived reliability (0.02, 95% CI [-0.06, 0.10], p = 0.685) and perceived usefulness (0.01, 95% CI [-0.06, 0.08], p = 0.781).

Unlike what the literature suggests about the familiarity effect, a cognitive phenomenon where people tend to prefer things they are familiar with [64], our results showed inconclusive evidence that familiarity in the prediction sources (AI, astrology, and personality) neither increased nor decreased belief in the respective predictions.

Other results

Gullibility was not found to be a significant predictor of perceived validity of AI predictions (-0.02, 95% CI [-0.04, 0.01], p = 0.128). While the interaction between gullibility and Personalization subscale was significant (0.01, 95% CI [0.00, 0.02], p = 0.021), the results were inconclusive when compared to the main effects. Interactions between gullibility and prophecy source were not significant (Astrology: -0.00, 95% CI [-0.02, 0.02], p = 0.977, Personality: -0.02, 95% CI [-0.04, 0.00], p = 0.096), suggesting the effects were similarly insignificant across different prophecy sources. Gullibility scale ranged from 6 to 39 (Mean = 14.68, SD = 7.69).

We found that older age was associated with a decrease in belief in predictions across all sources. One year increase in age (Mean = 41.04, SD = 12.38, Range = [19, 75]) was associated with a decrease of perceived validity of AI predictions by 0.02 points (95% CI [-0.03, -0.00], p = 0.014). Interactions between age and subscales were not significant, nor were interactions between age and prophecy source. The non-significant interaction terms suggested that the effect was consistent across subscale and prophecy source.

For gender, we did not find significant main effects relative to the Female reference level. While male participants were associated with lower perceived validity scores for AI predictions than female participants, it was not statistically significant (-0.31, 95% CI [-0.69, 0.06], p = 0.103). However, they tend to perceive them as more reliable than female participants (0.41, 95% CI [0.14, 0.68], p = 0.003). Interactions with other subscales were not significant, suggesting they were similar to the perceived validity baseline. Interactions between gender and prophecy source were not significant.

There were no significant main effects for level of education compared to the Bachelor reference level. Interactions between education level and subscale showed some significant effects between High school and Personalization (-0.23, 95% CI [-0.40, -0.06], p = 0.009), and Less than high school and Reliability (-0.89, 95% CI [-1.56, -0.22], p = 0.009), but these results did not lead to conclusive results. The detailed results table can be found in Appendix B.

2.4 Discussion

Belief is subjective and context-dependent; it is also socially constructed through external stimuli such as media narratives, contextual factors, and internal disposition [65, 66]. When looking at belief, trust, and reliance in AI predictions, it is important to consider not just the features of the AI system, but also the characteristics of the users and the broader context. The results from our study show that people who are more likely to believe in astrology- and personality-based predictions were more likely to believe in AI predictions. Moreover, we found that belief in AI predictions on personal behavior can be influenced by various psychological, cognitive, and contextual factors that also drive belief in astrology- or personality-based predictions. By breaking down the concept of believability into four subscales (perceived validity, reliability, usefulness, and personalization), we were able to observe interesting patterns and interactions with different subscales.

Cognitive style

While prior literature seemed to suggest that an analytic cognitive style would lead to more skepticism in predictions based on astrology and possibly personality [32, 34, 36], our results did not show significant evidence to support these claims. Moreover, cognitive style was not a significant predictor in belief in AI predictions. Thus, our hypothesis (H2) that people with more analytic cognitive style would be more likely to believe in AI predictions than predictions based on astrology and personality was not supported by our findings. This suggests that analytic cognitive style may not necessarily lead to rational skepticism of fictitious predictions, and perhaps there may be a missing link or mechanism between cognitive style and belief/trust in AI predictions, which could be a topic of future study.

Paranormal beliefs

Paranormal beliefs were found to be an influential predictor of belief in predictions across all sources, including AI predictions. While the association may seem unlikely at first, it can be interpreted within the context of the influence of popular AI narratives. The popular portrayal of AI as rational and objective creates the impression that people would also treat AI in a rational way, and thus more

scientifically-inclined people would find AI predictions more believable. However, our results show that belief in AI predictions is more closely associated with paranormal beliefs than one's cognitive style. This result points to the existence of the phenomenon of "rational superstition" in AI that was described in the introduction. Moreover, while the effect of paranormal beliefs on belief in AI predictions was significant across all subscales, the effect was larger for perceived usefulness and reliability. Social science research suggests that the rise in belief in astrology is an indicator of rising uncertainties and anxieties as a result of disintegration of community in the modern era [67]. As such, people turn to pseudo-scientific approaches such as astrology for insights to help them navigate their lives. Our findings suggest that people may perceive AI predictions in a way that is similar to the way they perceive astrology-based predictions, which could help explain the results.

AI Attitude/Trust in AI

Our results showed that people with positive attitudes towards AI found AI predictions more valid, reliable, useful, and personalized. Among the subscales, perceived reliability was found to be most closely related to attitude towards AI. This supports findings from prior literature that positive attitudes toward AI lead to higher reliance in AI predictions [63, 68].

Personality

Our findings using the Big Five personality traits [69] show that conscientiousness had a negative influence on perception of validity, personalization, reliability, and usefulness, while other traits did not have significant effects. This contrasts previous findings by Riedl [31], who found that agreeableness, openness, and extraversion positively influenced trust and high neuroticism negatively impacted trust, while results on conscientious were mixed.

Conscientiousness is defined as "socially prescribed impulse control that facilitates task- and goal-directed behavior, such as thinking before acting, delaying gratification, following norms and rules and planning, organizing, and prioritizing tasks" [70, p. 120]. Prior studies have found a positive relationship between conscientiousness and cognitive ability [71, 72]. Contrary to our expectations that these two traits may mirror each other in direction, it was not the case—cognitive style did not have significant effect on believability in our study, while conscientiousness had a significant negative influence on perception of validity, personalization, reliability, and usefulness.

Interest in the topic of prediction

Our findings show that people's interest in the topic of prediction is a strong predictor when it comes to their belief in predictions, regardless of the source. While our results do not confirm any causal relationship, it could suggest that the more interested one is in a topic, the more exposure they may

have to information about it, and the more likely they may be susceptible to cognitive biases that could influence their belief. Some biases that could explain this tendency are confirmation bias and belief bias. Confirmation bias is when individuals seek out and remember information that confirm their existing beliefs [73]. Belief bias occurs when individuals judge the strength of an argument based on the believability of its conclusion rather than the logical structure of the argument itself [74, 75].

This has implications in the context of AI-assisted decision-making in different fields; e.g. in marketing, companies may take advantage of people's interest in a certain topic and expose them to more targeted information (e.g. advertisements) that may create an impression of hyper-personalization and lead to more positive attitudes toward the brand [76]. Moreover, in the medical field, an individual may be more prone to believing AI-based predictions about their health if it confirms their prior beliefs.

Familiarity/Prior knowledge

Familiarity with the prediction source was not found to be a significant predictor of belief in predictions. Earlier, we discussed how people's positive AI attitudes and interest in the topic of the prediction were positively correlated with belief in the predictions, which may be explained by its likelihood to fuel certain cognitive biases such as belief bias. On the other hand, an individual's level of familiarity in AI, astrology, and personality psychology may not be necessarily indicative of their attitude towards the field. For instance, someone who is not acquainted with the details of latest AI developments may have either a utopian or dystopian view of AI's impact on society, as mentioned in the introduction. Similarly, just because an individual is very familiar with astrology does not inherently mean they are more likely to believe or adopt the predictions, as that may depend on other personal and contextual factors. This suggests two things: 1) simply knowing more about the prediction source and underlying technology does not predict whether one finds a prediction more or less believable, and 2) it does not make one better at calibrating one's trust, providing implications for the design of trustworthy AI systems.

Gullibility

While prior literature suggested a positive relationship between gullibility and paranormal beliefs [32, 35], our findings showed inconclusive results on whether self-reported gullibility had a positive or negative effect on the belief of fictitious predictions based on AI, astrology, and personality. There may be a few potential explanations for this. One is that the order of the study potentially biased participants' answers. The survey was completed after participants had seen and evaluated the predictions, at which point their perception of the believability of the predictions may shadow their

self-perception of gullibility—in other words, they did not want to contradict their decisions so quickly by admitting that they are more prone to being fooled. Another possible explanation is the limitation of a self-report nature of the gullibility scale. Some people may be hesitant to admit that they are gullible for different reasons. As such, future studies can further explore the relationship between gullibility and belief in AI predictions.

Demographic factors

Our results found that older people are more skeptical of predictions about personal behavior, across AI, astrology, and personality sources. This is aligned with prior literature that seems to suggest that older age is associated with lower trust in AI [77], due to many reasons including bias, lack of learning avenues, and concerns about privacy [78]. For gender, while the main effects were insignificant, interactions revealed that male participants were more likely to perceive AI predictions as reliable than female participants. This is supported by a prior study that found that being male with higher education and with a Western background were predictors of trust in AI among the general population [79]. Moreover, while prior work suggested a positive association between level of education and belief in AI [77], we did not find significant main effects for level of education, and the interaction terms did not lead to conclusive results.

2.4.1 General Discussion

Our results highlight the irrational side of how humans perceive and believe in AI predictions by comparing the perception of fictitious predictions based on AI, astrology, and personality. Our analysis showed that belief in AI prediction about personal behavior may be more related to paranormal beliefs than cognitive style, and that positive attitudes toward AI and interest in the prediction topic enhance this belief, regardless of familiarity with the sources.

These findings suggest that highly accurate and reliable performance is not a prerequisite for people to put high trust in predictions; models do not need to be completely trustworthy for users to trust them. This disconnect between perception and actual performance raises important discussions about how people perceive and engage with AI systems.

1) Incorrect mental models of AI can override perception of validity.

Belief in AI predictions is often rooted in mental models rather than actual performance. Hoff and Bashir's three-layer framework of trust in automation—dispositional, situational, and learned— illustrates that system performance contributes to learned trust through active engagement with the system [80]. Our study shows that even without validating system performance, people rated fictitious AI predictions as highly valid, reliable, useful, and personalized. This placebo effect of AI

has been observed by other researchers, where mere priming with AI descriptions alters perception and behavior [43, 81–83].

This underscores the significant role mental models play in shaping trust in AI [23, 84–88]. Mental models, comprising beliefs and expectations about AI, are influenced by prior beliefs and ongoing interactions [23, 84]. By framing the scenario in certain ways, including the task domain and the information provided about the AI system, we observed that people may perceive fictitious AI predictions to be accurate, personalized, and trustworthy [23].

Moreover, our results suggest that people might view AI predictions similarly to astrology, supporting the idea of "rational superstition" in AI and automation [1]. We argue that this phenomenon potentially stems from the idealized portrayal of AI as enchanted yet rational and deterministic [89], which can bias interactions and reinforce existing beliefs about AI.

The dual-process theory in psychology provides insight into how larger narratives can bias perceptions of AI. This theory describes two types of thinking: fast and intuitive (System 1) and slow and reflective (System 2) [20, 21]. Mental models, shaped by public narratives, may act as heuristics that trigger System 1 thinking, leading to speculation rather than informed reasoning [22]. Incomplete or incorrect mental models, shaped by speculative claims rather than evidence, can lead to overreliance or miscalibrated trust, affecting effective AI use [18, 19, 23, 60].

This highlights the need to demystify AI in popular narratives and promote awareness of both its capabilities and limitations to foster accurate mental models. This leads to the importance of AI literacy efforts and how narratives are shaped by providers, media, etc. As technology advances, users' mental models should evolve with a balanced level of skepticism to prevent overreliance or undue skepticism.

2) Mental models can influence and be influenced by cognitive biases.

Cognitive biases help explain the irrational belief in AI predictions. Belief bias, where individuals evaluate statements based on prior beliefs rather than logical validity [74, 75], is largely driven by System 1 thinking (especially under time pressure [90]) but can also involve System 2, reinforcing mental models aligned with existing beliefs [21, 91]. This selective processing model explains why people may fail to engage logical thinking when a statement aligns with their beliefs, especially concerning self-related beliefs [92].

The Barnum effect, where individuals believe vague, general statements to be highly accurate for them personally, also plays a role [93]. This effect, studied in contexts like astrology and personality tests [94, 95], can explain how people view generic AI-generated predictions as personalized and valid. Cognitive biases like belief bias and the Barnum effect shape mental models and influence AI reliance [85], while incorrect mental models can, in turn, induce these biases.

36

3) Design of AI systems must consider mental models and cognitive biases to support the development of an appropriate level of trust.

AI systems should be designed to help users form accurate mental models. In contexts like mental health chatbots, incorrect mental models can lead to unrealistic expectations and greater disappointment among users. Thus, these systems should include mechanisms that promote an appropriate level of trust to avoid over or under-reliance, which should be accompanied by efforts to increase AI literacy [63]. Fortunately, mental models are dynamic and can evolve with more interaction [23, 96].

Awareness of the potential impact of cognitive biases on human-AI interaction is crucial to reduce the risk of negatively exploiting these human tendencies. AI systems that merely validate users' pre-existing beliefs may be favored but may not support better decision-making, particularly in critical areas like personal finance or health. Thus, we could think about how to design systems that can present information in a way that helps people think critically.

There are also important considerations for designing explainable AI systems. While explanations are intended to build trust, they can inadvertently increase over-trust, especially if users have inaccurate mental models [97]. Studies have shown that belief bias can lead to overreliance on AI explanations when they align with prior beliefs [98]. To mitigate this, explanations should be designed to engage users in critical thinking, perhaps by questioning the user or presenting examples that clarify the AI model's reasoning [99, 100].

4) Our understanding of human-AI interaction needs to include psychological and contextual factors.

How people believe, trust, and rely on AI predictions for their decision-making should be seen in a larger context [63]. Contextual factors, including individual differences, task domain, information provided, media discourse, organizational and social environment, and more are all factors that could influence user perceptions and behavior in AI-assisted decision-making [12, 23].

Our findings emphasize that beyond system performance, user-related factors are critical in evaluating human-AI interaction. Psychological factors, including prior attitudes toward AI, paranormal beliefs, personality traits, and interest in the topic, significantly influence how people perceive AI predictions. These findings must be considered within the broader societal narratives, mental models, and cognitive biases that shape human-AI interactions.

2.5   Limitations and Future Work

While our study has presented valuable insights into the phenomenon of "irrational superstition" in AI predictions and cognitive biases, there are several important limitations that can be addressed in future work. First, our experiment was specifically designed to provide predictions on personal behavior. There are some limitations to the extent to which our results can be generalized to the general use of AI. Therefore, future studies could look at how people perceive AI predictions in different contexts and prediction topics, while also exploring the relationship with underlying psychological and cognitive factors at play. For instance, the effect can be tested within the context of AI-assisted decision-making in personal health or well-being, or in educational settings using an AI tutor. We believe that our study opens up more questions about understanding the intricate ways in which humans perceive, trust, and interact with AI systems. Second, like most survey designs, our survey-based approach to measuring psychological and cognitive traits may be affected by social desirability bias (the tendency to respond in a way that seems socially favorable) [101], as well as common method biases [102], and reliance on self-reported data. Therefore, we hope future studies can validate our findings using different methods and approaches. Lastly, while our study mainly focused on user-related factors, future studies could also look at other contextual factors such as social and organizational environments and their impact on user perception and trust.

2.6   Conclusion

Our study empirically investigated the phenomenon of "rational superstition" in AI by providing a side-by-side comparison of people's perceptions of fictitious predictions based on AI, astrology, and personality. Even without any validation of system performance, people perceived fictitious AI predictions as highly believable, and their perceptions were positively associated with their perception of astrology- and personality-based predictions. Our findings showed that people's belief in AI predictions was positively associated with paranormal beliefs, positive attitude towards AI, interest in the topic of prediction, and negatively associated with conscientiousness and age. We also explored interactions with other prediction sources and the subscales including perceived validity, reliability, usefulness, and personalization. In our discussion, we highlighted the role of mental models and cognitive biases in shaping people's perceptions of AI predictions, and the importance of including psychological and contextual factors in studying human-AI interaction.

Chapter 3

Personal Validation Effect in LLM Use: Positive AI Responses Bias Perceptions of Validity, Personalization, Reliability, and Usefulness of Fictitious Predictions

3.1 Introduction

Danish author Hans Christian Andersen published "The Emperor's New Clothes" in 1837. This cautionary tale highlights human vulnerability to suggestion and flattery. In the story, two swindlers persuade an emperor they can weave fabric invisible to the foolish. Through false praise, and always telling the emperor what he wants to hear, thereby gaining his trust, the swindlers convince the emperor to parade around naked, believing he's wearing magnificent attire. This psychological manipulation bears a striking resemblance to the phenomenon of "sycophancy" observed in Large Language Models (LLMs), where the use of human feedback to fine-tune AI assistants inadvertently encourages responses that align with users' preferences rather than truthful information, creating generative echo-chambers [103, 104].

The deliberate fine-tuning of AI models to consistently provide positive affirmation, while ostensibly aimed at building user trust and retention, may inadvertently serve as a subtle form of manipulation. This concern is underscored by studies revealing that people's evaluations of AI models are often irrational and subjective [43, 44, 82, 83]. Just as the emperor's subjects perceived nonexistent clothing because of suggestion, users may interpret AI-generated content as uniquely insightful, trustworthy, or personalized, even when it lacks substance.

In psychology, the "personal validation effect," also known as the Barnum or Forer effect, de-

scribes the tendency of individuals to accept vague and general personality descriptions as uniquely applicable to themselves [105]. Since its initial discovery in 1949, this phenomenon has been extensively studied and replicated, showing that people often rate supposedly personalized descriptions more favorably than general ones, even when the content is identical [106]. This effect has been explored in various contexts, including personality tests and astrology [94, 95, 107].

One of the working principles of the personal validation effect is the Pollyanna principle, or more generally known as the positivity bias, which refers to the tendency to focus on the positive and to recall positive memories more accurately than negative ones [108]. This effect is particularly concerning in the context of AI, especially in critical fields such as healthcare, where an AI assistant's praise of a physician's choices could unduly influence medical decision-making. In educational settings, AI systems offering unwarranted positive feedback to students may hinder genuine learning and critical self-assessment, potentially fostering a false sense of achievement reminiscent of "toxic positivity" [109].

As sophisticated AI systems become increasingly integrated into our daily lives, understanding the psychological dynamics at play in human-AI interactions is crucial. This paper investigates the phenomenon of personal validation in the context of LLM use, highlighting how positive AI responses can significantly influence users' perceptions of the validity, personalization, reliability, and usefulness of predictions—even when those predictions are demonstrably false. While the personal validation effect is a general phenomenon, its implications are especially dangerous in the context of AI, where users may unknowingly place undue trust in AI-generated content, leading to misguided decisions and beliefs.

The impetus for this research stemmed from our previous study [24] (see Chapter 2), which was inspired by the tech industry's quasi-religious narratives around AI development. This study explored how participants' trust in and responses to AI bear similarities to beliefs in astrology and pseudo-scientific personality predictions. The primary objective was to elucidate how people's reactions to AI predictions could be less rational and analytical than commonly assumed.

In the course of that research, a novel and intriguing research question was uncovered that extended beyond the initial investigation: *How does the valence (positive or negative) of AI-generated predictions influence perceptions of validity, personalization, reliability, and usefulness?* While this was initially pre-registered as a secondary hypothesis under our broader investigation, we realized that the implications and scope of this inquiry marked a significant shift from our prior focus on AI and superstition. It became clear that the distinctive nature of this inquiry, coupled with its potential for wide-ranging implications for real-world contexts, necessitated a separate, dedicated study.

As a result, this study aims to uncover the nuanced ways in which the valence, or positive or negative framing, of AI communications affects user perceptions and decision-making in real-world

scenarios[1]. Based on prior literature, we formulated the following hypothesis: Positive responses will significantly increase the perceived validity, personalization, reliability, and usefulness of fictitious predictions, compared to negative ones.

In addition to the pre-registered main hypothesis, we also explored the following research questions for exploratory analysis: To what extent do cognitive style, paranormal beliefs, gullibility, AI attitude/trust in AI, and demographic factors such as age and gender moderate the effect of the *valence of predictions* on people's belief in those predictions?

To explore this hypothesis and research questions, we built upon the approach from the previous study [24] to draw additional insights related to the Pollyanna principle or positivity bias in the context of human-LLM interaction and the psychological factors that influence belief in AI predictions. We employed an experimental design where participants are exposed to fictitious AI-generated predictions with either positive and negative valences to systematically examine these effects. This study contributes to the growing body of knowledge on human-AI interaction in three ways: 1) examining the extent to which the Barnum effect manifests in human-AI interactions, 2) quantifying the impact of positive framing on users' perceptions of AI-generated information, and 3) proposing strategies to mitigate the negative effects of the personal validation effect in AI design and implementation.

The results from our study (N=238) revealed that positive AI-generated predictions are perceived as significantly more valid (36% increase), personalized (42% increase), reliable (27% increase), and useful (22% increase) compared to negative predictions. These findings underscore the substantial impact of prediction valence on user perceptions, with implications for the design and deployment of AI systems across various applications.

The implications of this research reach areas of AI ethics, interaction design, and digital literacy. By understanding how personal validation and other cognitive biases operate in the context of human-LLM interaction, we can develop more responsible AI systems that balance the benefits of personalization with the need for accuracy and objectivity. Moreover, this work emphasizes the importance of educating users about the limitations and potential biases of AI systems, fostering a more critical and informed approach to human-AI interaction. Furthermore, as we continue to navigate the evolving landscape of AI technology, this work highlights that it is crucial to consider not only the technical capabilities of these systems but also the psychological factors that shape our interactions with them.

---

[1]This work is co-authored with Pat Pataranutaporn, Judith Amores, and Pattie Maes, 2024.

3.2 Related Works

3.2.1 Psychology of Human-AI Interaction

The evolution of human-AI interaction has been shaped by the complex interplay between human psychology and our interactions with technology, beginning with the foundational work in Human-Computer Interaction (HCI) and advancing through the emergence of Human-Centered AI (HCAI), which emphasizes balancing human control with automation to enhance capabilities and promote ethical, responsible outcomes [110]. The rapid development and widespread adoption of technologies, particularly Large Language Models (LLMs), have dramatically transformed human-AI interaction, presenting new opportunities and challenges in designing conversational AI interfaces [111, 112] and requiring more dynamic evaluation methods beyond static assessments [113]. In this context, studying the psychology of human-AI interaction is more important than ever, as it influences how people engage with AI systems, impacting user trust, decision-making, and the responsible use of AI systems in daily life. Understanding these psychological dynamics is essential for designing AI that is intuitive, ethical, and aligned with human needs.

A fundamental concept in this field is the Media Equation theory, proposed in 1996 [114, 115]. This research revealed that people often treat computers and other media as if they were real social actors, unconsciously assigning them human characteristics. This phenomenon highlights the human tendency to form social relationships and anthropomorphize non-human entities, a concept also known as the "Eliza Effect," named after the ELIZA chatbot created in the 1960s [116, 117]. Anthropomorphism significantly influences human-AI interaction by shaping how users perceive and engage with AI systems, as these human-like characteristics can lead to stronger emotional connections and altered expectations.

Researchers have identified various observable factors that influence human-AI interaction, including appearance (especially personified appearance) [118–124], voice characteristics like tone, pitch, and style [125–135], dialogue [136–138], movement and behavior [138–140], and emotional and social expression [138, 141–145]. Fine-tuning these factors allows AI system designers to enhance the impact of virtual agents or physical robots and strengthen their relationship with users. However, observable factors are only part of the equation. The internal processes of human psychology play a crucial role, as people construct "mental models" of the world and the entities they interact with [146–152]. These models shape perceptions and expectations, influenced by cultural context, individual beliefs, and specific use cases.

Recent studies have shown that altering a person's mental model of an AI system can affect their interaction with it. For example, one study found that participants primed to believe an AI had caring motives perceived it as more trustworthy, empathetic, and better-performing [83]. This

research highlighted a feedback loop where the AI and the user reinforce the user's mental model, emphasizing the importance of how AI systems are introduced.

The role of expectations in human-AI interaction has also been explored. One study demonstrated that heightened AI expectations can improve performance through placebo effects [82]. Participants with high expectations performed better when a sham AI system was present. Other studies have shown similar placebo effects in adaptive interfaces, where the belief in receiving adaptive AI support boosts expectations about task performance, positively correlating with actual performance and behavior [43, 44]. These findings suggest that system descriptions can elicit placebo effects, potentially biasing the outcomes of user studies and influencing the future assessment of AI-based interfaces and technologies.

In addition, previous work [103] has investigated sycophantic model behavior in LLMs that try to win the users' favor, exerting greater influence on user perception and behavior. A recent paper [153] has studied the persuasive influence of AI ingratiation on consumer behavior, and found that such behavior leads to increased acceptance of AI-based product recommendations. The authors propose that the effect is caused by the potential bias of consumers viewing ingratiating comments as more accurate and objective.

As conversational agents powered by large language models become more human-like, users are starting to view them as companions rather than mere assistants. This shift in perception underscores the importance of understanding the psychological factors influencing human-AI interaction, as they play a crucial role in shaping user experiences, expectations, and the overall effectiveness of AI systems.

3.2.2   Personal Validation Effect and Related Cognitive Biases

The personal validation effect is a psychological phenomenon where individuals highly rate vague personality descriptions as personalized and accurate, despite these descriptions being general enough to apply to many people. Forer's seminal study in 1949 provided early evidence of this effect, demonstrating that participants rated generic personality statements as highly accurate when told they were tailored specifically for them [105]. Subsequent research in the early 1970s further corroborated these findings, showing that students often accept generalized personality interpretations as accurate when they believe the feedback is personally tailored to them [106]. The role of favorability, referring to whether personality descriptions are positive or negative, has been studied, revealing that participants perceived favorable descriptions as more personally true and attributed greater skill to the source of these descriptions [154]. The study also found that the valence of the descriptions was more influential on acceptance than the modality (whether presented orally or in writing).

Underpinning this preference toward positive descriptions is the Pollyanna principle, or the positivity bias, which has been studied in the context of self-report measures, such as quality of life [155], future expectations [156], and when describing other people [157], where people generally gravitate towards positively-valenced responses. A study on social networks found that observers were more accurate in perceiving positive interpersonal ties than negative ones, with negative ties being more frequently overlooked or imagined, which aligns with the Pollyanna principle [158].

More recently, the influence of the personal validation effect have extended to various domains, including recommendation systems. A recent crowd-sourced experiment examined the perceived quality of personalized versus non-personalized movie recommendations [159]. Contrary to expectations, results showed numerically lower mean quality scores for personalized recommendations, albeit without statistical significance. This suggests that Barnum-like effects of personalization may have limited influence on perceived quality in recommender systems.

In the context of large language models (LLMs), research has explored people's perceptions of LLM-generated advice and the role of advice style and user characteristics in shaping these perceptions. A recent study found that different LLM advice styles influenced user perceptions, with "skeptical" style rated as most unpredictable and "whimsical" style as least malicious [160]. Moreover, the study also revealed that individuals who are more agreeable and with higher technological insecurity tend to like the advice more and find it more useful. While this study touches upon the style of LLM outputs, the effect of different valence (positive or negative) of the AI predictions on user perception have yet to be directly studied, which leads to the goal of this study.

Furthermore, cognitive biases including belief bias and confirmation bias could potentially explain the mechanisms that power the effect of personal validation. Belief bias occurs when people evaluate an argument based on how believable its conclusion is, rather than on its logical validity [74, 75]. According to dual process theory in psychology [20, 21], this bias is mainly associated with System 1 thinking, especially when people are under time pressure [90]. However, System 2 thinking can also play a role by creating biased mental models based on pre-existing beliefs, sometimes overriding logical reasoning [21, 91]. The selective processing model [92] suggests that people are more likely to apply logic to arguments that contradict their beliefs, but may neglect logical thinking for arguments that confirm their beliefs, particularly regarding self-related beliefs.

Moreover, confirmation bias is the tendency for individuals to seek out and recall information that supports their existing beliefs [73]. Confirmation bias has been studied in the context of AI-assisted decision-making [161, 162]. In human-AI interaction, this bias can cause people to perceive statements that align with their existing beliefs as more trustworthy and accept them [161]. This selective attention and preference can intensify the personal validation effect, where people give undue credibility to AI outputs that match their expectations

These studies collectively highlight the complex interplay between personal validation effect,

perception of personalized information, and the influence of positive versus negative information in various contexts, from personality assessments to social network perceptions and LLM-generated advice. Building on the rich literature in social and cognitive psychology, this study uniquely contributes by empirically testing how the valence of AI-generated predictions affects user perception and trust, specifically in human-LLM interactions. It offers novel insights into the potential biases at play in human-LLM interaction and suggests strategies to mitigate the negative effects of the personal validation effect in designing AI systems.

3.3 Methods

To address our research question, we utilize the study design, data collection, and analysis framework established in the previous study [24] (See Section 2.2). This section provides a brief overview of the overall approach and the modifications specifically made for this study to address the distinct aspects of the current research question.

3.3.1 Participants

Participants and data collection procedures were consistent with our previous study (See Section 2.2.1). Briefly, data were collected from 238 participants aged 18 and over from various socioeconomic backgrounds, recruited on Prolific. Participants were randomly assigned to either the "Positive" prediction group (N=119) or the "Negative" prediction group (N=119). See Table 2.1 for an overview of participants' demographic information.

3.3.2 Experiment Protocol

The experimental protocol, including the initial astrology and personality assessments and simulated investment game, followed the procedures outlined in our previous study [24] (see Section 2.2.2).

The study used a simulated investment game that was designed to elicit interactions comparable to those of modern robo-advising platforms [40], but in a simplified way. Participants were given $10,000 in virtual currency to allocate across three investment categories (high risk/return, medium risk/return, and low risk/return) over ten rounds (representing years). They could invest up to 100% of their assets across the three categories at each round, with the goal of maximizing the total returns. The game simulated market fluctuations through a series of pre-determined "bull" (gains) and "bear" (losses) market scenarios, affecting potential investment outcomes.

After the final round, participants were exposed to predictions about their future investment behavior. We employed a placebo approach that was proven effective in previous studies [43,

45

44] where participants were informed that predictions were AI-generated, but were in fact pre-determined (either positive or negative), and randomly assigned. The predictions were either of positive valence (e.g., rational investor, higher returns) or negative valence (e.g., impulsive investor, lower returns), with participants randomly assigned to each condition. Each person received predictions from three sources (AI, astrology, personality-based) which were presented in a random order. Examples of these predictions are listed in Table 2.2. Participants then evaluated each prediction for its perceived validity, personalization, reliability, and usefulness (see Section 2.2.3).

Following the experiment, participants were asked to respond to a series of questionnaires on participant characteristics (cognitive style, paranormal beliefs, gullibility, trust in AI/attitudes towards AI, personality, etc.) and demographic information (age, gender, education level). A summary of the scales used in the questionnaires is provided in Table 2.3. Detailed questionnaires can be found in Appendix A.

3.3.3 Analysis

We employed an approach similar to the analysis outlined in our previous study, with some major changes. The variables included in this analysis were the same as those used in the previous study, with subscale scores as the dependent variable and predictors such as prophecy source, prophecy group, cognitive style (composite score), paranormal beliefs, AI attitude/trust in AI, gullibility, Big Five personality traits, interest in the prediction topic, familiarity with prediction sources, and demographic variables (age, gender, and education level). See Section 2.2.5 for the description of key variables.

A mixed-effects model was fitted to the data in a nested long format with a three-level hierarchical structure. Participants were divided into two prediction groups (positive, negative), and each subject engaged in three types of predictions (AI, astrology, personality) and evaluated across four subscales (perceived validity, reliability, usefulness, and personalization). The outcome variable was the *subscale_score*, which represented the scores for each subscale on a 7-point Likert scale. The mixed effects model included fixed effects for subscale, prophecy source, prophecy group, cognitive style, paranormal beliefs, gullibility, AI attitude/trust in AI, big five personality, familiarity with the prediction sources, interest in the prediction topic, age, gender, education level. Interaction terms were included to assess the combined effect of prophecy group and select moderating factors on subscale scores. For the random effects, random intercepts and slopes were included to account for the variability between subjects across different conditions.

A few changes were made to the model specification from the previous study to focus on the impact of prophecy group based on the new hypotheses. Specifically, the focus shifted from *prophecy_source* to *prophecy_group* for interactions with the moderator variables, allowing for an

examination of how different prediction groups influence believability, moderated by these variables. To ensure consistency and enable valid comparisons across models, the random effects structure, correlation, and weighting schemes were retained from the previous model, appropriately accounting for the data structure.

The mixed effects model was specified as follows (changes in the model specification are in bold). The fixed effects were defined as:

$$
\begin{aligned}
\text{subscale\_score} \sim \text{subscale} * (&\text{prophecy\_group} * \text{prophecy\_source} \\
&+ \textbf{prophecy\_group} * \text{composite\_score} + \textbf{prophecy\_group} * \text{paranormal\_score} \\
&+ \textbf{prophecy\_group} * \text{aias\_score} + \textbf{prophecy\_group} * \text{gullibility\_score} \\
&+ \text{big5\_extraversion} + \text{big5\_openness} + \text{big5\_agreeableness} + \text{big5\_conscientiousness} \\
&+ \text{big5\_emotional\_stability} + \text{interest\_behavior} + \text{familiarity} \\
&+ \textbf{prophecy\_group} * \text{Age} + \text{education} + \textbf{prophecy\_group} * \text{gender})
\end{aligned}
$$

(3.1)

For the random effects, we included random intercepts for each participant (*qualtrics_code*) to account for between-subject variability and random slopes for prophecy source to capture the variability in responses across participants to different conditions (*prophecy_source*):

$$
\text{random} = \ \sim \text{prophecy\_source} \mid \text{qualtrics\_code} \tag{3.2}
$$

Additionally, the model incorporated a correlation structure and variance weights to further account for the hierarchical structure and heteroscedasticity in the data:

$$
\begin{aligned}
&\text{correlation} = \text{corSymm(form} = \ 1 \mid \text{qualtrics\_code/prophecy\_source)}, \\
&\quad \text{weights} = \text{varIdent(form} = \ 1 \mid \text{subscale * prophecy\_source * prophecy\_group)}
\end{aligned} \tag{3.3}
$$

Continuous predictors were centered to reduce multicollinearity and to stabilize coefficient estimates. The model was fit using the *lme* function from the *nlme* package in R, with settings to exclude missing values and optimize convergence.

Several diagnostic tests were conducted to ensure the validity of the model. Linearity was confirmed through a residuals vs. fitted values plot, which showed no clear patterns, indicating that the linearity assumption was met. Homoscedasticity was supported by a random spread around zero in the residuals plot, although the multilevel nature of the outcome variable caused some diagonal patterns. The normality of residuals was assessed with a Q-Q plot, showing slight deviations at the extremes but overall consistency with normal distribution. Multicollinearity was addressed by

standardizing the continuous predictor variables, reducing Variance Inflation Factor (VIF) values to acceptable levels (below 5). Independence of residuals was evaluated using the Durbin-Watson test, which revealed significant positive autocorrelation at the subscale level. An autoregressive correlation structure (corAR1) initially helped with the "Validity" subscale but wasn't fully effective for others. Switching to an unstructured symmetric correlation (corSymm) improved the model fit but didn't completely resolve the autocorrelation. This unresolved autocorrelation may pose a limitation in our analysis, suggesting the need for alternative methods or additional data in future studies. Normality of random effects was confirmed through a Q-Q plot, with minor deviations noted. The evaluation of random effects, including variance components and Intraclass Correlation Coefficient (ICC), demonstrated significant variability attributable to these effects.

The results of the model diagnostics and the full results table of the analysis are provided in Appendix C. The code for analysis is in Appendix E, and the full code for implementation is available on GitHub (https://github.com/mitmedialab/ai-prophecy).

3.4 Results

The results from our experiment (N=238) showed that positive predictions are perceived as significantly more valid, personalized, reliable, and useful.

3.4.1 Descriptive statistics

Table 3.1 provides the mean, standard deviation, and the 95% confidence interval for perceived validity, personalization, reliability, and usefulness across Positive and Negative prophecy groups. The Positive prophecy group had higher mean scores across all subscales compared to the Negative group. For perceived validity, the Negative prophecy group had a mean score of 3.70 (SD = 1.79, 95% CI [3.51, 3.88]), compared to 4.89 (SD = 1.32, 95% CI [4.75, 5.03]) in the Positive prophecy group. For perceived personalization, Negative prophecy group had a mean Personalization score of 3.70 (SD = 1.90, 95% CI [3.50, 3.89]), while the Positive prophecy group had a higher mean score of 5.18 (SD = 1.32, 95% CI [5.04, 5.32]). The mean score for perceived reliability in the Negative prophecy group was 3.32 (SD = 1.84, 95% CI [3.13, 3.51]), compared to 4.07 (SD = 1.62, 95% CI [3.90, 4.23]) in the Positive prophecy group. For perceived usefulness, the Negative prophecy group had a mean score of 3.61 (SD = 1.92, 95% CI [3.41, 3.81]), whereas the Positive prophecy group had a mean score of 4.34 (SD = 1.56, 95% CI [4.18, 4.51]).

Table 3.1: Descriptive statistics by prediction valence (prophecy group) and subscale

prophecy_group	subscale	Mean	SD	95% CI
Positive (N=119)	Validity	4.89	1.32	[4.75, 5.03]
	Personalization	5.18	1.32	[5.04, 5.32]
	Reliability	4.07	1.62	[3.90, 4.23]
	Usefulness	4.34	1.56	[4.18, 4.51]
Negative (N=119)	Validity	3.70	1.79	[3.51, 3.88]
	Personalization	3.70	1.90	[3.50, 3.89]
	Reliability	3.32	1.84	[3.13, 3.51]
	Usefulness	3.61	1.92	[3.41, 3.81]

3.4.2 Positive predictions are perceived as significantly more valid, personalized, reliable, and useful than negative predictions.

The mixed effects model analysis showed that positive AI predictions generally elicited significantly more favorable responses compared to negative ones, consistent with the Pollyanna principle. The intercept was 5.20 (95% CI [4.81, 5.59], p < 0.001) on a 7-point Likert scale, and this was associated with "AI" prophecy source, "Validity" subscale, "Positive" prediction group, "Female" gender, and "Bachelor" education level, which served as the reference level for subsequent interpretation. The main effects of the prediction valence on perceived validity score showed that compared to the baseline Positive prediction group, the Negative prediction group was associated with a decrease in perceived validity by 1.37 points (95% CI [-1.83, -0.91], p < 0.001).

Interactions between the prediction groups and subscales revealed significant results that suggested differential impact of prediction valence across subscales. Compared to perceived validity baseline, perceived personalization further decreased by 0.25, which was significant (95% CI [-0.43, -0.06], p = 0.008). On the other hand, perceived reliability significantly increased by 0.46, partially reversing the negative main effect (95% CI [0.17, 0.75], p = 0.002). Similarly, perceived usefulness significantly increased by 0.56 (95% CI [0.31, 0.81], p < 0.001). This suggested that while all subscales were impacted by the valence of predictions, the influence was stronger for perceived personalization and less strong for perceived reliability and usefulness compared to the baseline.

There were significant main effects of subscales on the perception of positive AI predictions. On average, perceived personalization was rated 0.30 point higher than perceived validity (95% CI [0.15, 0.44], p < 0.001). On the other hand, perceived reliability was rated 0.90 points lower than perceived validity (95% CI [-1.12, -0.68], p < 0.001), and perceived usefulness was rated 0.67 points lower than perceived validity (95% CI [-0.86, -0.48], p < 0.001). The detailed results are

Table 3.2: Percentage differences of estimates for each subscale by prediction valence

	Negative	Positive	Δ Percentage
Validity	3.83	5.20	36%
Personalization	3.88	5.50	42%
Reliability	3.39	4.30	27%
Usefulness	3.72	4.53	22%

provided in the supplementary materials.

To get a sense of the magnitude of the impact of prediction valence across subscales, the percentage differences in the estimated values of subscale scores were calculated based on the coefficients from the mixed effects model results. Positive predictions increased perceived validity by 36% and perceived personalization by 42%. Perceived reliability increased by 27% and perceived usefulness was increased by 22% compared to negative predictions. This comparison is visualized in Figure 3.1 and Table 3.2.

Figure 3.1: Comparison of estimates by prediction valence for each subscale

Standard errors are represented as error bars, and subscale score range is from 1 to 7. All differences were significant ($p < 0.01$**).
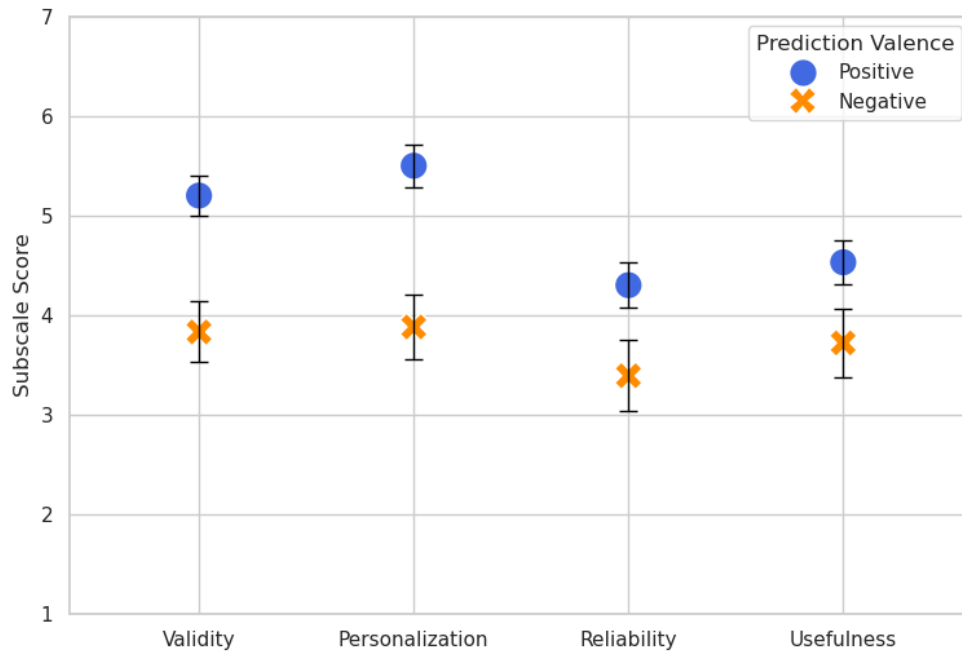
Moreover, the main effects of prophecy source on believability of predictions showed that on average, perceived validity of astrology-based predictions was 0.66 points lower compared to AI predictions (95% CI [-0.89 -0.43], $p < 0.001$). The difference was not statistically significant for

personality-based predictions (0.06, 95% CI [-0.14, 0.27], p = 0.549). The interaction between Astrology prediction source and Negative prediction group was not significant (0.09, 95% CI [-0.23, 0.41], p = 0.598) as well as the interaction between Personality prediction source and Negative prediction group (0.02, 95% CI [-0.27, 0.31], p = 0.881), which suggested that the negative effect of receiving the negative predictions was comparable across different prediction sources.

3.4.3   Analysis of moderating factors

Although not initially pre-registered, the inclusion of psychological factors such as paranormal belief provided valuable insights. We observed that these factors moderated the effect of AI prediction valence on believability, highlighting a complex interplay between individual predispositions and the nature of AI communications.

Cognitive style

The composite cognitive score was measured with a combined score of the cognitive reflection test (CRT-2) [45] and need for cognition (NCS-6) [46], and this continuous variable ranged from -4.29 to 2.61 (Mean = 0.0, SD = 1.44). Cognitive style was not found to be a significant moderator of the impact of AI prediction valence on perceived validity (Estimate = -0.16, 95% CI [-0.39, 0.08], p = 0.196).

However, interactions across subscales revealed that cognitive style significantly moderated the impact of negative prediction valence on perceived reliability. The previous section discussed that perceived reliability was significantly reduced with negative predictions, but this was less compared to the perceived validity baseline, as seen in the slight increase in the estimate (Estimate = 0.46, 95% CI [0.17, 0.75], p = 0.002). However, this effect was slightly lessened with higher composite scores, which represents a more analytic cognitive style (Estimate = 0.11, 95% CI [0.00, 0.22], p = 0.046). This suggests that those with more analytic cognitive style are less influenced by the difference in prediction valence, thus less prone to the positivity bias.

Paranormal beliefs

We explored whether paranormal beliefs moderated the relationship between prediction valence and believability. Paranormal beliefs, as measured by a revised version of R-PBS [47], were considered as a continuous moderating variable, with scores ranging from 15 to 95 (Mean = 45.6, SD = 20.08).

The interaction between prediction valence and paranormal beliefs was not significant (Estimate = 0.01, 95% CI [-0.01, 0.02], p = 0.539), indicating that the effect of prediction valence on perceived validity does not depend on the level of paranormal beliefs.

Looking at interactions across subscales, paranormal beliefs significantly moderated the effect of prediction valence on perceived personalization. While negative predictions generally led to a larger decrease in perceived personalization compared to perceived validity baseline (Estimate = -0.25, 95% CI [-0.43, -0.06], p = 0.008), this effect was slightly attenuated by higher paranormal beliefs score (Estimate = 0.01, 95% CI [0.00, 0.01], p = 0.022). On the other hand, the effect of negative valence on perceived reliability (Estimate = 0.46, 95% CI [0.17, 0.75], p = 0.002) was slightly exacerbated by higher paranormal beliefs score (Estimate = -0.01, 95% CI [-0.02, -0.00], p = 0.028). This indicates that the level of paranormal beliefs slightly moderates the effect of prediction valence on perceived personalization and reliability.

Gullibility

The moderating effect of gullibility on the impact of prediction valence on believability was assessed. Gullibility, as measured by a self-report gullibility scale [49], was a continuous moderating variable, with scores ranging from 6 to 39 (Mean = 14.68, SD = 7.69).

The significant main effect of negative predictions on perceived validity (Estimate = -1.37, 95% CI [-1.83, -0.91], p < 0.001) as described in Section 3.4.2 was slightly reduced for individuals with higher gullibility score, as seen in the significant interaction between Negative prophecy group and gullibility score (Estimate = 0.06, 95% CI [0.01, 0.10], p = 0.010), suggesting a moderating role of gullibility. However, this effect did not vary significantly across subscales. This implies that the more gullible an individual is, the less likely that the valence of predictions will influence their belief in those predictions, which may be counter-intuitive.

Additionally, contrary to the findings from the previous study [24], we found the negative relationship between gullibility and perceived validity to be significant (Estimate = -0.05, 95% CI [-0.09, -0.02], p = 0.001). However, this effect was reversed for those who received negative predictions (Estimate = 0.06, 95% CI [0.01, 0.10], p = 0.010). These seemingly contradictory results for gullibility are discussed in Section 3.5.

AI attitude/Trust in AI

We investigated whether AI attitude/trust in AI influenced the relationship between prediction valence and believability. Attitude toward AI was measured using AIAS-4 [48], and this continuous moderating variable ranged from 4 to 40 (Mean = 25.79, SD = 8.68), where higher scores indicated more positive attitudes toward AI.

While the interaction between AIAS score and Negative prophecy group approached significance (Estimate = 0.04, 95% CI [-0.00, 0.07], p = 0.058), it was not statistically significant. This indicates that the effect of prediction valence on perceived validity does not depends on the level

of AI attitude/trust.

Interactions with other subscales revealed significant moderating effects of AI attitude on the effect of negative predictions on perceived reliability and usefulness. Higher AIAS scores, indicating more positive attitudes toward AI, led to a slight decrease in the impact of negative predictions on perceived reliability (Estimate = -0.02, 95% CI [-0.04, -0.00], p = 0.021) as well as perceived usefulness (Estimate = -0.02, 95% CI [-0.03, -0.00], p = 0.034). This significant interaction suggests that for perceived reliability and usefulness, higher AI attitude actually amplifies the negative effect of negative predictions.

Age

Interaction between age and the negative prophecy group was not significant (Estimate = 0.01, 95% CI [-0.02, 0.04], p = 0.519), suggesting that age is not a significant moderator on the impact of prediction valence on perceived validity. Age is a continuous variable and ranged from 19 to 75 (Mean = 41.04, SD = 12.38).

While the impact of negative predictions on perceived reliability was significant but less pronounced than the impact on perceived validity or personalization (Estimate = 0.46, 95% CI [0.17, 0.75], p = 0.002), older age further reduced the impact of negative predictions on perceived reliability (Estimate = 0.02, 95% CI [0.00, 0.03], p = 0.010). This suggests that older age may lead to decreased influence of the positivity bias.

Gender

Gender was not a significant moderator of the impact of prediction valence on believability of AI predictions. Interactions between the Negative prophecy group and gender was not significant (Male: Estimate = 0.35, 95% CI [-0.35, 1.05], p = 0.322; Other: Estimate = 0.83, 95% CI [-0.97, 2.64], p = 0.365) compared to the female baseline, suggesting that the impact of prediction valence on perceived validity is consistent across gender. Similarly, there were no significant interaction effects across other subscales. The detailed results table can be found in Appendix C

3.5 Discussion

Our results demonstrate that positive AI-generated predictions are perceived as significantly more valid, personalized, reliable, and useful compared to negative predictions. This finding aligns with the personal validation effect, also known as the Barnum effect, where individuals tend to accept vague and general personality descriptions as uniquely applicable to themselves. Moreover, the findings confirm the Pollyanna principle, or positivity bias, that people find positive pre-

dictions more believable than negative ones. The significant main effect of prediction group and the non-significant interaction with astrology and personality prediction sources together support the hypothesis that people generally tend to believe in positive predictions more, irrespective of the source of the predictions.

Perceived validity

The substantial difference in perceived validity between positive and negative predictions suggests that participants were more likely to believe in the accuracy of favorable forecasts. This tendency may be attributed to a combination of cognitive biases, including confirmation bias and belief bias, as described in [24]. Belief bias refers to the tendency to evaluate statements based on their believability and alignment with their existing beliefs rather than their logical validity [74, 75]. This can lead to people to accept AI predictions as believable even when they are fictitious or misleading. Moreover, in terms of predictions about their personal behavior, people generally prefer information that confirms their existing beliefs or desires about themselves, and positive predictions likely align more closely with individuals' hopes and self-perceptions.

Perceived personalization

While the positivity bias can be observed across all subscales, the marked increase in perceived personalization for positive predictions is particularly noteworthy. The perception of personalization is stronger than other subscales when predictions are positive, but similar to others when predictions are negative, as seen in Figure 3.1. This finding validates that the positivity bias is an influential working principle of the personal validation effect.

This effect may also be explained by the self-serving bias [163, 164], the tendency to attribute our successes to internal, personal factors, and our failures to external, situational factors. When presented with favorable predictions, participants may have been more inclined to view them as tailored to their unique qualities or circumstances.

Perceived reliability and usefulness

In addition, the higher ratings for reliability and usefulness of positive predictions further underscore the influence of personal validation. Participants may have judged favorable predictions as more dependable and actionable, possibly due to the illusion of control [165]. Positive forecasts might be seen as more reliable because they align with desired outcomes, and more useful because they support optimistic future planning.

54

Factors that moderate the personal validation effect

Our analysis revealed several factors that moderate the impact of the personal validation effect on the perception of AI predictions. These findings provide nuanced insights into how individual differences and psychological factors influence the susceptibility to positive bias in AI-generated predictions.

Interestingly, we found that cognitive style significantly moderated the effect of prediction valence on perceived reliability, but not on other subscales. This indicates that individuals with a more analytical cognitive style are less likely to have changed perceptions of reliability when the predictions are negative versus positive, possibly making them less susceptible to the personal validation effect. While cognitive style was not found to a significant predictor of believability of predictions based on the previous study [24], our results reveal that it could moderate the level of influence of the positivity bias and the personal validation effect.

While paranormal beliefs did not significantly moderate the overall effect of prediction valence on believability, they differentially moderated the impact of negative predictions across subscales, particularly for perceived personalization and reliability. Paranormal beliefs were found to mitigate the negative impact of negative predictions for perceived personalization, suggesting that those who hold stronger paranormal beliefs were less likely to perceive personalized predictions negatively when faced with negative predictions. On the other hand, paranormal beliefs amplified the negative impact of negative predictions, meaning that individuals with higher paranormal beliefs saw negative predictions as more damaging to the reliability of the information. This highlights the role of individual differences on the effect of personal validation.

Gullibility emerged as a significant moderator of the effect of prediction valence on believability. Those who reported to be more gullible were less influenced by the different valence of predictions, suggesting they are more resistant to positivity bias. While the results seem counter-intuitive, this contradiction may stem from the self-report nature of the gullibility scale as well as the fact that it was administered after participants had rated their belief in the predictions, making them biased in their self-perception of gullibility. Moreover, the negative impact of gullibility on perceived believability was reversed when predictions were negative. This could suggest that people were more likely to be biased in their responses to the self-reported gullibility scale when predictions were positive, but not when predictions were negative, as they were less likely to admit that their trust in the positive predictions were unfounded.

Attitudes toward AI played a role in moderating the effect of prediction valence on perceived reliability and usefulness. Participants with more positive attitudes toward AI showed a stronger effect of prediction valence on these dimensions. This suggests that individuals who are more favorable towards AI technology may be more susceptible to the positivity bias of the personal

validation effect, potentially due to higher trust or expectations in AI capabilities.

Age was found to moderate the impact of prediction valence on perceived reliability, with older participants showing a reduced effect of negative predictions. This finding implies that older individuals might be less swayed by the positivity bias in AI interactions, possibly due to greater life experience or a more critical perspective developed over time.

Notably, gender did not significantly moderate the impact of prediction valence on the believability of AI predictions. This suggests that the personal validation effect and its influence on perceptions of AI-generated content may be relatively consistent across gender identities.

The exploratory findings on the role of moderating factors highlight the complex interplay between individual characteristics and the personal validation effect in AI interactions. Understanding these nuances is crucial for developing more effective and ethical AI systems that can adapt to individual differences and mitigate potential biases in user perceptions. Future research should further explore these moderating factors and their implications for designing AI interfaces and interaction paradigms that can accommodate diverse user characteristics and cognitive styles.

3.5.1 General Discussion

Our study contributes to and expands upon the existing body of knowledge in human-AI interaction. While previous studies have focused on how users' perceptions of AI motives affect their evaluations of AI [83], our research shows that the content itself—specifically its positive or negative framing—can have a profound impact on user judgments, even in the absence of perceived AI intentions.

Furthermore, our study contributes to the growing body of research on placebo effects in AI interactions [43, 44, 82]. By demonstrating that users' biases extend beyond performance expectations to their evaluation of AI-generated content, we provide a more comprehensive understanding of how psychological factors shape human-AI interactions. The observed preference for positive predictions in our study adds a new dimension to recent investigations into sycophantic behavior in language models [103, 153]. Our work empirically demonstrates that positively-biased model behavior affects user perception beyond simple misinformation, potentially reinforcing users' positivity bias and creating a feedback loop. Models that prioritize user favor over accuracy can be more persuasive and influential, affecting users beyond the accuracy of their outputs.

These findings have important implications for AI system design and deployment, particularly in generating predictions or personalized recommendations. Users' tendency to favor positive AI-generated content could lead to over-reliance on AI, potentially resulting in poor decisions if the predictions are inaccurate or biased. This bias could also be exploited by malicious actors to manipulate user behavior or beliefs through strategically positive messaging.

In the context of human-AI interaction, these results highlight the need for increased aware-

ness and education about cognitive biases. Users should be encouraged to critically evaluate AI-generated predictions, regardless of their valence. Additionally, designers of AI systems should consider implementing mechanisms to mitigate the effects of personal validation, such as using neutral tones in predictions, providing balanced perspectives, or explicitly highlighting the potential for bias in user perceptions.

3.5.2 Limitations and Future Directions

While our study provides valuable insights into the influence of personal validation on human-AI interaction, it has several limitations that point to important avenues for future research.

First, our experiment focused on single-turn interactions with the predictions. In real-world scenarios, human-AI interactions often involve multiple turns of dialogue, which could potentially amplify the effects we observed. Future studies should investigate how the personal validation effect evolves over extended interactions, similar [83] examining whether the effect strengthens or diminishes with repeated exposure to positive or negative predictions. This longitudinal approach could reveal important dynamics in the formation and persistence of biases in human-AI relationships.

Second, the predictions used in our experiment were presented in a uniform format. However, AI systems in practice often employ various features that could influence user perceptions. Future studies should investigate how AI characteristics such as personalization, the use of virtual avatars, or speech-based interactions could potentially augment or alter the observed personal validation effect. For instance, would an embodied AI agent delivering positive predictions have a stronger influence on user perceptions compared to text-based interactions? Understanding these nuances could provide valuable insights for designing more effective and responsible AI interfaces.

Third, our study was limited to a specific domain of predictions, which emphasizes the need to extend this research into high-impact fields where the consequences of biased perception could be more severe. For instance, in the medical field, an AI system that consistently reinforces a doctor's decisions might inadvertently foster a personal bias, potentially leading to skewed medical judgments and compromising patient outcomes. Similarly, in the realm of education, an AI tutor that provides continuous positive reinforcement without accurate assessment of a student's progress could lead to overconfidence, undermining genuine skill development and long-term learning. These scenarios represent critical areas for future investigation, as they could have significant real-world implications for patient care and student development.

Furthermore, while this study focused on the perception of AI predictions, it raises important questions about how these perceptions influence actual user behavior. Expanding the experimental framework to include measures of behavioral changes and decision-making outcomes could provide valuable insights into the broader impact of cognitive and psychological factors in human-AI

interaction. Understanding this could help in designing AI systems that not only foster appropriate trust but also encourage sound decision-making in users.

Additionally, future research should explore potential mitigation strategies for the personal validation effect in AI interactions. This could involve developing and testing different approaches to presenting AI-generated predictions. Investigating the effectiveness of user education programs in reducing susceptibility to this bias could also yield valuable insights for improving AI literacy.

Lastly, cross-cultural studies could provide a more comprehensive understanding of personal validation effect in AI interactions. Cultural differences in communication styles, attitudes towards technology, and responses to positive feedback might influence how this bias manifests across different populations. Such research could inform the development of culturally sensitive AI systems and interaction guidelines.

By addressing these limitations and pursuing these future directions, researchers can build a more comprehensive understanding of the personal validation effect in human-AI interaction, ultimately contributing to the development of more effective, ethical, and user-centered AI systems.

3.6 Conclusion

In conclusion, our study demonstrates that the personal validation effect and positivity bias significantly influence how users perceive AI-generated predictions. Positive predictions are consistently rated as more valid, personalized, reliable, and useful compared to negative predictions. Moreover, the study presents an exploratory analysis of how cognitive style, paranormal beliefs, gullibility, AI attitude/trust, age, and gender moderate the negative impact of negative predictions on perceived validity, personalization, reliability, and usefulness, producing some nuanced findings that open up opportunities for future work. The findings of this study underscore the complex psychological dynamics at play in human-AI interaction and highlight the need for careful consideration of cognitive biases in the design and deployment of AI systems. As AI continues to play an increasingly prominent role in decision-making and personal assistance, understanding and mitigating the effects of personal validation will be crucial for fostering more balanced and effective human-AI collaboration.

Chapter 4

Future Directions: Investigating Self-Fulfilling Prophecies in Human-AI Interaction

With the rapid advancement and adoption of AI in society, individuals are increasingly exposed to algorithms that make personalized predictions and recommendations based on their data. These applications range from user modeling based on web interactions for tailored experiences [166] to sophisticated AI assistants that continuously adapt user representations through ongoing interactions [167]. Such technologies are widely applied in e-commerce, adaptive learning systems, targeted advertising, personalized content delivery, and healthcare. Consequently, users are constantly subjected to the "expectations" set by AI systems regarding their preferences, behaviors, and future actions.

The concept of the self-fulfilling prophecy refers to situations where prior expectations influence subsequent behavior in a way that ultimately confirms those expectations [168, 169]. Originally conceptualized by sociologist Robert K. Merton in 1948, this phenomenon has been widely studied across disciplines such as psychology, sociology, business, medicine, education, economics, and law [170, 171]. A related concept, the Pygmalion effect, occurs on an interpersonal level when one person's expectations influence another's behavior, leading to the fulfillment of the initial expectations [169, 172, 173]. This effect has been notably explored in education, where teachers' expectations significantly impact students' academic performance [169, 173, 174]. Despite the broad exploration of these concepts, their implications within human-AI interaction remain underexplored, offering a valuable opportunity to investigate how AI's expectations might influence user behavior. This raises a critical question: Can AI's expectations of individuals shape their actual behavior in a way that fulfills those predictions?

This exploration is particularly relevant in light of concerns about advanced AI models, such

as large language models (LLMs). These models have been critiqued for displaying sycophantic behavior, where they tend to agree with users' inputs regardless of accuracy [103], and for creating echo chambers that reinforce existing beliefs rather than challenging them [104]. Such tendencies can exacerbate the self-fulfilling prophecy effect, as users might increasingly rely on AI-generated feedback that mirrors their existing views, further entrenching those views. For instance, if an LLM consistently affirms a user's overconfidence in investment decisions, it could lead the user to make riskier choices, thereby validating the AI's predictions and perpetuating a cycle of reinforcement. These concerns highlight the importance of critically examining how AI systems influence user behavior and the potential for these technologies to inadvertently shape, rather than simply reflect, human actions. Understanding these dynamics is crucial for designing AI systems that avoid the pitfalls of manipulation and instead foster more balanced and beneficial interactions.

This chapter outlines the direction of a future work in studying the phenomenon of self-fulfilling prophecy in the context of human-AI interaction. Through an experimental design that appears to provide personalized predictions on their investment behavior and future outcome, we investigate the extent to which fictitious AI predictions can serve as self-fulfilling prophecies. More specifically, we aim to investigate whether AI predictions on one's emotional discipline in an simulated investment game context can impact their actual emotional discipline and investment behavior in later stages. The study also explores the role of moderating factors such as affective state and trust in the realization of AI predictions. By studying the phenomenon of self-fulfilling prophecy and key moderating factors in the context of human-AI interaction, we hope to draw generalizable insights on the design of AI systems in a way that minimizes the risk of manipulation and increases potential benefits.

4.1 Related Works

4.1.1 Self-fulfilling prophecy in human-AI interaction

The concept of self-fulfilling prophecies in human-AI interaction is an emerging area of study, with significant implications for how AI systems influence user behavior. Lawrence Switzky discusses the "ELIZA Effect" in his work Pygmalion and the Early Development of Artificial Intelligence, highlighting the human tendency to attribute greater intelligence and intention to responsive computers than is warranted [175]. This phenomenon underscores the potential for AI systems to shape user perceptions and actions in ways that align with the AI's predictions, even when those predictions are based on flawed or simplistic models.

De-Arteaga and Elmer (2023) have identified several ways in which machine learning systems can perpetuate or create self-fulfilling prophecies, particularly in healthcare. These include encod-

ing flawed human beliefs into algorithms, reinforcing biases through human-machine interaction cycles, making inaccurate predictions that users act upon, and relying on outdated clinical decisions despite advancements in medical knowledge [176]. Such processes highlight the potential for AI systems to not only reflect but also amplify existing biases, resulting in outcomes that may seem accurate simply because they have shaped user behavior to match the predictions.

Feedback loops play a critical role in the self-fulfilling prophecy effect within AI systems. King and Mertens (2023) argue that misunderstandings between predictions and their subjects can lead to accountability issues and feedback loops that reinforce these prophecies [177]. Similarly, Bauer (2023) discusses how algorithmic score disclosures can influence individual behavior, creating self-fulfilling prophecies where people align their actions with erroneous assessments [178]. This phenomenon is particularly concerning in settings where biased machine learning models are continuously retrained on new data that the models themselves help generate. In such cases, the AI's predictions may appear increasingly accurate, not because they reflect reality, but because users adjust their behavior to fit the AI's expectations, reinforcing the model's biases.

Similarly, Bauer and colleagues (2023) found that disclosing algorithmic scores can lead to anchoring bias, where individuals adjust their behavior to align with the scores, thus reinforcing the predictions and creating a feedback loop that perpetuates bias [178]. This is particularly problematic in scenarios where machine learning models are continuously retrained on data they have influenced, as biased outputs can lead to behaviors that further entrench the biases in the model.

The implications of these feedback loops are evident in areas like predictive policing, where biased models can lead to runaway cycles of reinforcement, sending police to the same neighborhoods repeatedly regardless of the true crime rate [179]. Similarly, in economic decision-making, speculative forecasts can influence behavior before any actual changes occur, as demonstrated by Petropoulos Petalas et al. [180]. This study, though not directly related to AI, highlights how predictions can alter decision-making and risk-taking, further illustrating the broader impact of self-fulfilling prophecies in contexts where accurate predictions are critical.

Self-fulfilling prophecies have also been observed in the context of placebo AI treatments. Recent studies have revealed that individuals' beliefs about an AI's capabilities can influence their responses to the system, even when the AI is not functioning as claimed [43, 44, 82]. This highlights the power of perception in shaping user behavior and the potential for AI systems to unintentionally manipulate or reinforce certain actions.

Given this backdrop, it is crucial to explore the phenomenon of self-fulfilling prophecies in human-AI interaction, particularly in contexts like financial decision-making, where risk and outcomes are heavily influenced by user behavior. Understanding how AI expectations influence attitudes toward risk can shed light on the complex interplay between human psychology and artificial intelligence, offering insights into how to design AI systems that minimize the risk of manipulation

and maximize their benefits.

4.1.2 Emotions and risk-taking behavior

Emotions play a significant role in influencing risk-taking behavior, often serving as a critical determinant in decision-making processes. One key theory is the "mood maintenance hypothesis," which posits that individuals in a positive emotional state are less likely to engage in risky behavior, as they wish to preserve their current affective state. This theory is supported by findings that suggest individuals experiencing positive affect exhibit more conservative decision-making behavior compared to those in neutral or negative states [181–184]. Conversely, those in a negative emotional state may engage in riskier behavior in an attempt to improve their mood, often focusing on the potential positive outcomes rather than the more probable negative ones [185, 186].

Integral emotions, which arise directly from the decision at hand, play a crucial role in shaping risk-taking behavior. Unlike incidental emotions, which are unrelated to the decision but still influence it, integral emotions are considered a normatively defensible input to judgment and decision-making [187]. Immediate emotions, a subset of integral emotions, refer to the feelings experienced in response to the options being considered, such as fear or excitement, which can motivate individuals to either approach or avoid certain risks. These immediate emotions differ from anticipated emotions, which involve predictions about how one will feel after the outcomes of a decision [187–189]. The framing of risky choices also affects immediate emotional responses, which can influence decision-making. When a choice is framed as a potential loss, individuals are more likely to experience negative emotions, leading to a higher likelihood of risk-taking as they seek to avoid the perceived loss [190]. Together, these emotional processes highlight the complexity of how emotions influence risk-taking behavior, often dictating the choices individuals make under uncertainty.

4.1.3 Emotional regulation in investment decision-making

Emotion regulation involves efforts to alter the course of an ongoing emotional response [191]. Individuals typically employ various strategies to manage their emotions, which significantly affect their emotional well-being, including cognitive reappraisal, where individuals modify their thoughts about a situation to change its emotional impact, and expressive suppression, which involves restraining the external display of emotions gross_individual_2003.

Emotional regulation plays a critical role in decision-making, particularly in high-stakes environments like investing. Emotional regulation involves a sequence of stages, including the identification of the need to regulate emotions, selecting an appropriate strategy, and implementing that strategy effectively [192]. The ability to regulate emotions is vital for maintaining goal-directed behavior, as difficulties in this area can impair performance in cognitive tasks [193]. For instance, a study

suggests that effective emotion regulation strategies support better decision-making by promoting goal-oriented actions [194].

In the context of investing, emotional regulation becomes even more crucial due to the complexity and uncertainty involved in financial decisions. Overconfidence, often driven by strong positive emotions like pride, can lead to detrimental outcomes such as increased trading volumes and a stronger disposition effect [195]. When emotions are not adequately regulated, they can overwhelm cognitive processes, leading to poor decision-making and increased risk-taking behaviors [196]. As investment decisions are inherently complex, the lack of professional expertise often forces individuals to rely on their emotions rather than rational analysis, further underscoring the importance of emotional regulation in financial contexts [197].

4.2 Proposed Study Design

4.2.1 Participants

The study will involve participants aged 18 and above, representing a diverse range of financial literacy, investment experience, emotional regulation skills, and familiarity with AI. Participants will be randomly assigned to one of two groups: positive prediction or negative prediction. This randomization ensures a balanced comparison of the effects of prediction valence on participant behavior and emotions.

4.2.2 Experiment protocol

The study will utilize a simulated investment game, building on the design of previous studies. Each participant will be provided with a virtual currency of $10,000 and will be asked to invest this amount across three categories: high risk/high return, medium risk/medium return, and low risk/low return. The investment game will consist of two phases:

- **Pre-prophecy phase:** Participants will complete 10 rounds of investment decisions.

- **Post-prophecy phase:** Participants will receive an "AI-generated" prophecy regarding their emotional regulation in investing, after which they will complete additional rounds of investment decisions. We are considering two possible designs for this phase:

    - **Single prophecy:** The prophecy is provided once after the first 10 rounds, followed by 10 more rounds of investment.

– **Multiple prophecies:** The prophecy is provided 2-3 times throughout the second phase, allowing us to observe how repeated feedback might affect behavior and emotions over time.

The proposed experimental protocol is provided in Figure 4.1.
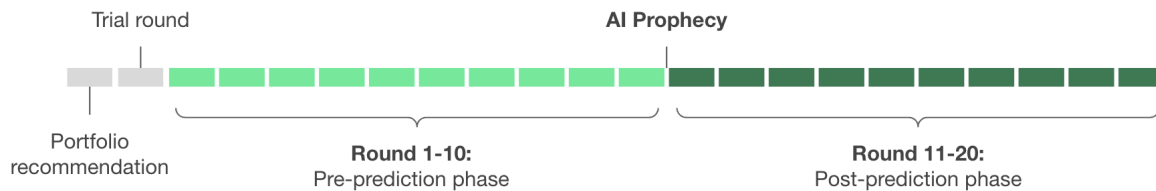
Figure 4.1: Proposed study procedure

We hypothesize that the valence of the prediction will significantly influence both the participants' investment behavior and emotional responses. Specifically, positive predictions are expected to lead to behavior that reflects higher emotional regulation, while negative predictions may result in behavior consistent with lower emotional regulation, demonstrating a self-fulfilling prophecy effect. Additionally, we anticipate to see significant differences in allocation behavior, emotional state, and decision time post-prediction. While affective state will be measured as a dependent variable, we will also explore its potential role as a mediator or moderator in the relationship between prediction valence and investment behavior, along with the potential influences of other individual differences.

4.2.3 Data collection

The study will collect data on the following variables based on the game interaction:

- Shift in investment allocation behavior before and after receiving the prediction
- Affective state (self-reported integral emotion at each round)
- Decision time (time taken to make investment decisions at each round)

Additionally, participants will complete surveys assessing various individual differences, including but not limited to:

- Self-reported emotional regulation (before and after the game)
- Belief in prediction (validity, reliability, usefulness, personalization)
- Risk preference
- Level of financial literacy/investing experience

- Attitude toward AI

- Personality traits

- Incidental emotion/state affect

4.2.4 Data analysis

The data will be analyzed to assess the impact of fictitious AI predictions on investment behavior, emotional responses, and self-perception of emotional regulation. Additionally, we will explore how affective state and changes in self-perception may serve as both mediators and moderators in the relationship between AI predictions and behavioral outcomes.

Dependent Variables

We will first examine how AI predictions influence the following key outcomes:

- **Behavioral impact:** The shift in investment allocation before and after receiving the prediction, as well as decision time.

- **Emotional impact:** Differences in affective state (integral emotion) and decision time between the positive and negative prediction groups.

- **Self-perception impact:** Changes in self-perception of emotional regulation before and after the study, assessing the reported influence of the prophecy.

Mixed effects model

We will employ mixed-effects models to account for repeated measures within participants and the hierarchical nature of the data. This approach will allow us to model individual differences (random effects) and the overall effects of the AI prediction, prediction valence, and other relevant factors on our dependent variables: investment behavior, affective state, decision time, and change in self-perception of emotional regulation.

- **Fixed effects:** The primary fixed effects will include the presence of a prediction (prediction vs. no prediction), prediction valence (positive vs. negative), and their interaction. Additional factors, such as emotional regulation skills, risk preference, trust, and perceived belief in the AI-generated prophecy (including subscales for validity, personalization, reliability, and usefulness), will also be included to explore their influence on the impact of the AI predictions.

- **Random effects:** Random intercepts will be included to account for individual differences in baseline levels of the dependent variables. If variability in responses to the AI predictions is observed, random slopes may be included to model this variability over time.

Mediation analysis

We will explore whether changes in self-perception of emotional regulation and affective state mediate the relationship between receiving an AI prediction and investment behavior. This analysis will help determine if these internal psychological changes explain how AI predictions influence behavior.

Moderation analysis

We will investigate whether changes in self-perception, alongside other factors such as emotional regulation skills, risk preference, trust, and perceived belief, moderate the impact of AI predictions on behavior and affective state. This analysis will help identify if the effects of AI predictions on behavior differ depending on how participants' self-perception changes or how they perceive the AI-generated prophecy.

This mixed-effects modeling approach, combined with mediation and moderation analysis, will provide a robust framework for understanding how fictitious AI-generated predictions influence participant behavior and emotional responses, and how these effects are shaped by the valence of the predictions, perceived belief of the predictions, and other individual differences. The analysis plan is flexible and may be adjusted if the data suggest a different approach is more suitable.

Chapter 5

Conclusion

This thesis explored the complex psychological dynamics underlying human-AI interaction, with a focus on the interplay between belief, perception, and behavior.

Chapter 2 investigated the phenomenon of "rational superstition" in the context of human-AI interaction, revealing that people's perceptions of AI predictions can be irrationally influenced by psychological factors such as paranormal beliefs, cognitive biases, and incorrect mental models. The findings demonstrated that, much like astrology, AI predictions can be seen as trustworthy and reliable even without validation, driven by a combination of idealized portrayals of AI and mental heuristics. This underscores the necessity of critically assessing how AI systems are presented and understood, to avoid unwarranted trust and over-reliance.

Chapter 3 delved into the personal validation effect within AI-generated predictions, particularly focusing on how positive framing can bias user perceptions of validity, personalization, reliability, and usefulness. The study highlighted the substantial influence that positive AI responses have on user perception, which can lead to overconfidence in AI systems and poor decision-making. These findings stress the importance of designing AI systems that are not only accurate but also transparent and balanced, mitigating the risks associated with cognitive biases.

Chapter 4 proposed future work to explore the concept of the self-fulfilling prophecy in human-AI interaction. The proposed study aims to investigate how AI predictions might shape user behavior, particularly in contexts where the predictions are about themselves. This exploration is crucial as it addresses the potential for AI systems to inadvertently influence behavior in ways that reinforce existing biases or expectations, further highlighting the importance of responsible AI design.

Together, this thesis underscores the critical role that psychological factors play in shaping human-AI interaction and highlights the importance of accounting for human irrationality in AI system design. As AI systems become increasingly integrated into daily life, it is essential to design them and educate users in ways that foster informed and critical engagement, rather than blind trust or over-reliance. By understanding and mitigating the effects of cognitive biases and psychological

influences, we can ensure that AI serves as a tool for enhancing human decision-making, rather than undermining it.

Appendix A

Survey Instruments

A.1 Consent Form

This consent form was administered on Qualtrics.

Welcome! This study aims to understand what makes people believe in predictions about themselves.

The study should take around 30 min to complete. You are eligible to participate in this study if you are 18 or above, and you have normal or corrected-to-normal vision and hearing.

Your responses will be kept confidential. Results from this study may be used for future academic research publications.

Please note that:
1. Your participation is voluntary.
2. Some questions are mandatory. However, if you do not want to answer them, you can exit the study.
3. At any time, you may decline further participation without adverse consequences. To do this, simply close out of the study without submitting your answers.
4. You will remain anonymous in any publications or presentations based on the results of this study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

By clicking on "Next" below, you acknowledge: Your participation in the study is voluntary. You are 18 years of age or above. You are aware that you may choose to terminate your participation at any time for any reason. Please select if you consent to begin the study.

A.2   Experiment instructions

These instructions were provided on a web app embedded on Qualtrics.

You will complete a short questionnaire about your birthdate and your personality, then participate in a simulated investing game

At the end, you will be provided with three different versions of predictions about your future investment behavior, based on analysis based on: 1) your zodiac sign, 2) your personality, and 3) your interaction data monitored by our AI model.

Please answer honestly and to the best of your ability for accurate predictions. The information you provide in the questionnaire will only be used for the purpose of creating personalized answers for you and will not be shared with anyone else.

A.3   Astrology and Personality Questionnaires

Zodiac sign

What is your birthdate? MM/DD/YYYY

(Optional) What is the local time you were born? hh/mm

(Optional) Which city were you born in?

MBTI Personality test (shortened version)

Likert scale of 1 (strongly disagree) to 7 (strongly agree)

You regularly make new friends.

You spend a lot of your free time exploring various random topics that pique your interest.

Seeing other people cry can easily make you feel like you want to cry too.

You often make a backup plan for a backup plan.

You usually stay calm, even under a lot of pressure.

At social events, you rarely try to introduce yourself to new people and mostly talk to the ones you already know.

You prefer to completely finish one project before starting another.

You are very sentimental.

You like to use organizing tools like schedules and lists.

Even a small mistake can cause you to doubt your overall abilities and knowledge.

You feel comfortable just walking up to someone you find interesting and striking up a conversation.

You are not too interested in discussing various interpretations and analyses of creative works.

A.4 Investment game instructions

These instructions were provided on a web app embedded on Qualtrics.

You will be provided with virtual currency of $10,000. You may invest across three investment categories (high risk/return, medium risk/return, low risk/return). You will invest over 10 rounds (representing years).

In the next page, you will answer a short risk preference questionnaire and receive a recommended portfolio allocation (%). You may revise your allocation at any round. After each round, we will ask you how confident you feel about your choices. Good luck!

Important note: Your investment performance will increase your total compensation for participating in the study (10% increase for every $100,000).

A.5 Investment preference survey

The survey was administered through a web app embedded in Qualtrics.

1. When making a long-term investment, I plan to keep the money invested for

    • 1 to 2 years

    • 3 to 5 years

    • 6 to 8 years

    • More than 8 years

2. What's your main priority when investing?

    • Avoiding losing money

    • Preferring safety over returns

- Low risk over more money

- More money over low risk

- Maximizing profits, no matter what

3. How do you feel about taking risks, in general?

- I avoid risks. Better safe than sorry.

- Little risk is okay. I'm a bit cautious.

- I like to take risks sometimes.

- I love risk. I'm a bit of a gambler.

4. During market declines, I would sell portions of my riskier assets and invest the money in safe assets. (5-point Likert scale, 1=Strongly disagree to 5=Strongly agree)

A.6   Simulated investment game

The game was administered through a web app embedded in Qualtrics.

Emotion survey (after each round)

How confident you feel about your choices? (7-point Likert scale, 1=Not at all confident to 7=Very confident)

A.7   Evaluation of prophecies

Administered through a web app embedded in Qualtrics.

Please tell us what you think about each of the following statements (1=Strongly disagree, 7=Strongly agree).

- I find the prediction convincing.

- I can identify with the prediction.

- The source of the prediction is reliable.

- I find the prediction helpful.

- The prediction is accurate.

- The prediction describes me very well.

- I trust the source of the prediction.

- The prediction is useful for making future decisions.

A.8   Post-study questionnaire

Administered on Qualtrics.

Cognitive Reflection Test (CRT-2) [45]:

- If you're running a race and you pass the person in second place, what place are you in?

- A farmer had 15 sheep and all but 8 died. How many are left?

- Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?

- How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long?

Need for Cognition (NCS-6) [46]:

- I would prefer complex to simple problems.

- I like to have the responsibility of handling a situation that requires a lot of thinking.

- Thinking is not my idea of fun. (R)

- I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. (R)

- I really enjoy a task that involves coming up with new solutions to problems.

- I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

Revised Paranormal Belief Scale (R-PBS) [47]:
To what extent do you agree with the following (7-point Likert scale, 1=Strongly disagree, 7=Strongly agree):

- The soul continues to exist though the body may die. [religious beliefs]

- Black cats can bring bad luck. [superstition]

- Your mind or soul can leave your body and travel (astral projection). [spiritualism]

- Astrology is a way to accurately predict the future. [precognition]

- There is a devil. [religious beliefs]

- If you break a mirror, you will have bad luck. [superstition]

- During altered states, such as sleep or trances, the spirit can leave the body. [spiritualism]

- The horoscope accurately tells a person's future. [precognition]

- I believe in God. [religious beliefs]

- The number "13" is unlucky. [superstition]

- Reincarnation does occur. [spiritualism]

- Some psychics can accurately predict the future. [precognition]

- There is a heaven and a hell. [religious beliefs]

- It is possible to communicate with the dead. [spiritualism]

- Some people have an unexplained ability to predict the future. [precognition]

AI attitude scale (AIAS-4) [48]:

To what extent do you agree or disagree with the following statements? (1=Not at all, 10=Completely agree)

- I believe that AI will improve my life.

- I believe that AI will improve my work.

- I think I will use AI technology in the future.

- I think AI technology is positive for humanity.

Gullibility Scale [49]:

To what extent do you agree or disagree with the following statements? (7-point Likert scale, 1=Strongly disagree, 7=Strongly agree)

- (G4) I'm not that good at reading the signs that someone is trying to manipulate me. [insensitivity]

- (G5) I'm pretty poor at working out if someone is tricking me. [insensitivity]

- (G6) It usually takes me a while to "catch on" when someone is deceiving me. [insensitivity]

- (G1) I guess I am more gullible than the average person. [persuadability]

- (G9) My friends think I'm easily fooled. [persuadability]

- (G12) Overall, I'm pretty easily manipulated. [persuadability]

Big Five Personality Inventory (TIPI) [50]:
(7-point Likert scale, 1=Strongly disagree, 7=Strongly agree)
I see myself as:

- Extraverted, enthusiastic.

- Critical, quarrelsome.

- Dependable, self-disciplined.

- Anxious, easily upset.

- Open to new experiences, complex.

- Reserved, quiet.

- Sympathetic, warm.

- Disorganized, careless.

- Calm, emotionally stable.

- Conventional, uncreative.

A.9   Study debrief

This debrief and continued participation agreement form was administered on Qualtrics.

You are at the end of the study! We appreciate your participation.

Full disclosure: The predictions shown in the experiment were randomly generated and are unrelated to your actual behavior. The data collected for the predictions (birthdate, personality, etc.) are discarded and not used for the study.

If you have any concerns or questions, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787

Given this information, if you wish to withdraw your participation and data from this study, please select the option below. Otherwise, please click "Next" to continue.

[Checkbox] I would like to withdraw from the study. I understand that my responses will be deleted and I may not be eligible for any bonus payments based on my game performance.

Appendix B

Detailed Results and Model Diagnostics for Chapter 2

B.1 Multiple linear regression

Regression Equation

$$
\begin{aligned}
\text{ai\_overall\_score} = \beta_0 + \beta_1 (\text{zodiac\_overall\_score}) + \\
\beta_2 (\text{personality\_overall\_score}) + \cdots + \\
\beta_n (\text{prophecy\_group}) + \epsilon
\end{aligned}
\tag{B.1}
$$

Model results

Table B.1: Linear regression model summary

Statistic	Value
Residual standard error	0.7763 on 213 degrees of freedom
Multiple R-squared	0.7606
Adjusted R-squared	0.7337
F-statistic	28.2 on 24 and 213 DF
p-value	<2.2e-16

Table B.2: Linear regression residuals

Min	1Q	Median	3Q	Max
-2.52323	-0.40173	-0.03516	0.34271	3.05145

Table B.3: Detailed results of multiple linear regression

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.15	0.55	0.273	0.785	
zodiac_overall_score	0.31	0.05	6.690	0.000	***
personality_overall_score	0.46	0.05	9.585	<0.000	***
composite_score	0.05	0.04	1.187	0.237	
paranormal_score	0.00	0.00	0.970	0.333	
gullibility_score	0.00	0.01	0.455	0.650	
aias_score	0.02	0.01	2.302	0.022	*
interest_behavior	0.10	0.05	1.889	0.060	.
familiarity_ai	0.06	0.07	0.854	0.394	
Age	-0.00	0.00	-0.395	0.693	
big5_extraversion	-0.02	0.03	-0.486	0.627	
big5_openness	-0.03	0.05	-0.536	0.593	
big5_agreeableness	0.06	0.05	1.286	0.200	
big5_conscientiousness	0.01	0.05	0.114	0.909	
big5_emotional_stability	-0.01	0.04	-0.169	0.866	
educationBachelor	-0.29	0.23	-1.241	0.216	
educationDoctorate	0.26	0.61	0.430	0.668	
educationHigh school	-0.36	0.27	-1.325	0.187	
educationLess than high school	0.63	0.52	1.212	0.227	
educationMaster	-0.59	0.27	-2.206	0.029	*
educationProfessional degree	-0.65	0.52	-1.262	0.208	
educationSome college	-0.22	0.25	-0.915	0.361	
genderMale	0.06	0.12	0.478	0.633	
genderOther	0.03	0.31	0.086	0.932	
prophecy_grouprational	0.24	0.11	2.154	0.032	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

B.2 Multiple linear regression: Model diagnostics

Figure B.1: Residuals vs Fitted Plot
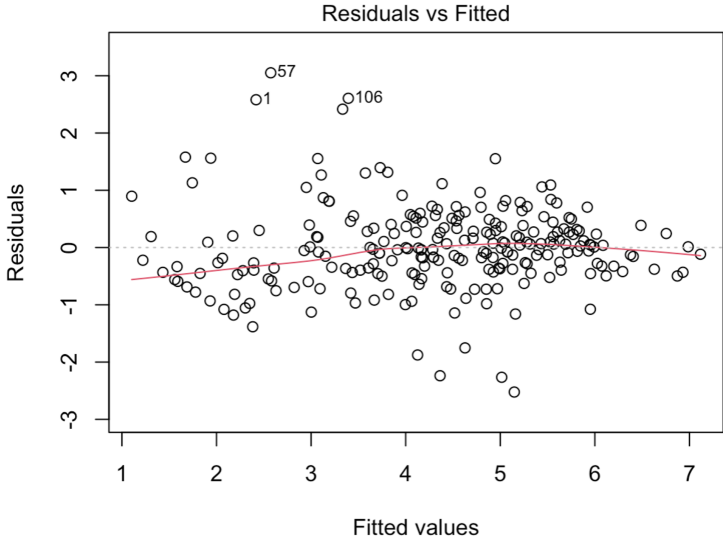
The residuals scatter randomly around the horizontal line (y=0), indicating that the linearity assumption holds. While there are a few outliers, the plot overall shows no strong patterns, suggesting the assumptions are reasonably met.

Figure B.2: Q-Q plot of the fitted model
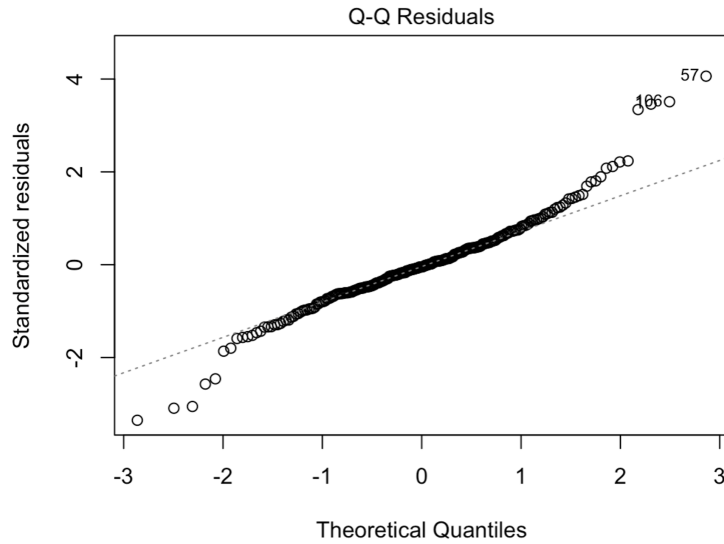
The points mostly follow the diagonal line, which suggests that the residuals are approximately normally distributed. There are some deviations at the tails, which may indicate some mild non-normality in the residuals.

Table B.4: VIF values for LM

Predictor	VIF
zodiac_overall_score	1.553862
personality_overall_score	1.570028
composite_score	1.181011
paranormal_score	1.234361
gullibility_score	1.213760
aias_score	1.257553
interest_behavior	1.143645
familiarity_ai	1.275727
Age	1.121698
big5_extraversion	1.173883
big5_openness	1.205639
big5_agreeableness	1.195784
big5_conscientiousness	1.226641
big5_emotional_stability	1.372926
education	1.062415
gender	1.111215
prophecy_group	1.109082

Table B.5: Durbin-Watson test results for LM

Lag	Autocorrelation	D-W Statistic	p-value
1	0.0676	1.8128	0.118

The residuals show no significant autocorrelation, as indicated by a low lag 1 autocorrelation (0.0676), a near-ideal Durbin-Watson statistic (1.8128), and a p-value (0.118) that suggests the independence of errors in the regression model is likely satisfied.

B.3 Mixed effects model: Results

This is the detailed results table from the model used in Section 2.2. All reported coefficients are based on the reference categories: Prophecy Source = AI, Prophecy Group = Positive, Subscale = Validity, Gender = Female, and Education = Bachelor.

Table B.6: Detailed mixed effects model results for Chapter 2

Predictor	Estimate	PValue	Lower	Upper
(Intercept)	5.10	0.000	4.75	5.44
subscalePersonalization	0.23	0.000	0.08	0.37
subscaleReliability	-0.91	0.000	-1.13	-0.68
subscaleUsefulness	-0.63	0.000	-0.83	-0.44
prophecy_sourceAstrology	-0.66	0.000	-0.93	-0.39
prophecy_sourcePersonality	0.06	0.620	-0.19	0.31
prophecy_groupNegative	-1.19	0.000	-1.53	-0.85
composite_score	0.13	0.060	-0.01	0.26
paranormal_score	0.02	0.000	0.01	0.03
aias_score	0.04	0.000	0.01	0.06
gullibility_score	-0.02	0.130	-0.04	0.01
big5_extraversion	-0.02	0.650	-0.12	0.08
big5_openness	-0.10	0.210	-0.25	0.05
big5_agreeableness	0.09	0.240	-0.06	0.24
big5_conscientiousness	-0.15	0.030	-0.30	-0.01
big5_emotional_stability	0.00	0.980	-0.13	0.13
interest_behavior	0.27	0.000	0.11	0.43
familiarity	-0.02	0.650	-0.11	0.07
Age	-0.02	0.010	-0.03	-0.00
educationAssociate	0.33	0.360	-0.38	1.03
educationDoctorate	0.91	0.310	-0.84	2.65
educationHigh school	0.39	0.170	-0.18	0.96
educationLess than high school	0.66	0.370	-0.80	2.12
educationMaster	0.19	0.460	-0.31	0.69
educationProfessional degree	-0.27	0.720	-1.71	1.18
educationSome college	0.02	0.910	-0.40	0.44
genderMale	-0.31	0.100	-0.69	0.06

Predictor	Estimate	PValue	Lower	Upper
genderOther	-0.06	0.910	-1.06	0.95
prophecy_sourceAstrology:prophecy_groupNegative	0.14	0.360	-0.17	0.45
prophecy_sourcePersonality:prophecy_groupNegative	-0.00	0.980	-0.29	0.28
prophecy_sourceAstrology:composite_score	-0.10	0.080	-0.21	0.01
prophecy_sourcePersonality:composite_score	-0.07	0.170	-0.18	0.03
prophecy_sourceAstrology:paranormal_score	0.01	0.000	0.01	0.02
prophecy_sourcePersonality:paranormal_score	-0.00	0.270	-0.01	0.00
prophecy_sourceAstrology:aias_score	-0.03	0.000	-0.05	-0.01
prophecy_sourcePersonality:aias_score	0.01	0.570	-0.01	0.02
prophecy_sourceAstrology:gullibility_score	-0.00	0.980	-0.02	0.02
prophecy_sourcePersonality:gullibility_score	-0.02	0.100	-0.04	0.00
prophecy_sourceAstrology:Age	-0.00	0.530	-0.02	0.01
prophecy_sourcePersonality:Age	0.01	0.380	-0.01	0.02
prophecy_sourceAstrology:genderMale	-0.06	0.740	-0.39	0.27
prophecy_sourcePersonality:genderMale	0.01	0.930	-0.30	0.32
prophecy_sourceAstrology:genderOther	-0.20	0.650	-1.08	0.68
prophecy_sourcePersonality:genderOther	0.20	0.630	-0.62	1.01
subscalePersonalization:prophecy_sourceAstrology	0.23	0.030	0.02	0.44
subscaleReliability:prophecy_sourceAstrology	-0.31	0.050	-0.62	-0.00
subscaleUsefulness:prophecy_sourceAstrology	-0.20	0.150	-0.47	0.07
subscalePersonalization:prophecy_sourcePersonality	0.08	0.440	-0.12	0.27
subscaleReliability:prophecy_sourcePersonality	0.18	0.240	-0.12	0.47
subscaleUsefulness:prophecy_sourcePersonality	0.10	0.440	-0.16	0.36
subscalePersonalization:prophecy_groupNegative	-0.24	0.000	-0.40	-0.08
subscaleReliability:prophecy_groupNegative	0.44	0.000	0.19	0.69
subscaleUsefulness:prophecy_groupNegative	0.53	0.000	0.31	0.74
subscalePersonalization:composite_score	-0.05	0.090	-0.11	0.01
subscaleReliability:composite_score	-0.11	0.020	-0.20	-0.02
subscaleUsefulness:composite_score	-0.12	0.000	-0.20	-0.04
subscalePersonalization:paranormal_score	0.00	0.430	-0.00	0.01
subscaleReliability:paranormal_score	0.01	0.020	0.00	0.01
subscaleUsefulness:paranormal_score	0.01	0.000	0.01	0.02
subscalePersonalization:aias_score	-0.00	0.940	-0.01	0.01

Predictor	Estimate	PValue	Lower	Upper
subscaleReliability:aias_score	0.03	0.000	0.02	0.05
subscaleUsefulness:aias_score	0.01	0.070	-0.00	0.03
subscalePersonalization:gullibility_score	0.01	0.020	0.00	0.02
subscaleReliability:gullibility_score	-0.01	0.480	-0.02	0.01
subscaleUsefulness:gullibility_score	-0.01	0.070	-0.03	0.00
subscalePersonalization:big5_extraversion	0.00	0.890	-0.03	0.03
subscaleReliability:big5_extraversion	0.04	0.090	-0.01	0.09
subscaleUsefulness:big5_extraversion	0.06	0.000	0.02	0.10
subscalePersonalization:big5_openness	-0.06	0.010	-0.10	-0.01
subscaleReliability:big5_openness	-0.02	0.640	-0.09	0.05
subscaleUsefulness:big5_openness	-0.02	0.550	-0.08	0.04
subscalePersonalization:big5_agreeableness	0.02	0.370	-0.02	0.07
subscaleReliability:big5_agreeableness	-0.01	0.690	-0.08	0.06
subscaleUsefulness:big5_agreeableness	0.02	0.470	-0.04	0.08
subscalePersonalization:big5_conscientiousness	0.01	0.770	-0.04	0.05
subscaleReliability:big5_conscientiousness	-0.00	0.890	-0.07	0.06
subscaleUsefulness:big5_conscientiousness	-0.00	0.920	-0.06	0.05
subscalePersonalization:big5_emotional_stability	0.02	0.390	-0.02	0.05
subscaleReliability:big5_emotional_stability	0.01	0.700	-0.05	0.07
subscaleUsefulness:big5_emotional_stability	0.00	0.870	-0.05	0.06
subscalePersonalization:interest_behavior	-0.06	0.020	-0.11	-0.01
subscaleReliability:interest_behavior	-0.07	0.070	-0.14	0.01
subscaleUsefulness:interest_behavior	0.02	0.480	-0.04	0.09
subscalePersonalization:familiarity	0.05	0.040	0.00	0.11
subscaleReliability:familiarity	0.02	0.690	-0.06	0.10
subscaleUsefulness:familiarity	0.01	0.780	-0.06	0.08
subscalePersonalization:Age	0.00	0.720	-0.01	0.01
subscaleReliability:Age	0.01	0.200	-0.00	0.02
subscaleUsefulness:Age	0.00	0.420	-0.01	0.01
subscalePersonalization:educationAssociate	-0.03	0.780	-0.25	0.18
subscaleReliability:educationAssociate	0.11	0.500	-0.22	0.44
subscaleUsefulness:educationAssociate	0.21	0.140	-0.07	0.49
subscalePersonalization:educationDoctorate	0.39	0.140	-0.13	0.91

Table B.6 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
subscaleReliability:educationDoctorate	-0.36	0.390	-1.17	0.46
subscaleUsefulness:educationDoctorate	-0.30	0.400	-0.99	0.40
subscalePersonalization:educationHigh school	-0.23	0.010	-0.40	-0.06
subscaleReliability:educationHigh school	-0.06	0.680	-0.32	0.21
subscaleUsefulness:educationHigh school	-0.01	0.960	-0.23	0.22
subscalePersonalization:educationLess than high school	-0.03	0.880	-0.49	0.43
subscaleReliability:educationLess than high school	-0.89	0.010	-1.56	-0.22
subscaleUsefulness:educationLess than high school	0.08	0.790	-0.51	0.67
subscalePersonalization:educationMaster	-0.03	0.680	-0.18	0.12
subscaleReliability:educationMaster	0.15	0.210	-0.08	0.39
subscaleUsefulness:educationMaster	0.05	0.630	-0.15	0.25
subscalePersonalization:educationProfessional degree	0.07	0.760	-0.37	0.51
subscaleReliability:educationProfessional degree	-0.35	0.310	-1.02	0.32
subscaleUsefulness:educationProfessional degree	-0.17	0.570	-0.74	0.41
subscalePersonalization:educationSome college	-0.04	0.510	-0.17	0.08
subscaleReliability:educationSome college	-0.05	0.650	-0.24	0.15
subscaleUsefulness:educationSome college	0.03	0.730	-0.14	0.20
subscalePersonalization:genderMale	0.11	0.210	-0.06	0.28
subscaleReliability:genderMale	0.41	0.000	0.14	0.68
subscaleUsefulness:genderMale	0.18	0.130	-0.05	0.42
subscalePersonalization:genderOther	0.38	0.100	-0.07	0.84
subscaleReliability:genderOther	-0.15	0.680	-0.87	0.57
subscaleUsefulness:genderOther	-0.14	0.670	-0.76	0.49
subscalePersonalization:prophecy_sourceAstrology: prophecy_groupNegative	-0.10	0.420	-0.33	0.14
subscaleReliability:prophecy_sourceAstrology: prophecy_groupNegative	0.11	0.530	-0.24	0.47
subscaleUsefulness:prophecy_sourceAstrology: prophecy_groupNegative	0.07	0.630	-0.23	0.38
subscalePersonalization:prophecy_sourcePersonality: prophecy_groupNegative	-0.07	0.550	-0.29	0.16
subscaleReliability:prophecy_sourcePersonality: prophecy_groupNegative	-0.09	0.630	-0.44	0.27

Predictor	Estimate	PValue	Lower	Upper
subscaleUsefulness:prophecy_sourcePersonality: prophecy_groupNegative	-0.19	0.230	-0.49	0.12
subscalePersonalization:prophecy_sourceAstrology:composite_score	0.04	0.400	-0.05	0.12
subscaleReliability:prophecy_sourceAstrology:composite_score	0.01	0.870	-0.12	0.14
subscaleUsefulness:prophecy_sourceAstrology:composite_score	0.06	0.320	-0.05	0.16
subscalePersonalization:prophecy_sourcePersonality:composite_score	0.04	0.360	-0.04	0.12
subscaleReliability:prophecy_sourcePersonality:composite_score	-0.03	0.640	-0.16	0.10
subscaleUsefulness:prophecy_sourcePersonality:composite_score	0.05	0.410	-0.06	0.16
subscalePersonalization:prophecy_sourceAstrology:paranormal_score	-0.00	0.060	-0.01	0.00
subscaleReliability:prophecy_sourceAstrology:paranormal_score	0.00	0.840	-0.01	0.01
subscaleUsefulness:prophecy_sourceAstrology:paranormal_score	0.00	0.530	-0.01	0.01
subscalePersonalization:prophecy_sourcePersonality:paranormal_score	-0.00	0.130	-0.01	0.00
subscaleReliability:prophecy_sourcePersonality:paranormal_score	0.00	0.830	-0.01	0.01
subscaleUsefulness:prophecy_sourcePersonality:paranormal_score	0.00	0.600	-0.01	0.01
subscalePersonalization:prophecy_sourceAstrology:aias_score	0.01	0.110	-0.00	0.03
subscaleReliability:prophecy_sourceAstrology:aias_score	-0.01	0.190	-0.04	0.01
subscaleUsefulness:prophecy_sourceAstrology:aias_score	0.00	0.810	-0.02	0.02
subscalePersonalization:prophecy_sourcePersonality:aias_score	0.00	0.690	-0.01	0.02
subscaleReliability:prophecy_sourcePersonality:aias_score	-0.02	0.110	-0.04	0.00
subscaleUsefulness:prophecy_sourcePersonality:aias_score	-0.01	0.330	-0.03	0.01
subscalePersonalization:prophecy_sourceAstrology:gullibility_score	0.00	0.950	-0.01	0.02
subscaleReliability:prophecy_sourceAstrology:gullibility_score	-0.01	0.560	-0.03	0.02
subscaleUsefulness:prophecy_sourceAstrology:gullibility_score	0.01	0.370	-0.01	0.03
subscalePersonalization:prophecy_sourcePersonality:gullibility_score	-0.02	0.030	-0.03	-0.00
subscaleReliability:prophecy_sourcePersonality:gullibility_score	-0.01	0.520	-0.03	0.02
subscaleUsefulness:prophecy_sourcePersonality:gullibility_score	0.01	0.610	-0.02	0.03
subscalePersonalization:prophecy_sourceAstrology:Age	-0.00	0.570	-0.01	0.01
subscaleReliability:prophecy_sourceAstrology:Age	0.01	0.420	-0.01	0.02
subscaleUsefulness:prophecy_sourceAstrology:Age	-0.00	0.830	-0.01	0.01
subscalePersonalization:prophecy_sourcePersonality:Age	-0.00	0.940	-0.01	0.01
subscaleReliability:prophecy_sourcePersonality:Age	-0.00	0.750	-0.02	0.01
subscaleUsefulness:prophecy_sourcePersonality:Age	-0.01	0.330	-0.02	0.01
subscalePersonalization:prophecy_sourceAstrology:genderMale	-0.21	0.100	-0.45	0.04

Table B.6 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
subscaleReliability:prophecy_sourceAstrology:genderMale	0.00	1.000	-0.38	0.38
subscaleUsefulness:prophecy_sourceAstrology:genderMale	0.08	0.630	-0.24	0.40
subscalePersonalization:prophecy_sourcePersonality:genderMale	0.17	0.150	-0.41	0.07
subscaleReliability:prophecy_sourcePersonality:genderMale	-0.29	0.140	-0.66	0.09
subscaleUsefulness:prophecy_sourcePersonality:genderMale	-0.08	0.650	-0.40	0.25
subscalePersonalization:prophecy_sourceAstrology:genderOther	0.06	0.870	-0.71	0.59
subscaleReliability:prophecy_sourceAstrology:genderOther	0.36	0.490	-0.65	1.37
subscaleUsefulness:prophecy_sourceAstrology:genderOther	0.48	0.270	-0.37	1.33
subscalePersonalization:prophecy_sourcePersonality:genderOther	0.12	0.720	-0.75	0.52
subscaleReliability:prophecy_sourcePersonality:genderOther	-0.17	0.730	-1.18	0.83
subscaleUsefulness:prophecy_sourcePersonality:genderOther	0.03	0.940	-0.84	0.91

Table B.7: Model fit statistics for mixed effects model

Statistic	Value
AIC	7949.6
BIC	9129.2
Log-Likelihood	-3774.8
Number of Observations	2856
Number of Groups	238

Table B.8: Standardized residuals

Min	Q1	Med	Q3	Max
-4.5316	-0.5581	-0.0065	0.5654	3.8666

Table B.9: Correlation structure specified at the lowest level

	1	2	3
2	-0.474		
3	-0.737	0.263	
4	0.131	-0.365	-0.596

This models the correlations within `prophecy_source` nested within `qualtrics_code`. This structure accounts for the dependency of observations grouped by `prophecy_source` within each `qualtrics_code`, allowing for different intercepts and slopes across these nested groups.

B.4 Mixed effects model: Model diagnostics for Section 2.2

Figure B.3: Q-Q plot of the fitted model
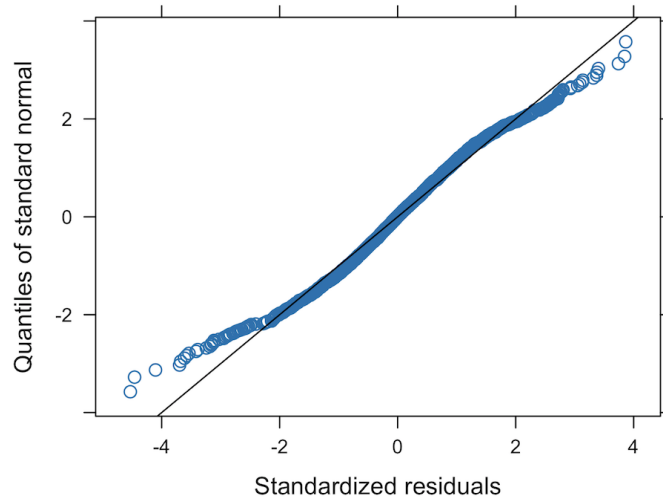
Normality of residuals was assessed using a Q-Q plot, using standardized residuals to account for both fixed and random effects. While there were slight deviations at the extremes, they were not severe and thus considered as normal deviations from real-world data.

Table B.10: Variance Inflation Factors (VIFs) values

Predictor	VIF
subscale	3.285782
prophecy_source	3.509332
prophecy_group	3.511723
composite_score	3.776002
paranormal_score	3.771773
aias_score	3.762766
gullibility_score	3.734846
big5_extraversion	2.268579
big5_openness	2.383225
big5_agreeableness	2.361957
big5_conscientiousness	2.420150

Table B.10 continued from previous page

Predictors	VIF
big5_emotional_stability	2.723458
interest_behavior	2.253065
familiarity	2.264958
Age	3.618152
education	2.107854
gender	3.674199
prophecy_source:prophecy_group	3.386351
prophecy_source:composite_score	2.745626
prophecy_source:paranormal_score	2.846607
prophecy_source:aias_score	2.758390
prophecy_source:gullibility_score	2.691879
prophecy_source:Age	2.697924
prophecy_source:gender	3.048713
subscale:prophecy_source	2.746200
subscale:prophecy_group	2.892868
subscale:composite_score	2.378733
subscale:paranormal_score	2.376068
subscale:aias_score	2.370394
subscale:gullibility_score	2.352806
subscale:big5_extraversion	1.429115
subscale:big5_openness	1.501338
subscale:big5_agreeableness	1.487939
subscale:big5_conscientiousness	1.524599
subscale:big5_emotional_stability	1.715671
subscale:interest_behavior	1.419342
subscale:familiarity	1.426834
subscale:Age	2.279293
subscale:education	1.391251
subscale:gender	2.612706
subscale:prophecy_source:prophecy_group	2.304924
subscale:prophecy_source:composite_score	1.729636
subscale:prophecy_source:paranormal_score	1.793250
subscale:prophecy_source:aias_score	1.737677

Table B.10 continued from previous page

| Predictors	VIF
subscale:prophecy_source:gullibility_score	1.695778
subscale:prophecy_source:Age	1.699586
subscale:prophecy_source:gender	1.989948

All VIF values after centering the continuous variables were under 5, suggesting that multicollinearity was not a concern. This method adjusts the GVIF for predictors with multiple degrees of freedom, making multicollinearity measures more comparable across variables.

Figure B.4: A residuals plot comparing standardized residuals and fitted values
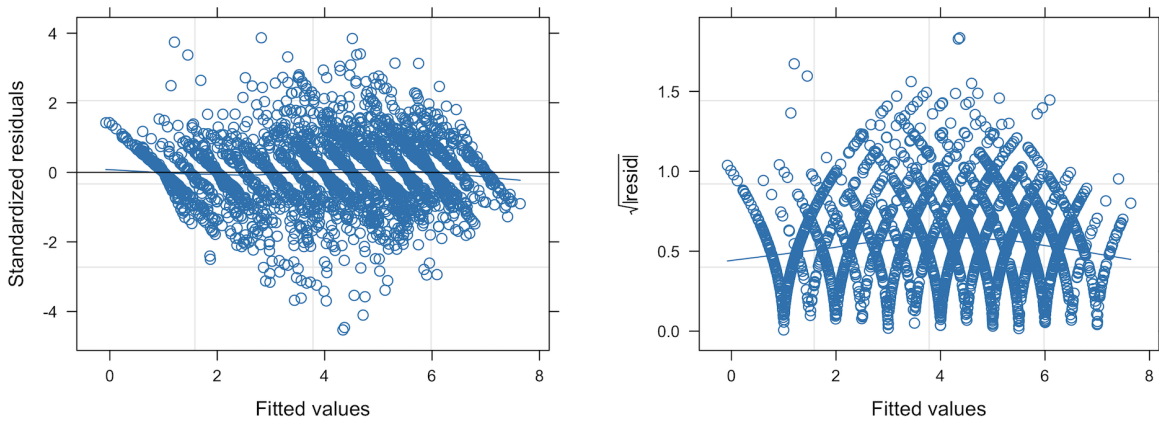
The plot showed a random spread around the horizontal line at zero, confirming that heteroscedasticity was successfully addressed using the variance function (varIdent) and that the linearity assumption is satisfied. One limitation to this diagnostics was the less continuous nature of the outcome variable (subscale score), which was multilevel with 13 levels (value of 1 to 7 with an increment of 0.5), which caused the residual plot to exhibit a slight diagonal pattern.

Figure B.5: Q-Q plot of the random effects
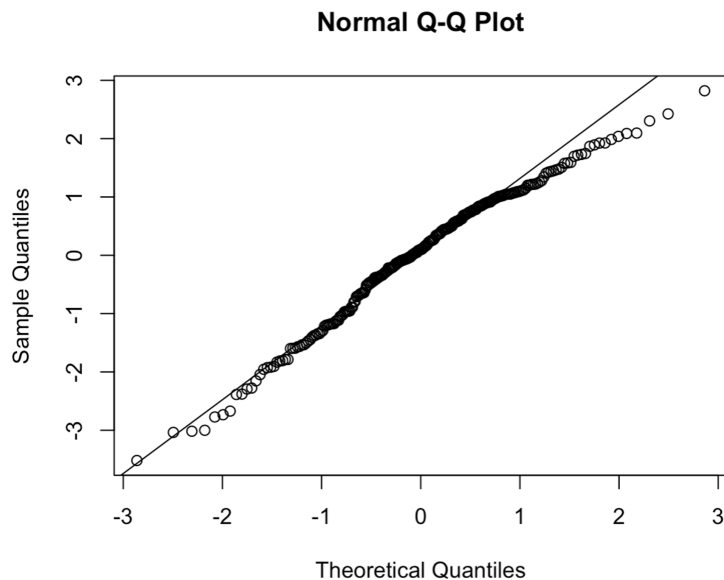
Normal Q-Q Plot

This indicates that the random effects are approximately normally distributed, as the points closely follow the theoretical normal line. However, there are slight deviations at the tails, suggesting minor departures from normality. These deviations were not deemed significant enough to invalidate the model assumptions.

Table B.11: Results of the Durbin-Watson test

Group	DW Statistic	p-value
Validity	1.6175	1.547e-07
Personalization	1.4457	6.027e-14
Reliability	1.3706	2.2e-16
Usefulness	1.4651	4.158e-13

We conducted the Durbin-Watson test within each subscale group to check for autocorrelation, finding significant positive autocorrelation in all groups. We initially used an autoregressive correlation structure (corAR1) to mitigate this, but model fit statistics indicated that an unstructured symmetric correlation structure (corSymm) provided a better fit, leading us to adopt corSymm as the final model. Although corSymm improved the model fit, it did not fully resolve the autocorrelation, which could limit the validity of our results and suggests a need for alternative approaches in future research.

Appendix C

Detailed Results and Model Diagnostics for Chapter 3

C.1 Model results

This is the detailed results table from the model used in Section 3.3. All reported coefficients are based on the reference categories: Prophecy Source = AI, Prophecy Group = Positive, Subscale = Validity, Gender = Female, and Education = Bachelor.

Table C.1: Detailed mixed effects model results for Chapter 3

Predictor	Estimate	PValue	Lower	Upper
(Intercept)	5.10	0.000	4.75	5.44
subscalePersonalization	0.23	0.000	0.08	0.37
subscaleReliability	-0.91	0.000	-1.13	-0.68
subscaleUsefulness	-0.63	0.000	-0.83	-0.44
prophecy_sourceAstrology	-0.66	0.000	-0.93	-0.39
prophecy_sourcePersonality	0.06	0.620	-0.19	0.31
prophecy_groupNegative	-1.19	0.000	-1.53	-0.85
composite_score	0.13	0.060	-0.01	0.26
paranormal_score	0.02	0.000	0.01	0.03
aias_score	0.04	0.000	0.01	0.06
gullibility_score	-0.02	0.130	-0.04	0.01
big5_extraversion	-0.02	0.650	-0.12	0.08
big5_openness	-0.10	0.210	-0.25	0.05
big5_agreeableness	0.09	0.240	-0.06	0.24

Table C.1 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
big5_conscientiousness	-0.15	0.030	-0.30	-0.01
big5_emotional_stability	0.00	0.980	-0.13	0.13
interest_behavior	0.27	0.000	0.11	0.43
familiarity	-0.02	0.650	-0.11	0.07
Age	-0.02	0.010	-0.03	-0.00
educationAssociate	0.33	0.360	-0.38	1.03
educationDoctorate	0.91	0.310	-0.84	2.65
educationHigh school	0.39	0.170	-0.18	0.96
educationLess than high school	0.66	0.370	-0.80	2.12
educationMaster	0.19	0.460	-0.31	0.69
educationProfessional degree	-0.27	0.720	-1.71	1.18
educationSome college	0.02	0.910	-0.40	0.44
genderMale	-0.31	0.100	-0.69	0.06
genderOther	-0.06	0.910	-1.06	0.95
prophecy_sourceAstrology:prophecy_groupNegative	0.14	0.360	-0.17	0.45
prophecy_sourcePersonality:prophecy_groupNegative	-0.00	0.980	-0.29	0.28
prophecy_sourceAstrology:composite_score	-0.10	0.080	-0.21	0.01
prophecy_sourcePersonality:composite_score	-0.07	0.170	-0.18	0.03
prophecy_sourceAstrology:paranormal_score	0.01	0.000	0.01	0.02
prophecy_sourcePersonality:paranormal_score	-0.00	0.270	-0.01	0.00
prophecy_sourceAstrology:aias_score	-0.03	0.000	-0.05	-0.01
prophecy_sourcePersonality:aias_score	0.01	0.570	-0.01	0.02
prophecy_sourceAstrology:gullibility_score	-0.00	0.980	-0.02	0.02
prophecy_sourcePersonality:gullibility_score	-0.02	0.100	-0.04	0.00
prophecy_sourceAstrology:Age	-0.00	0.530	-0.02	0.01
prophecy_sourcePersonality:Age	0.01	0.380	-0.01	0.02
prophecy_sourceAstrology:genderMale	-0.06	0.740	-0.39	0.27
prophecy_sourcePersonality:genderMale	0.01	0.930	-0.30	0.32
prophecy_sourceAstrology:genderOther	-0.20	0.650	-1.08	0.68
prophecy_sourcePersonality:genderOther	0.20	0.630	-0.62	1.01
subscalePersonalization:prophecy_sourceAstrology	0.23	0.030	0.02	0.44
subscaleReliability:prophecy_sourceAstrology	-0.31	0.050	-0.62	-0.00
subscaleUsefulness:prophecy_sourceAstrology	-0.20	0.150	-0.47	0.07

Table C.1 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
subscalePersonalization:prophecy_sourcePersonality	0.08	0.440	-0.12	0.27
subscaleReliability:prophecy_sourcePersonality	0.18	0.240	-0.12	0.47
subscaleUsefulness:prophecy_sourcePersonality	0.10	0.440	-0.16	0.36
subscalePersonalization:prophecy_groupNegative	-0.24	0.000	-0.40	-0.08
subscaleReliability:prophecy_groupNegative	0.44	0.000	0.19	0.69
subscaleUsefulness:prophecy_groupNegative	0.53	0.000	0.31	0.74
subscalePersonalization:composite_score	-0.05	0.090	-0.11	0.01
subscaleReliability:composite_score	-0.11	0.020	-0.20	-0.02
subscaleUsefulness:composite_score	-0.12	0.000	-0.20	-0.04
subscalePersonalization:paranormal_score	0.00	0.430	-0.00	0.01
subscaleReliability:paranormal_score	0.01	0.020	0.00	0.01
subscaleUsefulness:paranormal_score	0.01	0.000	0.01	0.02
subscalePersonalization:aias_score	-0.00	0.940	-0.01	0.01
subscaleReliability:aias_score	0.03	0.000	0.02	0.05
subscaleUsefulness:aias_score	0.01	0.070	-0.00	0.03
subscalePersonalization:gullibility_score	0.01	0.020	0.00	0.02
subscaleReliability:gullibility_score	-0.01	0.480	-0.02	0.01
subscaleUsefulness:gullibility_score	-0.01	0.070	-0.03	0.00
subscalePersonalization:big5_extraversion	0.00	0.890	-0.03	0.03
subscaleReliability:big5_extraversion	0.04	0.090	-0.01	0.09
subscaleUsefulness:big5_extraversion	0.06	0.000	0.02	0.10
subscalePersonalization:big5_openness	-0.06	0.010	-0.10	-0.01
subscaleReliability:big5_openness	-0.02	0.640	-0.09	0.05
subscaleUsefulness:big5_openness	-0.02	0.550	-0.08	0.04
subscalePersonalization:big5_agreeableness	0.02	0.370	-0.02	0.07
subscaleReliability:big5_agreeableness	-0.01	0.690	-0.08	0.06
subscaleUsefulness:big5_agreeableness	0.02	0.470	-0.04	0.08
subscalePersonalization:big5_conscientiousness	0.01	0.770	-0.04	0.05
subscaleReliability:big5_conscientiousness	-0.00	0.890	-0.07	0.06
subscaleUsefulness:big5_conscientiousness	-0.00	0.920	-0.06	0.05
subscalePersonalization:big5_emotional_stability	0.02	0.390	-0.02	0.05
subscaleReliability:big5_emotional_stability	0.01	0.700	-0.05	0.07
subscaleUsefulness:big5_emotional_stability	0.00	0.870	-0.05	0.06

Table C.1 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
subscalePersonalization:interest_behavior	-0.06	0.020	-0.11	-0.01
subscaleReliability:interest_behavior	-0.07	0.070	-0.14	0.01
subscaleUsefulness:interest_behavior	0.02	0.480	-0.04	0.09
subscalePersonalization:familiarity	0.05	0.040	0.00	0.11
subscaleReliability:familiarity	0.02	0.690	-0.06	0.10
subscaleUsefulness:familiarity	0.01	0.780	-0.06	0.08
subscalePersonalization:Age	0.00	0.720	-0.01	0.01
subscaleReliability:Age	0.01	0.200	-0.00	0.02
subscaleUsefulness:Age	0.00	0.420	-0.01	0.01
subscalePersonalization:educationAssociate	-0.03	0.780	-0.25	0.18
subscaleReliability:educationAssociate	0.11	0.500	-0.22	0.44
subscaleUsefulness:educationAssociate	0.21	0.140	-0.07	0.49
subscalePersonalization:educationDoctorate	0.39	0.140	-0.13	0.91
subscaleReliability:educationDoctorate	-0.36	0.390	-1.17	0.46
subscaleUsefulness:educationDoctorate	-0.30	0.400	-0.99	0.40
subscalePersonalization:educationHigh school	-0.23	0.010	-0.40	-0.06
subscaleReliability:educationHigh school	-0.06	0.680	-0.32	0.21
subscaleUsefulness:educationHigh school	-0.01	0.960	-0.23	0.22
subscalePersonalization:educationLess than high school	-0.03	0.880	-0.49	0.43
subscaleReliability:educationLess than high school	-0.89	0.010	-1.56	-0.22
subscaleUsefulness:educationLess than high school	0.08	0.790	-0.51	0.67
subscalePersonalization:educationMaster	-0.03	0.680	-0.18	0.12
subscaleReliability:educationMaster	0.15	0.210	-0.08	0.39
subscaleUsefulness:educationMaster	0.05	0.630	-0.15	0.25
subscalePersonalization:educationProfessional degree	0.07	0.760	-0.37	0.51
subscaleReliability:educationProfessional degree	-0.35	0.310	-1.02	0.32
subscaleUsefulness:educationProfessional degree	-0.17	0.570	-0.74	0.41
subscalePersonalization:educationSome college	-0.04	0.510	-0.17	0.08
subscaleReliability:educationSome college	-0.05	0.650	-0.24	0.15
subscaleUsefulness:educationSome college	0.03	0.730	-0.14	0.20
subscalePersonalization:genderMale	0.11	0.210	-0.06	0.28
subscaleReliability:genderMale	0.41	0.000	0.14	0.68
subscaleUsefulness:genderMale	0.18	0.130	-0.05	0.42

Table C.1 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
subscalePersonalization:genderOther	0.38	0.100	-0.07	0.84
subscaleReliability:genderOther	-0.15	0.680	-0.87	0.57
subscaleUsefulness:genderOther	-0.14	0.670	-0.76	0.49
subscalePersonalization:prophecy_sourceAstrology:prophecy_groupNegative	-0.10	0.420	-0.33	0.14
subscaleReliability:prophecy_sourceAstrology:prophecy_groupNegative	0.11	0.530	-0.24	0.47
subscaleUsefulness:prophecy_sourceAstrology:prophecy_groupNegative	0.07	0.630	-0.23	0.38
subscalePersonalization:prophecy_sourcePersonality:prophecy_groupNegative	-0.07	0.550	-0.29	0.16
subscaleReliability:prophecy_sourcePersonality:prophecy_groupNegative	-0.09	0.630	-0.44	0.27
subscaleUsefulness:prophecy_sourcePersonality:prophecy_groupNegative	-0.19	0.230	-0.49	0.12
subscalePersonalization:prophecy_sourceAstrology:composite_score	0.04	0.400	-0.05	0.12
subscaleReliability:prophecy_sourceAstrology:composite_score	0.01	0.870	-0.12	0.14
subscaleUsefulness:prophecy_sourceAstrology:composite_score	0.06	0.320	-0.05	0.16
subscalePersonalization:prophecy_sourcePersonality:composite_score	0.04	0.360	-0.04	0.12
subscaleReliability:prophecy_sourcePersonality:composite_score	-0.03	0.640	-0.16	0.10
subscaleUsefulness:prophecy_sourcePersonality:composite_score	0.05	0.410	-0.06	0.16
subscalePersonalization:prophecy_sourceAstrology:paranormal_score	-0.01	0.060	-0.01	0.00
subscaleReliability:prophecy_sourceAstrology:paranormal_score	0.00	0.840	-0.01	0.01
subscaleUsefulness:prophecy_sourceAstrology:paranormal_score	0.00	0.530	-0.01	0.01
subscalePersonalization:prophecy_sourcePersonality:paranormal_score	-0.01	0.130	-0.01	0.00
subscaleReliability:prophecy_sourcePersonality:paranormal_score	0.00	0.830	-0.01	0.01
subscaleUsefulness:prophecy_sourcePersonality:paranormal_score	0.00	0.600	-0.01	0.01
subscalePersonalization:prophecy_sourceAstrology:aias_score	0.01	0.110	-0.00	0.03
subscaleReliability:prophecy_sourceAstrology:aias_score	-0.01	0.190	-0.04	0.01
subscaleUsefulness:prophecy_sourceAstrology:aias_score	0.00	0.810	-0.02	0.02
subscalePersonalization:prophecy_sourcePersonality:aias_score	0.00	0.690	-0.01	0.02
subscaleReliability:prophecy_sourcePersonality:aias_score	-0.02	0.110	-0.04	0.00
subscaleUsefulness:prophecy_sourcePersonality:aias_score	-0.01	0.330	-0.03	0.01
subscalePersonalization:prophecy_sourceAstrology:gullibility_score	0.00	0.950	-0.01	0.02
subscaleReliability:prophecy_sourceAstrology:gullibility_score	-0.01	0.560	-0.03	0.02
subscaleUsefulness:prophecy_sourceAstrology:gullibility_score	0.01	0.370	-0.01	0.03
subscalePersonalization:prophecy_sourcePersonality:gullibility_score	-0.02	0.030	-0.03	-0.00
subscaleReliability:prophecy_sourcePersonality:gullibility_score	-0.01	0.520	-0.03	0.02
subscaleUsefulness:prophecy_sourcePersonality:gullibility_score	0.01	0.610	-0.02	0.03

Table C.1 continued from previous page

Predictor	Estimate	PValue	Lower	Upper
subscalePersonalization:prophecy_sourceAstrology:Age	-0.00	0.570	-0.01	0.01
subscaleReliability:prophecy_sourceAstrology:Age	0.01	0.420	-0.01	0.02
subscaleUsefulness:prophecy_sourceAstrology:Age	-0.00	0.830	-0.01	0.01
subscalePersonalization:prophecy_sourcePersonality:Age	-0.00	0.940	-0.01	0.01
subscaleReliability:prophecy_sourcePersonality:Age	-0.00	0.750	-0.02	0.01
subscaleUsefulness:prophecy_sourcePersonality:Age	-0.01	0.330	-0.02	0.01
subscalePersonalization:prophecy_sourceAstrology:genderMale	-0.21	0.100	-0.45	0.04
subscaleReliability:prophecy_sourceAstrology:genderMale	0.00	1.000	-0.38	0.38
subscaleUsefulness:prophecy_sourceAstrology:genderMale	0.08	0.630	-0.24	0.40
subscalePersonalization:prophecy_sourcePersonality:genderMale	-0.17	0.150	-0.41	0.07
subscaleReliability:prophecy_sourcePersonality:genderMale	-0.29	0.140	-0.66	0.09
subscaleUsefulness:prophecy_sourcePersonality:genderMale	-0.08	0.650	-0.40	0.25
subscalePersonalization:prophecy_sourceAstrology:genderOther	-0.06	0.870	-0.71	0.59
subscaleReliability:prophecy_sourceAstrology:genderOther	0.36	0.490	-0.65	1.37
subscaleUsefulness:prophecy_sourceAstrology:genderOther	0.48	0.270	-0.37	1.33
subscalePersonalization:prophecy_sourcePersonality:genderOther	-0.12	0.720	-0.75	0.52
subscaleReliability:prophecy_sourcePersonality:genderOther	-0.17	0.730	-1.18	0.83
subscaleUsefulness:prophecy_sourcePersonality:genderOther	0.03	0.940	-0.84	0.91

C.2 Model diagnostics

Table C.2: Variance Inflation Factors (VIFs) values

Predictor	VIF
subscale	3.209298
prophecy_group	4.055251
prophecy_source	2.829113
composite_score	3.265803
paranormal_score	3.141498
aias_score	3.319988
gullibility_score	3.214735
big5_extraversion	2.363846

Table C.2 continued from previous page

Predictors	VIF
big5_openness	2.398541
big5_agreeableness	2.420504
big5_conscientiousness	2.465333
big5_emotional_stability	2.754594
interest_behavior	2.283370
familiarity	2.120407
Age	2.989669
education	2.163890
gender	3.146641
prophecy_group:prophecy_source	3.365217
prophecy_group:composite_score	3.169956
prophecy_group:paranormal_score	3.159193
prophecy_group:aias_score	3.188012
prophecy_group:gullibility_score	3.077286
prophecy_group:Age	3.053249
prophecy_group:gender	3.342625
subscale:prophecy_group	3.340613
subscale:prophecy_source	2.213901
subscale:composite_score	2.057327
subscale:paranormal_score	1.979020
subscale:aias_score	2.091462
subscale:gullibility_score	2.025156
subscale:big5_extraversion	1.489130
subscale:big5_openness	1.510986
subscale:big5_agreeableness	1.524822
subscale:big5_conscientiousness	1.553062
subscale:big5_emotional_stability	1.735285
subscale:interest_behavior	1.438433
subscale:familiarity	1.335773
subscale:Age	1.883374
subscale:education	1.428237
subscale:gender	2.237562
subscale:prophecy_group:prophecy_source	2.290539

Table C.2 continued from previous page

Predictors	VIF
subscale:prophecy_group:composite_score	1.997536
subscale:prophecy_group:paranormal_score	1.992019
subscale:prophecy_group:aias_score	2.008683
subscale:prophecy_group:gullibility_score	1.940316
subscale:prophecy_group:Age	1.923432
subscale:prophecy_group:gender	2.211006

All VIF values after centering the continuous variables were under 5, suggesting that multicollinearity was not a concern. This method adjusts the GVIF for predictors with multiple degrees of freedom, making multicollinearity measures more comparable across variables.

Table C.3: Results of the Durbin-Watson test

Group	DW Statistic	p-value
Validity	1.6917	1.86e-05
Personalization	1.5129	3.593e-11
Reliability	1.4308	1.314e-14
Usefulness	1.5136	3.813e-11

We conducted the Durbin-Watson test within each subscale group to check for autocorrelation, finding significant positive autocorrelation in all groups. We initially used an autoregressive correlation structure (corAR1) to mitigate this, but model fit statistics indicated that an unstructured symmetric correlation structure (corSymm) provided a better fit, leading us to adopt corSymm as the final model. Although corSymm improved the model fit, it did not fully resolve the autocorrelation, which could limit the validity of our results and suggests a need for alternative approaches in future research.

Figure C.1: Q-Q plot of the fitted model
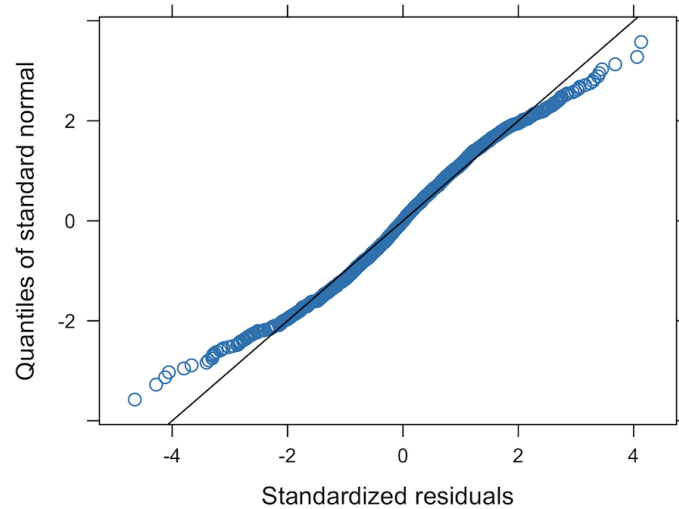
Normality of residuals was assessed using a Q-Q plot, using standardized residuals to account for both fixed and random effects. While there were slight deviations at the extremes, they were not severe and thus considered as normal deviations from real-world data.

Figure C.2: A residuals plot comparing standardized residuals and fitted values
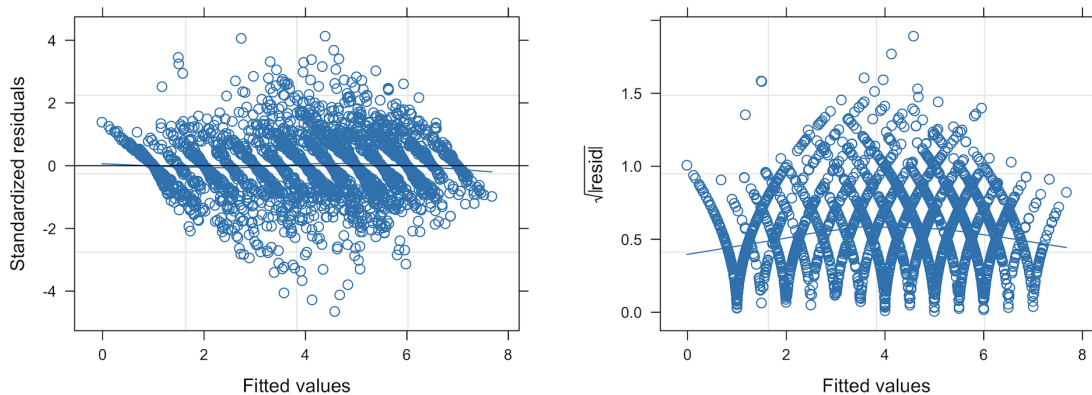
The plot showed a random spread around the horizontal line at zero, confirming that heteroscedasticity was successfully addressed using the variance function (varIdent) and that the linearity assumption is satisfied. One limitation to this diagnostics was the less continuous nature of the outcome variable (subscale score), which was multilevel with 13 levels (value of 1 to 7 with an increment of 0.5), which caused the residual plot to exhibit a slight diagonal pattern.

Figure C.3: Q-Q plots of the random effects
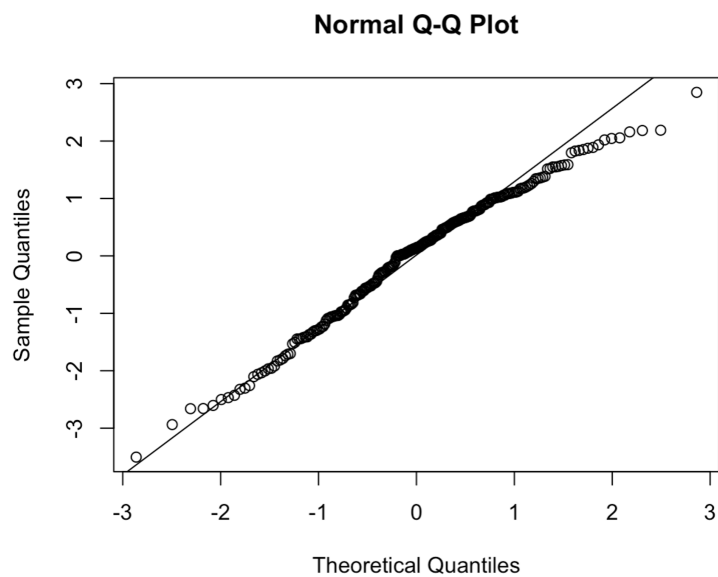
Normal Q-Q Plot

This indicates that the random effects are approximately normally distributed, as the points closely follow the theoretical normal line. However, there are slight deviations at the tails, suggesting minor departures from normality. These deviations were not deemed significant enough to invalidate the model assumptions.

Appendix D

Simulated investment game design: Additional details

D.1 Porfolio allocation scenario

A dynamic allocation strategy was chosen as the default user action scenario for the simulated investment game. In contrast to a fixed allocation approach, a dynamic allocation strategy allows users to adjust their portfolio based on outcomes over time. In this scenario, investors react to market changes, new information, or shifts in their investment thesis by modifying their asset allocation. Dynamic allocation is particularly suitable for simulating tactical asset allocation, sector rotation, or active management strategies. It mirrors real-world practices employed by both funds and individual investors, especially in volatile markets.

D.2 Market scenario

The investment game scenario was set to be consistent for everyone to control the condition. The sequence of user actions for each round included:

1. User sees the investment news.

2. User sees the market change.

3. User edits allocation based for that round.

D.3 Investment News

Table D.1: Overview of the news sentiment, expected user behavior, and market scenario in the simulated investment game

Round	News sentiment	Expected behavior	Market Scenario
1	n/a	n/a	Bull
2	Positive	Increase high risk; Decrease low risk	Bull
3	Positive	Increase high risk; Decrease low risk	Bull
4	Positive	Increase high risk; Decrease low risk	Bear
5	Negative	Decrease high risk; Increase low risk	Bear
6	Negative	Decrease high risk; Increase low risk	Bull
7	Positive	Increase high risk; Decrease low risk	Bull
8	Positive	Increase high risk; Decrease low risk	Bear
9	Negative	Decrease high risk; Increase low risk	Bear
10	Negative	Decrease high risk; Increase low risk	Bear

Table D.2: Example market news that was presented to users at each round

Positive sentiment	Negative sentiment
The high risk high return category is expected to generate significant growth in the next few months.	Due to an unexpected market downturn, some projects in the high risk high return category appear to be much riskier; some investors are selling.
Financial experts project a strong bull market for the high risk high return category for the next few months.	Due to continuing volatile market conditions, the performance of high risk high return investment category looks likely to plummet.
Warren Buffett reveals that he has just invested $1 million in a company in the high risk high return category.	A leading company that has attracted huge amounts of high risk high return investment looks shaky as their stock price takes a dive after the recession announcement.

Appendix E

Code listing

The full code is available on GitHub (https://github.com/mitmedialab/ai-prophecy).

E.1 Multiple Linear Regression (Chapter 2)

```
1 model <- lm(ai_overall_score ~ zodiac_overall_score + personality_
     overall_score + composite_score + paranormal_score + gullibility_
     score + interest_behavior  + Age + education + gender + prophecy_
     group, data = wide_df)
```

E.2 Mixed Effects Model (Chapter 2)

```
1 model <- lme(fixed = subscale_score ~ subscale * (prophecy_source *
     prophecy_group +
2               prophecy_source * composite_score +
3               prophecy_source * paranormal_score +
4               prophecy_source * aias_score +
5               prophecy_source * gullibility_score +
6               big5_extraversion +
7               big5_openness +
8               big5_agreeableness +
9               big5_conscientiousness +
10              big5_emotional_stability +
11              interest_behavior +
12              familiarity +
```

```
13                    prophecy_source * Age +
14                    education +
15                    prophecy_source * gender),
16      random = ~ prophecy_source | qualtrics_code,
17      correlation = corSymm(form = ~ 1 | qualtrics_code/prophecy_source),
18      weights = varIdent(form = ~1 | subscale * prophecy_source *
            prophecy_group),
19      data = long_df_centered,
20      na.action = na.exclude,
21      control = list(msMaxIter = 1000, msMaxEval = 1000, opt="optim"))
```

E.3 Mixed Effects Model (Chapter 3)

```
1  model <- lme(fixed = subscale_score ~ subscale * (prophecy_group *
      prophecy_source +
2                   prophecy_group * composite_score +
3                   prophecy_group * paranormal_score +
4                   prophecy_group * aias_score +
5                   prophecy_group * gullibility_score +
6                   big5_extraversion +
7                   big5_openness +
8                   big5_agreeableness +
9                   big5_conscientiousness +
10                  big5_emotional_stability +
11                  interest_behavior +
12                  familiarity +
13                  prophecy_group * Age +
14                  education +
15                  prophecy_group * gender),
16      random = ~ prophecy_source | qualtrics_code,
17      correlation = corSymm(form = ~1 | qualtrics_code/prophecy_source),
18      weights = varIdent(form = ~1 | subscale * prophecy_source *
            prophecy_group),
19      data = long_df_centered,
20      na.action = na.exclude,
21      control = list(msMaxIter = 1000, msMaxEval = 1000, opt="optim"))
```

References

[1] A. Wilson. "Techno-Optimism and Rational Superstition". In: *Techné: Research in Philosophy and Technology* 21.2 (2017), pp. 342–362. DOI: 10.5840/techne201711977. URL: https://www.pdcnet.org/techne/content/techne_2017_0021_42769_0342_0362.

[2] R. Evans. *The Cult of AI*. Rolling Stone, Jan. 2024. URL: https://www.rollingstone.com/culture/culture-features/ai-companies-advocates-cult-1234954528/.

[3] S. Sigal. "Silicon Valley's AI religion: transhumanism, longtermism, and effective altruism". In: *Vox* (Sept. 7, 2023). URL: https://www.vox.com/the-highlight/23779413/silicon-valleys-ai-religion-transhumanism-longtermism-ea.

[4] J. Danaher. "Techno-optimism: an Analysis, an Evaluation and a Modest Defence". In: *Philosophy & Technology* 35.2 (June 6, 2022), p. 54. ISSN: 2210-5441. DOI: 10.1007/s13347-022-00550-2. URL: https://doi.org/10.1007/s13347-022-00550-2 (visited on 07/10/2024).

[5] P. Königs. "What is Techno-Optimism?" In: *Philosophy & Technology* 35.3 (July 13, 2022), p. 63. ISSN: 2210-5441. DOI: 10.1007/s13347-022-00555-x. URL: https://doi.org/10.1007/s13347-022-00555-x (visited on 07/10/2024).

[6] P. Brauner, A. Hick, R. Philipsen, and M. Ziefle. "What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI". In: *Frontiers in Computer Science* 5 (2023), p. 1113903. DOI: 10.3389/fcomp.2023.1113903. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1113903/full.

[7] L. Sartori and G. Bocca. "Minding the gap(s): public perceptions of AI and socio-technical imaginaries". en. In: *AI & SOCIETY* 38.2 (Apr. 2023), pp. 443–458. ISSN: 1435-5655. DOI: 10.1007/s00146-022-01422-1. URL: https://doi.org/10.1007/s00146-022-01422-1 (visited on 07/28/2024).

[8] L. Nikolić. "ECS-Ecrea Early Career Scholar Prize winner - An astrological genealogy of artificial intelligence: From 'pseudo-sciences' of divination to sciences of prediction". In: *European Journal of Cultural Studies* 26.2 (Apr. 1, 2023). Publisher: SAGE Publications

Ltd, pp. 131–146. ISSN: 1367-5494. DOI: 10.1177/13675494231164874. URL: https://doi.org/10.1177/13675494231164874 (visited on 07/09/2024).

[9]   C. Lazaro. "Le pouvoir « divinatoire » des algorithmes : de la prédiction à la préemption du futur". In: *Anthropologie et Sociétés* 42.2 (2018), pp. 127–150.

[10] H. Steyerl. "A Sea of Data: Pattern Recognition and Corporate Animism (Forked Version)". In: in collab. with Mediarep.Org, C. Apprich, W. H. K. Chun, F. Cramer, and H. Steyerl. meson press, 2018. ISBN: 978-3-95796-145-7. DOI: 10.25969/MEDIAREP/12348. URL: https://mediarep.org/handle/doc/13257 (visited on 07/10/2024).

[11] F. Pethig and J. Kroenung. "Biased Humans, (Un)Biased Algorithms?" In: *Journal of Business Ethics* 183.3 (2023), pp. 637–652. DOI: 10.1007/s10551-022-05071-8. URL: https://link.springer.com/article/10.1007/s10551-022-05071-8.

[12] N. Kordzadeh and M. Ghasemaghaei. "Algorithmic bias: review, synthesis, and future research directions". In: *European Journal of Information Systems* 31.3 (2022), pp. 388–409. DOI: 10.1080/0960085X.2021.1927212. URL: https://www.tandfonline.com/doi/full/10.1080/0960085X.2021.1927212.

[13] G. M. Johnson. "Algorithmic bias: on the implicit biases of social technology". In: *Synthese* 199.1-2 (2021), pp. 1–23. DOI: 10.1007/s11229-020-02696-y. URL: https://link.springer.com/article/10.1007/s11229-020-02696-y.

[14] W. J. von Eschenbach. "Transparency and the Black Box Problem: Why We Do Not Trust AI". In: *Philosophy & Technology* 34.4 (2021), pp. 1607–1622. DOI: 10.1007/s13347-021-00477-0. URL: https://link.springer.com/article/10.1007/s13347-021-00477-0.

[15] A. B. Arrieta et al. "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence". In: *Cognitive Computation* 16 (2024), pp. 45–74. DOI: 10.1007/s12559-023-10179-8. URL: https://link.springer.com/article/10.1007/s12559-023-10179-8.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2021), p. 18. DOI: 10.3390/e23010018. URL: https://pubmed.ncbi.nlm.nih.gov/33375658/.

[17] D. Castelvecchi. "Can we open the black box of AI?" In: *Nature* 538 (2016), pp. 20–23. DOI: 10.1038/538020a. URL: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[18] J. M. Logg, J. A. Minson, and D. A. Moore. "Algorithm appreciation: People prefer algorithmic to human judgment". In: *Organizational Behavior and Human Decision Processes* 151 (Mar. 1, 2019), pp. 90–103. ISSN: 0749-5978. DOI: 10.1016/j.obhdp.2018.12.005. URL: https://www.sciencedirect.com/science/article/pii/S0749597818303388 (visited on 07/08/2024).

[19] H. Mahmud, A. K. M. N. Islam, X. (Luo, and P. Mikalef. "Decoding algorithm appreciation: Unveiling the impact of familiarity with algorithms, tasks, and algorithm performance". In: *Decision Support Systems* 179 (Apr. 1, 2024), p. 114168. ISSN: 0167-9236. DOI: 10.1016/j.dss.2024.114168. URL: https://www.sciencedirect.com/science/article/pii/S0167923624000010 (visited on 07/08/2024).

[20] D. Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011. ISBN: 978-0-374-27563-1.

[21] J. Evans and K. Frankish. *In two minds: Dual processes and beyond*. Oxford University Press, Jan. 2009. ISBN: 978-0-19-923016-7. DOI: 10.1093/acprof:oso/9780199230167.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199230167.001.0001.

[22] S. Dale. "Heuristics and biases: The science of decision-making". en. In: *Business Information Review* 32.2 (June 2015). Publisher: SAGE Publications Ltd, pp. 93–99. ISSN: 0266-3821. DOI: 10.1177/0266382115592536. URL: https://doi.org/10.1177/0266382115592536 (visited on 07/28/2024).

[23] M. Steyvers and A. Kumar. "Three Challenges for AI-Assisted Decision-Making". en. In: *Perspectives on Psychological Science* (July 2023). Publisher: SAGE Publications Inc, p. 17456916231181102 ISSN: 1745-6916. DOI: 10.1177/17456916231181102. URL: https://doi.org/10.1177/17456916231181102 (visited on 08/03/2024).

[24] E. Lee, P. Pataranutaporn, J. Amores, and P. Maes. *Super-intelligence or Superstition? Exploring Psychological Factors Underlying Unwarranted Belief in AI Predictions*. 2024. arXiv: 2408.06602 [cs.HC]. URL: https://arxiv.org/abs/2408.06602.

[25] S. Page. *Why millennials and Gen Z are turning to astrology*. 2023. URL: https://www.washingtonpost.com/lifestyle/2023/06/13/astrology-millennials-gen-z-science/.

[26] A. R. Farrar. *Why is astrology making a twenty-first-century comeback?* 2022. URL: https://www.panmacmillan.com/blogs/lifestyle-wellbeing/the-popularity-of-astrology.

[27] J. H. McGrew and R. M. McFall. "A Scientific Inquiry Into the Validity of Astrology". In: *Journal of Scientific Exploration* 4.1 (1990). Publisher: Society for Scientific Exploration, pp. 75–83. URL: https://www.scientificexploration.org/docs/4/jse_04_1_mcgrew.pdf.

[28] H. Nyborg. "The relationship between date of birth and individual differences in personality and general intelligence: A large-scale study". In: *Personality and Individual Differences* 40 (2006). Publisher: Pergamon Press, pp. 1349–1362. ISSN: 0191-8869. DOI: 10.1016/j.paid.2005.11.017.

[29] R. Safier. *Written in the Stars? Millennials Most Likely to Make Financial Decisions Based on Their Horoscopes*. Ed. by D. Shepard and X. Martinez-White. Dec. 2021. URL: https://www.lendingtree.com/debt-consolidation/zodiac-finances-survey/#Keyfindings.

[30] J. Wester, S. de Jong, H. Pohl, and N. van Berkel. "Exploring people's perceptions of LLM-generated advice". In: *Computers in Human Behavior: Artificial Humans* 2.2 (Aug. 1, 2024), p. 100072. ISSN: 2949-8821. DOI: 10.1016/j.chbah.2024.100072. URL: https://www.sciencedirect.com/science/article/pii/S294988212400032X (visited on 07/05/2024).

[31] R. Riedl. "Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions". In: *Electronic Markets* 32.4 (Dec. 2022), pp. 2021–2051. ISSN: 1019-6781, 1422-8890. DOI: 10.1007/s12525-022-00594-4. URL: https://link.springer.com/10.1007/s12525-022-00594-4 (visited on 04/10/2024).

[32] D. A. Bensley. "Critical Thinking, Intelligence, and Unsubstantiated Beliefs: An Integrative Review". In: *Journal of Intelligence* 11.11 (Oct. 30, 2023), p. 207. ISSN: 2079-3200. DOI: 10.3390/jintelligence11110207. URL: https://www.mdpi.com/2079-3200/11/11/207 (visited on 03/30/2024).

[33] E. Ballová Mikušková and V. Čavojová. "The Effect of Analytic Cognitive Style on Credulity". In: *Frontiers in Psychology* 11 (Oct. 15, 2020), p. 584424. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.584424. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.584424/full (visited on 02/23/2024).

[34] G. Pennycook, J. A. Fugelsang, and D. J. Koehler. "Everyday Consequences of Analytic Thinking". In: (2015).

[35] M. N. Torres, I. Barberia, and J. Rodríguez-Ferreiro. "A validation of the Pseudoscience Endorsement Scale and assessment of the cognitive correlates of pseudoscientific beliefs". In: *Humanities and Social Sciences Communications* 10.1 (Apr. 22, 2023), p. 176. ISSN: 2662-9992. DOI: 10.1057/s41599-023-01681-3. URL: https://www.nature.com/articles/s41599-023-01681-3 (visited on 02/20/2024).

[36] G. Pennycook, J. A. Cheyne, P. Seli, D. J. Koehler, and J. A. Fugelsang. "Analytic cognitive style predicts religious and paranormal belief". In: *Cognition* 123.3 (2012), pp. 335–346. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2012.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0010027712000534.

[37] M. Mosleh, G. Pennycook, A. A. Arechar, and D. G. Rand. "Cognitive reflection correlates with behavior on Twitter". In: *Nature Communications* 12.1 (Feb. 10, 2021). Publisher: Nature Publishing Group, p. 921. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20043-0. URL: https://www.nature.com/articles/s41467-020-20043-0 (visited on 07/02/2024).

[38] D. A. Stecula and M. Pickup. "Social Media, Cognitive Reflection, and Conspiracy Beliefs". In: *Frontiers in Political Science* 3 (June 8, 2021), p. 647957. ISSN: 2673-3145. DOI: 10.3389/fpos.2021.647957. URL: https://www.frontiersin.org/articles/10.3389/fpos.2021.647957/full (visited on 03/30/2024).

[39] I. B. Myers. *The Myers-Briggs Type Indicator: Manual (1962)*. The Myers-Briggs Type Indicator: Manual (1962). Pages: ii, 110. Palo Alto, CA, US: Consulting Psychologists Press, 1962. ii, 110. DOI: 10.1037/14404-000.

[40] F. D'Acunto, N. Prabhala, and A. G. Rossi. "The Promises and Pitfalls of Robo-Advising". In: *The Review of Financial Studies* 32.5 (May 1, 2019), pp. 1983–2020. ISSN: 0893-9454, 1465-7368. DOI: 10.1093/rfs/hhz014. URL: https://academic.oup.com/rfs/article/32/5/1983/5427774 (visited on 08/03/2024).

[41] H. Wang, J. Zhang, L. Wang, and S. Liu. "Emotion and Investment Returns: Situation and Personality as Moderators in a Stock Market". In: *Social Behavior and Personality: an international journal* 42.4 (May 15, 2014), pp. 561–569. ISSN: 0301-2212. DOI: 10.2224/sbp.2014.42.4.561. URL: https://www.ingentaconnect.com/content/10.2224/sbp.2014.42.4.561 (visited on 03/14/2024).

[42] B. Shiv, G. Loewenstein, A. Bechara, H. Damasio, and A. R. Damasio. "Investment Behavior and the Negative Side of Emotion". In: 16.6 (2005).

[43] S. Villa, T. Kosch, F. Grelka, A. Schmidt, and R. Welsch. "The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior". In: *Computers in Human Behavior* 146 (Sept. 2023), p. 107787. ISSN: 07475632. DOI: 10.1016/j.chb.2023.107787. URL: https://linkinghub.elsevier.com/retrieve/pii/S0747563223001383 (visited on 12/03/2023).

[44] T. Kosch, R. Welsch, L. Chuang, and A. Schmidt. "The Placebo Effect of Artificial Intelligence in Human–Computer Interaction". In: *ACM Transactions on Computer-Human Interaction* 29.6 (Dec. 31, 2022), pp. 1–32. ISSN: 1073-0516, 1557-7325. DOI: 10.1145/3529225. URL: https://dl.acm.org/doi/10.1145/3529225 (visited on 12/03/2023).

[45] K. S. Thomson and D. M. Oppenheimer. "Investigating an alternate form of the cognitive reflection test". In: *Judgment and Decision Making* 11.1 (Jan. 2016), pp. 99–113. ISSN: 1930-2975. DOI: 10.1017/S1930297500007622. URL: https://www.cambridge.org/core/product/identifier/S1930297500007622/type/journal_article (visited on 02/23/2024).

[46] G. Lins De Holanda Coelho, P. H. P. Hanel, and L. J. Wolf. "The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version". In: *Assessment* 27.8 (Dec. 2020), pp. 1870–1885. ISSN: 1073-1911, 1552-3489. DOI: 10.1177/1073191118793208. URL: http://journals.sagepub.com/doi/10.1177/1073191118793208 (visited on 02/23/2024).

[47] J. J. Tobacyk. "A Revised Paranormal Belief Scale". In: *International Journal of Transpersonal Studies* 23.1 (Jan. 1, 2004), pp. 94–98. ISSN: 13210122, 19423241. DOI: 10.24972/ijts.2004.23.1.94. URL: https://digitalcommons.ciis.edu/cgi/viewcontent.cgi?article=1235&context=ijts-transpersonalstudies (visited on 03/28/2024).

[48] S. Grassini. "Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence". In: *Frontiers in Psychology* 14 (July 24, 2023), p. 1191628. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2023.1191628. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1191628/full (visited on 03/27/2024).

[49] A. K. Teunisse, T. I. Case, J. Fitness, and N. Sweller. "I Should Have Known Better: Development of a Self-Report Measure of Gullibility". In: *Personality and Social Psychology Bulletin* 46.3 (Mar. 2020), pp. 408–423. ISSN: 0146-1672, 1552-7433. DOI: 10.1177/0146167219858641. URL: http://journals.sagepub.com/doi/10.1177/0146167219858641 (visited on 03/28/2024).

[50] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. "A very brief measure of the Big-Five personality domains". In: *Journal of Research in Personality* 37.6 (Dec. 1, 2003), pp. 504–528. ISSN: 0092-6566. DOI: 10.1016/S0092-6566(03)00046-1. URL: https://www.sciencedirect.com/science/article/pii/S0092656603000461 (visited on 07/06/2024).

[51] R. E. Petty and J. T. Cacioppo. "The Elaboration Likelihood Model of Persuasion". en. In: *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Ed. by R. E. Petty and J. T. Cacioppo. New York, NY: Springer, 1986, pp. 1–24. ISBN: 978-1-4612-4964-1. DOI: 10.1007/978-1-4612-4964-1_1. URL: https://doi.org/10.1007/978-1-4612-4964-1_1 (visited on 08/11/2024).

[52] F. D. Davis. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In: *MIS Quarterly* 13.3 (1989). Publisher: Management Information Systems Research Center, University of Minnesota, pp. 319–340. ISSN: 0276-7783. DOI: 10.2307/249008. URL: https://www.jstor.org.libproxy.mit.edu/stable/249008 (visited on 08/11/2024).

[53] E. Maslowska, E. G. Smit, and B. van den Putte. "It Is All in the Name: A Study of Consumers' Responses to Personalized Communication". In: *Journal of Interactive Advertising* 16.1 (Jan. 2016). Publisher: Routledge _eprint: https://doi.org/10.1080/15252019.2016.1161568, pp. 74–85. ISSN: null. DOI: 10.1080/15252019.2016.1161568. URL: https://doi.org/10.1080/15252019.2016.1161568 (visited on 08/11/2024).

[54] S. Frederick. "Cognitive Reflection and Decision Making". In: *Journal of Economic Perspectives* 19.4 (Dec. 2005), pp. 25–42. ISSN: 0895-3309. DOI: 10.1257/089533005775196732. URL: https://www.aeaweb.org/articles?id=10.1257/089533005775196732 (visited on 07/02/2024).

[55] J. T. Cacioppo and R. E. Petty. "The need for cognition". In: *Journal of Personality and Social Psychology* 42 (1982). Publisher: American Psychological Association, pp. 116–131. DOI: 10.1037/0022-3514.42.1.116.

[56] G. Pennycook, J. A. Cheyne, D. J. Koehler, and J. A. Fugelsang. "Is the cognitive reflection test a measure of both reflection and intuition?" In: *Behavior Research Methods* 48.1 (Mar. 2016), pp. 341–348. ISSN: 1554-3528. DOI: 10.3758/s13428-015-0576-1. URL: http://link.springer.com/10.3758/s13428-015-0576-1 (visited on 02/23/2024).

[57] A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. London: Sage Publications, 2012. ISBN: 978-1-4462-0046-9.

[58] J. H. Andrew Gelman and M. Yajima. "Why We (Usually) Don't Have to Worry About Multiple Comparisons". In: *Journal of Research on Educational Effectiveness* 5.2 (2012), pp. 189–211. DOI: 10.1080/19345747.2011.618213.

[59] P. K. Kahr, G. Rooks, M. C. Willemsen, and C. C. P. Snijders. "Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions." In: *ACM Trans. Interact. Intell. Syst.* (Aug. 2024). Just Accepted. ISSN: 2160-6455. DOI: 10.1145/3686164. URL: https://dl.acm.org/doi/10.1145/3686164 (visited on 08/09/2024).

[60] A. Klingbeil, C. Grützner, and P. Schreck. "Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI". In: *Computers in Human Behavior* 160 (Nov. 2024), p. 108352. ISSN: 0747-5632. DOI: 10.1016/j.chb.2024.108352. URL: https://www.sciencedirect.com/science/article/pii/S0747563224002206 (visited on 08/09/2024).

[61] E. K. Chiou and J. D. Lee. "Trusting Automation: Designing for Responsivity and Resilience". eng. In: *Human Factors* 65.1 (Feb. 2023), pp. 137–165. ISSN: 1547-8181. DOI: 10.1177/00187208211009995.

[62] A. Küper and N. Krämer. "Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI". In: *International Journal of Human–Computer Interaction* 0.0 (2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2024.2348216, pp. 1–17. ISSN: 1044-7318. DOI: 10.1080/10447318.2024.2348216. URL: https://doi.org/10.1080/10447318.2024.2348216 (visited on 08/09/2024).

[63] J. D. Lee and K. A. See. "Trust in Automation: Designing for Appropriate Reliance". en. In: *Human Factors* 46.1 (Mar. 2004). Publisher: SAGE Publications Inc, pp. 50–80. ISSN: 0018-7208. DOI: 10.1518/hfes.46.1.50_30392. URL: https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392 (visited on 07/29/2024).

[64] J. Hansen and M. Wänke. "Liking What's Familiar: The Importance of Unconscious Familiarity in the Mere-Exposure Effect". In: *Social Cognition* 27.2 (2009), pp. 161–182. DOI: 10.1521/soco.2009.27.2.161. URL: https://guilfordjournals.com/doi/abs/10.1521/soco.2009.27.2.161.

[65] P. L. Berger and T. Luckmann. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. New York: Anchor Books, 1966. ISBN: 9780385058988.

[66] J. Bryant, M. B. Oliver, and A. A. Raney. *Media Effects: Advances in Theory and Research*. 4th. New York: Routledge, 2023. ISBN: 9781138480215.

[67] M. Bauer and J. Durant. "Belief in Astrology: a social-psychological analysis". en. In: *Culture and Cosmos* 1.01 (June 1997), pp. 55–71. DOI: 10.46472/CC.0101.0211. URL: http://www.cultureandcosmos.org/pdfs/1-1/CandC-1_1-Bauer-and-Durant-p55.pdf (visited on 07/29/2024).

[68] E. Solberg, M. Kaarstad, M. H. R. Eitrheim, R. Bisio, K. Reegård, and M. Bloch. "A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids". en. In: *Group & Organization Management* 47.2 (Apr. 2022). Publisher: SAGE Publications Inc, pp. 187–222. ISSN: 1059-6011. DOI: 10.1177/10596011221081238. URL: https://doi.org/10.1177/10596011221081238 (visited on 07/29/2024).

[69] R. R. McCrae and P. T. Costa. "Validation of the five-factor model of personality across instruments and observers". In: *Journal of Personality and Social Psychology* 52.1 (1987). Place: US Publisher: American Psychological Association, pp. 81–90. ISSN: 1939-1315. DOI: 10.1037/0022-3514.52.1.81.

[70] O. P. John, L. P. Naumann, and C. J. Soto. "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues". In: *Handbook of personality: Theory and research, 3rd ed*. New York, NY, US: The Guilford Press, 2008, pp. 114–158. ISBN: 978-1-59385-836-0.

[71] A. Corbeanu. "Conscientiousness and Cognitive Abilities: a Meta-Analysis". In: *Psihologia Resurselor Umane* 21.1 (May 28, 2023). Number: 1. ISSN: 2392-8077. DOI: 10.24837/pru.v21i1.531. URL: https://www.hrp-journal.com/index.php/pru/article/view/531 (visited on 07/02/2024).

[72] J. Meyer, O. Lüdtke, F. T. C. Schmidt, J. Fleckenstein, U. Trautwein, and O. Köller. "Conscientiousness and Cognitive Ability as Predictors of Academic Achievement: Evidence of Synergistic Effects From Integrative Data Analysis". en. In: *European Journal of Personality* 38.1 (Jan. 2024). Publisher: SAGE Publications Ltd, pp. 36–52. ISSN: 0890-2070. DOI: 10.1177/08902070221127065. URL: https://doi.org/10.1177/08902070221127065 (visited on 07/29/2024).

[73] R. F. Pohl. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. en. Google-Books-ID: MS5Fr8safgEC. Psychology Press, Dec. 2012. ISBN: 978-1-135-84495-0.

[74] J. S. B. Evans, J. L. Barston, and P. Pollard. "On the conflict between logic and belief in syllogistic reasoning". In: *Memory & Cognition* 11.3 (1983). Place: US Publisher: Psychonomic Society, pp. 295–306. ISSN: 1532-5946. DOI: 10.3758/BF03196976.

[75] K. C. Klauer, J. Musch, and B. Naumer. "On belief bias in syllogistic reasoning". In: *Psychological Review* 107.4 (Oct. 2000), pp. 852–884. ISSN: 0033-295X. DOI: 10.1037/0033-295x.107.4.852.

[76] S. Chandra, S. Verma, W. M. Lim, S. Kumar, and N. Donthu. "Personalization in personalized marketing: Trends and ways forward". In: *Psychology & Marketing* 39.5 (2022), pp. 881–897. DOI: 10.1002/mar.21670. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/mar.21670.

[77] C. H. Chu, S. Donato-Woodger, S. S. Khan, R. Nyrup, K. Leslie, A. Lyn, T. Shi, A. Bianchi, S. A. Rahimi, and A. Grenier. "Age-related bias and artificial intelligence: a scoping review". en. In: *Humanities and Social Sciences Communications* 10.1 (Aug. 2023). Publisher: Palgrave, pp. 1–17. ISSN: 2662-9992. DOI: 10.1057/s41599-023-01999-y. URL: https://www.nature.com/articles/s41599-023-01999-y (visited on 07/31/2024).

[78] E. Shandilya and M. Fan. "Understanding Older Adults' Perceptions and Challenges in Using AI-enabled Everyday Technologies". In: *Proceedings of the Tenth International Symposium of Chinese CHI*. Chinese CHI '22. New York, NY, USA: Association for Computing Machinery, Feb. 2024, pp. 105–116. ISBN: 978-1-4503-9869-5. DOI: 10.1145/3565698.3565774. URL: https://doi.org/10.1145/3565698.3565774 (visited on 07/31/2024).

[79] D. Yakar, Y. P. Ongena, T. C. Kwee, and M. Haan. "Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine". eng. In: *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research* 25.3 (Mar. 2022), pp. 374–381. ISSN: 1524-4733. DOI: 10.1016/j.jval.2021.09.004.

[80] K. A. Hoff and M. Bashir. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust". In: *Human Factors* 57.3 (May 2015). Publisher: SAGE Publications Inc, pp. 407–434. ISSN: 0018-7208. DOI: 10.1177/0018720814547570. URL: https://doi.org/10.1177/0018720814547570 (visited on 04/24/2023).

[81] S. Villa, R. Welsch, A. Denisova, and T. Kosch. "Evaluating Interactive AI: Understanding and Controlling Placebo Effects in Human-AI Interaction". In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. New York, NY, USA: Association for Computing Machinery, May 11, 2024, pp. 1–4. ISBN: 9798400703317. DOI: 10.1145/3613905.3636304. URL: https://doi.org/10.1145/3613905.3636304 (visited on 07/05/2024).

[82] A. M. Kloft, R. Welsch, T. Kosch, and S. Villa. ""AI enhances our performance, I have no doubt this one will do the same": The Placebo effect is robust to negative descriptions of AI". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. New York, NY, USA: Association for Computing Machinery, May 11, 2024, pp. 1–24. ISBN: 9798400703300. DOI: 10.1145/3613904.3642633. URL: https://doi.org/10.1145/3613904.3642633 (visited on 07/05/2024).

[83] P. Pataranutaporn, R. Liu, E. Finn, and P. Maes. "Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness". In: *Nature Machine Intelligence* 5.10 (2023), pp. 1076–1086.

[84] S. Larghi and E. Datteri. "Mentalistic Stances Towards AI Systems: Beyond the Intentional Stance". en. In: *Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops*. Ed. by A. Aldini. Cham: Springer Nature Switzerland, 2024, pp. 28–41. ISBN: 978-3-031-66021-4. DOI: 10.1007/978-3-031-66021-4_2.

[85] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate. "Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems". en. In: *26th International Conference on Intelligent User Interfaces*. College Station TX USA: ACM, Apr. 2021, pp. 340–350. ISBN: 978-1-4503-8017-1. DOI: 10.1145/3397481.3450639. URL: https://dl.acm.org/doi/10.1145/3397481.3450639 (visited on 04/23/2023).

[86] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance". en. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), pp. 2–11. ISSN: 2769-1349, 2769-1330. DOI: 10.1609/hcomp.v7i1.5285. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5285 (visited on 04/22/2023).

[87] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong. "Too much, too little, or just right? Ways explanations impact end users' mental models". en. In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. San Jose, CA, USA: IEEE, Sept. 2013, pp. 3–10. ISBN: 978-1-4799-0369-6. DOI: 10.1109/VLHCC.2013.6645235. URL: https://ieeexplore.ieee.org/document/6645235 (visited on 04/21/2023).

[88] W. B. Rouse and N. M. Morris. "On Looking Into the Black Box: Prospects and Limits in the Search for Mental Models". en. In: *Psychological Bulletin* 100 (1986), pp. 349–363.

[89] K. Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021. ISBN: 978-0-300-20957-0. URL: https://katecrawford.net/atlas.

[90] J. S. B. T. Evans and J. Curtis-Holmes. "Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning". In: *Thinking & Reasoning* 11.4 (Sept. 2005). Publisher: Routledge _eprint: https://doi.org/10.1080/13546780542000005, pp. 382–389. ISSN: 1354-6783. DOI: 10.1080/13546780542000005. URL: https://doi.org/10.1080/13546780542000005 (visited on 07/29/2024).

[91] L. J. Ball, V. A. Thompson, and E. J. N. Stupple. "Conflict and Dual Process Theory". en. In: *Dual Process Theory 2.0*. Ed. by W. De Neys. 1st ed. Routledge, Nov. 2017, pp. 100–120. ISBN: 978-1-315-20455-0. DOI: 10.4324/9781315204550-7. URL: https://www.taylorfrancis.com/books/9781351791373/chapters/10.4324/9781315204550-7 (visited on 07/29/2024).

[92] J. S. B. T. Evans. "Thinking and believing". In: *Mental models in reasoning*. Ed. by J. Garcìa-Madruga, N. Carriedo, and M. J. González-Labra. Madrid: UNED, 2000, pp. 41–56.

[93] B. R. Forer. "The fallacy of personal validation: a classroom demonstration of gullibility". In: *The Journal of Abnormal and Social Psychology* 44.1 (1949). Place: US Publisher: American Psychological Association, pp. 118–123. ISSN: 0096-851X. DOI: 10.1037/h0059240.

[94] P. Glick, D. Gottesman, and J. Jolton. "The Fault is not in the Stars: Susceptibility of Skeptics and Believers in Astrology to the Barnum Effect". In: *Personality and Social Psychology Bulletin* 15.4 (Dec. 1, 1989). Publisher: SAGE Publications Inc, pp. 572–583. ISSN: 0146-1672. DOI: 10.1177/0146167289154010. URL: https://doi.org/10.1177/0146167289154010 (visited on 05/23/2024).

[95] J. T. Johnson, L. M. Cain, T. L. Falke, J. Hayman, and E. Perillo. "The "Barnum effect" revisited: Cognitive and motivational factors in the acceptance of personality descriptions". In: *Journal of Personality and Social Psychology* 49.5 (1985). Place: US Publisher: American Psychological Association, pp. 1378–1391. ISSN: 1939-1315. DOI: 10.1037/0022-3514.49.5.1378.

[96] K. I. Gero et al. "Mental Models of AI Agents in a Cooperative Game Setting". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–12. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376316. URL: https://doi.org/10.1145/3313831.3376316 (visited on 08/06/2024).

[97] V. Danry, P. Pataranutaporn, Z. Epstein, M. Groh, and P. Maes. "Deceptive AI systems that give explanations are just as convincing as honest AI systems in human-machine decision making". In: *arXiv preprint arXiv:2210.08960* (2022). URL: https://arxiv.org/abs/2210.08960.

[98] A. V. Gonzalez, A. Rogers, and A. Søgaard. *On the Interaction of Belief Bias and Explanations*. arXiv:2106.15355 [cs]. June 2021. DOI: 10.48550/arXiv.2106.15355. URL: http://arxiv.org/abs/2106.15355 (visited on 07/29/2024).

[99] V. Danry, P. Pataranutaporn, Y. Mao, and P. Maes. "Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580672. URL: https://doi.org/10.1145/3544548.3580672.

[100] G. Nguyen, M. R. Taesiri, and A. Nguyen. *Visual correspondence-based explanations improve AI robustness and human-AI team accuracy*. 2023. arXiv: 2208.00780 [cs.CV]. URL: https://arxiv.org/abs/2208.00780.

[101] A. J. Nederhof. "Methods of coping with social desirability bias: A review". en. In: *European Journal of Social Psychology* 15.3 (1985). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420 pp. 263–280. ISSN: 1099-0992. DOI: 10.1002/ejsp.2420150303. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420150303 (visited on 07/31/2024).

[102] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff. "Common method biases in behavioral research: a critical review of the literature and recommended remedies." In: *Journal of applied psychology* 88.5 (2003), p. 879.

[103] M. Sharma et al. *Towards Understanding Sycophancy in Language Models*. 2023. arXiv: 2310.13548 [cs.CL]. URL: https://arxiv.org/abs/2310.13548.

[104] N. Sharma, Q. V. Liao, and Z. Xiao. *Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking*. Feb. 10, 2024. DOI: 10.48550/arXiv.2402.05880. arXiv: 2402.05880[cs]. URL: http://arxiv.org/abs/2402.05880 (visited on 05/14/2024).

[105] B. R. Forer. "The fallacy of personal validation: A classroom demonstration of gullibility." In: *The Journal of Abnormal and Social Psychology* 44.1 (1949), p. 118.

[106] C. Snyder and G. R. Larson. "A further look at student acceptance of general personality interpretations." In: *Journal of Consulting and Clinical Psychology* 38.3 (1972), p. 384.

[107] D. Dickson and I. W. Kelly. "The 'Barnum Effect'in personality assessment: A review of the literature". In: *Psychological reports* 57.2 (1985), pp. 367–382.

[108] M. W. Matlin. "Pollyanna Principle". In: *Cognitive Illusions*. 2nd ed. Num Pages: 21. Psychology Press, 2016. ISBN: 978-1-315-69693-5.

[109] M. Lecompte-Van Poucke. "'You got this!': A critical discourse analysis of toxic positivity as a discursive construct on Facebook". In: *Applied Corpus Linguistics* 2.1 (2022), p. 100015.

[110] B. Shneiderman. "Human-centered artificial intelligence: Three fresh ideas". In: *AIS Transactions on Human-Computer Interaction* 12.3 (2020), pp. 109–124.

[111] D. Yang, S. T. Wu, and M. A. Hearst. "Human-AI Interaction in the Age of LLMs". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*. 2024, pp. 34–38.

[112] S. Marri. "A Conceptual Framework for Conversational Human-AI Interaction Design (CHAI)". In: *Design in the Era of Industry 4.0, Volume 1*. Ed. by A. Chakrabarti and V. Singh. Singapore: Springer Nature, 2023, pp. 627–640. ISBN: 978-981-9902-93-4. DOI: 10.1007/978-981-99-0293-4_50.

[113] L. Ibrahim, S. Huang, L. Ahmad, and M. Anderljung. "Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks". In: *arXiv preprint arXiv:2405.10632* (2024).

[114] K. M. Lee. "Media equation theory". In: *The international encyclopedia of communication* (2008).

[115] B. Reeves and C. Nass. "The media equation: How people treat computers, television, and new media like real people". In: *Cambridge, UK* 10 (1996), p. 236605.

[116] J. Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[117] S. Natale. "If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA". In: *new media & society* 21.3 (2019), pp. 712–728.

[118] B. P. Knijnenburg and M. C. Willemsen. "Inferring capabilities of intelligent agents from their external traits". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6.4 (2016), pp. 1–25.

[119] J. Złotowski, H. Sumioka, S. Nishio, D. F. Glas, C. Bartneck, and H. Ishiguro. "Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy". In: *Paladyn, Journal of Behavioral Robotics* 7.1 (2016).

[120] D. Li, P.-L. Rau, and Y. Li. "A cross-cultural study: Effect of robot appearance and task". In: *International Journal of Social Robotics* 2.2 (2010), pp. 175–186.

[121] T. Komatsu and S. Yamada. "Effect of agent appearance on people's interpretation of agent's attitude". In: *CHI'08 Extended Abstracts on Human Factors in Computing Systems*. 2008, pp. 2919–2924.

[122] Z. Pi, L. Deng, X. Wang, P. Guo, T. Xu, and Y. Zhou. "The influences of a virtual instructor's voice and appearance on learning from video lectures". In: *Journal of Computer Assisted Learning* (2022).

[123] M. Paetzel. "The influence of appearance and interaction strategy of a social robot on the feeling of uncanniness in humans". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 522–526.

[124] T. Koda and P. Maes. "Agents with faces: The effect of personification". In: *Proceedings 5th IEEE international workshop on robot and human communication. RO-MAN'96 TSUKUBA*. IEEE. 1996, pp. 189–194.

[125] S. B. H. Pias, R. Huang, D. S. Williamson, M. Kim, and A. Kapadia. "The Impact of Perceived Tone, Age, and Gender on Voice Assistant Persuasiveness in the Context of Product Recommendations". In: *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. CUI '24. New York, NY, USA: Association for Computing Machinery, July 2024, pp. 1–15. ISBN: 9798400705113. DOI: 10.1145/3640794.3665545. URL: https://dl.acm.org/doi/10.1145/3640794.3665545 (visited on 08/17/2024).

[126] K. L. Goodman and C. B. Mayhorn. "Pitch Perfect: Influence of Perceived Voice Agent Gender and Vocal Pitch on Trust and Reliance". en. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65.1 (Sept. 2021). Publisher: SAGE Publications Inc, pp. 1529–1530. ISSN: 1071-1813. DOI: 10.1177/1071181321651126. URL: https://doi.org/10.1177/1071181321651126 (visited on 08/17/2024).

[127] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura. "Voice in human–agent interaction: a survey". In: *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–43.

[128] K. Seaborn and J. Urakami. "Measuring voice UX quantitatively: A rapid review". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–8.

[129] J. Ehret, A. Bönsch, L. Aspöck, C. T. Röhr, S. Baumann, M. Grice, J. Fels, and T. W. Kuhlen. "Do prosody and embodiment influence the perceived naturalness of conversational agents' speech?" In: *ACM Transactions on Applied Perception (TAP)* 18.4 (2021), pp. 1–15.

[130] Y. Kim, M. Reza, J. McGrenere, and D. Yoon. "Designers characterize naturalness in voice user interfaces: their goals, practices, and challenges". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–13.

[131] M. P. Aylett, B. R. Cowan, and L. Clark. "Siri, Echo and performance: You have to suffer darling". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–10.

[132] J. R. Lewis and M. L. Hardzinski. "Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire". In: *International Journal of Speech Technology* 18.3 (2015), pp. 479–487.

[133] A. H.-C. Hwang and A. S. Won. "AI in Your Mind: Counterbalancing Perceived Agency and Experience in Human-AI Interaction". In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–10.

[134] M. Cohn and G. Zellou. "Expressiveness Influences Human Vocal Alignment Toward voice-AI". en. In: *Interspeech 2019*. ISCA, Sept. 2019, pp. 41–45. DOI: 10.21437/Interspeech.2019-1368. URL: https://www.isca-archive.org/interspeech_2019/cohn19_interspeech.html (visited on 08/17/2024).

[135] A. Niculescu, B. van Dijk, A. Nijholt, H. Li, and S. L. See. "Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy". en. In: *International Journal of Social Robotics* 5.2 (Apr. 2013), pp. 171–191. ISSN: 1875-4805. DOI: 10.1007/s12369-012-0171-x. URL: https://doi.org/10.1007/s12369-012-0171-x (visited on 08/17/2024).

[136] S. T. Völkel, D. Buschek, M. Eiband, B. R. Cowan, and H. Hussmann. "Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15.

[137] M. Kraus, N. Wagner, and W. Minker. "Effects of proactive dialogue strategies on human-computer trust". In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 2020, pp. 107–116.

[138] J. Feine, U. Gnewuch, S. Morana, and A. Maedche. "A taxonomy of social cues for conversational agents". In: *International Journal of Human-Computer Studies* 132 (2019), pp. 138–161.

[139] Á. Castro-González, H. Admoni, and B. Scassellati. "Effects of form and motion on judgments of social robots animacy, likability, trustworthiness and unpleasantness". In: *International Journal of Human-Computer Studies* 90 (2016), pp. 27–38.

[140] R. van den Brule, R. Dotsch, G. Bijlstra, D. H. Wigboldus, and P. Haselager. "Do robot performance and behavioral style affect human trust?" In: *International journal of social robotics* 6.4 (2014), pp. 519–531.

[141] S. Song and S. Yamada. "Expressing emotions through color, sound, and vibration with an appearance-constrained social robot". In: *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE. 2017, pp. 2–11.

[142] R. B. Paradeda, M. Hashemian, R. A. Rodrigues, and A. Paiva. "How facial expressions and small talk may influence trust in a robot". In: *International Conference on Social Robotics*. Springer. 2016, pp. 169–178.

[143] R. Mallick, C. Flathmann, W. Duan, B. G. Schelble, and N. J. McNeese. "What you say vs what you do: Utilizing positive emotional expressions to relay AI teammate intent within human-AI teams". In: *International Journal of Human-Computer Studies* (Aug. 13, 2024), p. 103355. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2024.103355. URL: https://www.sciencedirect.com/science/article/pii/S1071581924001381 (visited on 08/16/2024).

[144] C. N. Moridis and A. A. Economides. "Affective Learning: Empathetic Agents with Emotional Facial and Tone of Voice Expressions". In: *IEEE Transactions on Affective Computing* 3.3 (July 2012). Conference Name: IEEE Transactions on Affective Computing, pp. 260–272. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2012.6. URL: https://ieeexplore.ieee.org/abstract/document/6175887?casa_token=_cJji7esrA8AAAAA:EtLk8jLGfTP_zoyftPfcJ6CGXutiUlk3t9_VLj-0d4EUBDJn3qNuSI9bvyybrA8FFLoSU6s (visited on 08/17/2024).

[145] C. Breazeal. *Designing sociable robots*. MIT press, 2004.

[146] P. N. Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press, 1983.

[147] D. A. Norman. "Some observations on mental models". In: *Mental models*. Psychology Press, 2014, pp. 15–22.

[148] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. "Beyond accuracy: The role of mental models in human-AI team performance". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 2019, pp. 2–11.

[149] H. Rutjes, M. Willemsen, and W. IJsselsteijn. "Considerations on explainable AI and users' mental models". In: *CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*. Association for Computing Machinery, Inc. 2019.

[150] K. I. Gero, Z. Ashktorab, C. Dugan, Q. Pan, J. Johnson, W. Geyer, M. Ruiz, S. Miller, D. R. Millen, M. Campbell, et al. "Mental models of AI agents in a cooperative game setting". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.

[151] D. E. Kieras and S. Bovair. "The role of a mental model in learning to operate a device". In: *Cognitive science* 8.3 (1984), pp. 255–273.

[152] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan. "Tell me more? The effects of mental model soundness on personalizing an intelligent agent". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pp. 1–10.

[153] U. Usman, T. Kim, A. Garvey, and A. Duhachek. "The Persuasive Power of AI Ingratiation: A Persuasion Knowledge Theory Perspective". In: *Journal of the Association for Consumer Research* 9.3 (2024), pp. 319–331. DOI: 10.1086/730280. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/730280?journalCode=jacr.

[154] C. Snyder and R. J. Shenkel. "Effects of" favorability," modality, and relevance on acceptance of general personality interpretations prior to and after receiving diagnostic feedback." In: *Journal of Consulting and Clinical Psychology* 44.1 (1976), p. 34.

[155] R. A. Cummins and H. Nistico. "Maintaining Life Satisfaction: The Role of Positive Cognitive Bias". In: *Journal of Happiness Studies* 3.1 (Mar. 1, 2002), pp. 37–69. ISSN: 1573-7780. DOI: 10.1023/A:1015678915305. URL: https://doi.org/10.1023/A:1015678915305 (visited on 08/12/2024).

[156] H. C. Lench and S. W. Bench. "Automatic Optimism: Why People Assume Their Futures will be Bright". In: *Social and Personality Psychology Compass* 6.4 (2012). _eprint: https://onlinelibrary.wiley.com 9004.2012.00430.x, pp. 347–360. ISSN: 1751-9004. DOI: 10.1111/j.1751-9004.2012.00430.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9004.2012.00430.x (visited on 08/12/2024).

[157] H. B. Bergsieker, L. M. Leslie, V. S. Constantine, and S. T. Fiske. "Stereotyping by Omission: Eliminate the Negative, Accentuate the Positive". In: *Journal of personality and social psychology* 102.6 (June 2012), pp. 1214–1238. ISSN: 0022-3514. DOI: 10.1037/a0027717. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4251455/ (visited on 08/12/2024).

[158] J. E. Marineau and G. J. Labianca. "Positive and negative tie perceptual accuracy: Pollyanna principle vs. negative asymmetry explanations". In: *Social Networks* 64 (2021), pp. 83–98.

[159]   P. Suwanaposee, C. Gutwin, Z. Chen, and A. Cockburn. "'Specially For You'–Examining the Barnum Effect's Influence on the Perceived Quality of System Recommendations". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–11.

[160]   J. Wester, S. De Jong, H. Pohl, and N. Van Berkel. "Exploring People's Perceptions of LLM-generated Advice". In: *Computers in Human Behavior: Artificial Humans* (2024), p. 100072.

[161]   A. Bashkirova and D. Krpan. "Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance". In: *Computers in Human Behavior: Artificial Humans* 2.1 (Jan. 2024), p. 100066. ISSN: 2949-8821. DOI: 10.1016/j.chbah.2024.100066. URL: https://www.sciencedirect.com/science/article/pii/S2949882124000264 (visited on 08/12/2024).

[162]   T. Nazaretsky, M. Cukurova, M. Ariely, and G. Alexandron. *Confirmation bias and trust: Human factors that influence teachers' attitudes towards AI-based educational technology*. eng. Proceedings paper. Conference Name: AIBL 2021 Meeting Name: AIBL 2021 Place: Bozen-Bolzano, Italy Volume: 3042. Sept. 2021. URL: http://ceur-ws.org/Vol-3042/ (visited on 08/12/2024).

[163]   J. Shepperd, W. Malone, and K. Sweeny. "Exploring causes of the self-serving bias". In: *Social and Personality Psychology Compass* 2.2 (2008), pp. 895–908.

[164]   A. H. Mezulis, L. Y. Abramson, J. S. Hyde, and B. L. Hankin. "Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias". In: *Psychological Bulletin* 130.5 (Sept. 2004), pp. 711–747. ISSN: 0033-2909. DOI: 10.1037/0033-2909.130.5.711.

[165]   E. J. Langer. "The illusion of control." In: *Journal of personality and social psychology* 32.2 (1975), p. 311.

[166]   E. Purificato, L. Boratto, and E. W. De Luca. *User Modeling and User Profiling: A Comprehensive Survey*. arXiv:2402.09660 [cs]. Feb. 2024. DOI: 10.48550/arXiv.2402.09660. URL: http://arxiv.org/abs/2402.09660 (visited on 08/13/2024).

[167]   W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang. *MemoryBank: Enhancing Large Language Models with Long-Term Memory*. en. arXiv:2305.10250 [cs]. May 2023. URL: http://arxiv.org/abs/2305.10250 (visited on 11/09/2023).

[168]   R. K. Merton. "The Self-Fulfilling Prophecy". In: *Antioch Review* 8 (1948), pp. 193–210. DOI: 10.2307/4609267. URL: http://dx.doi.org/10.2307/4609267.

[169] L. Jussim. "Self-fulfilling prophecies: A theoretical and integrative review". In: *Psychological Review* 93.4 (1986). Place: US Publisher: American Psychological Association, pp. 429–445. ISSN: 1939-1471. DOI: 10.1037/0033-295X.93.4.429.

[170] S. Almadi. "The meta-narrative of self-fulfilling prophecy in the different research areas". In: *Society and Economy* (Mar. 8, 2022). ISSN: 1588-9726, 1588-970X. DOI: 10.1556/204.2022.00001. URL: https://akjournals.com/view/journals/204/aop/article-10.1556-204.2022.00001/article-10.1556-204.2022.00001.xml (visited on 12/03/2023).

[171] R. L. Henshel. "The Boundary of the Self-Fulfilling Prophecy and the Dilemma of Social Prediction". In: *The British Journal of Sociology* 33.4 (1982). Publisher: [Wiley, London School of Economics and Political Science, London School of Economics], pp. 511–528. ISSN: 0007-1315. DOI: 10.2307/589359. URL: https://www.jstor.org/stable/589359 (visited on 08/14/2024).

[172] R. Rosenthal and L. Jacobson. "Pygmalion in the classroom". en. In: *The Urban Review* 3.1 (Sept. 1968), pp. 16–20. ISSN: 1573-1960. DOI: 10.1007/BF02322211. URL: https://doi.org/10.1007/BF02322211 (visited on 08/11/2024).

[173] R. Rosenthal. "The Pygmalion effect and its mediating mechanisms". In: *Improving academic achievement*. Elsevier, 2002, pp. 25–36.

[174] A. Friedrich, B. Flunger, B. Nagengast, K. Jonkmann, and U. Trautwein. "Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement". In: *Contemporary Educational Psychology* 41 (2015), pp. 1–12.

[175] L. Switzky. "ELIZA Effects: Pygmalion and the Early Development of Artificial Intelligence". In: *Shaw: The Journal of Bernard Shaw Studies* 40.1 (2023), pp. 50–65. DOI: 10.5325/shaw.40.1.0050. URL: https://scholarlypublishingcollective.org/psup/shaw/article/40/1/50/199270/ELIZA-Effects-Pygmalion-and-the-Early-Development.

[176] M. De-Arteaga and J. Elmer. "Self-fulfilling prophecies and machine learning in resuscitation science". In: *Resuscitation* 183 (Feb. 2023), p. 109622. ISSN: 03009572. DOI: 10.1016/j.resuscitation.2022.10.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0300957222006931 (visited on 12/16/2023).

[177] O. C. King and M. Mertens. "Self-fulfilling Prophecy in Practical and Automated Prediction". In: *Ethical Theory and Moral Practice* 26.1 (Mar. 2023), pp. 127–152. ISSN: 1386-2820, 1572-8447. DOI: 10.1007/s10677-022-10359-9. URL: https://link.springer.com/10.1007/s10677-022-10359-9 (visited on 12/16/2023).

[178] K. Bauer and A. Gill. "Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies". In: *Information Systems Research* (May 3, 2023), isre.2023.1217. ISSN: 1047-7047, 1526-5536. DOI: 10.1287/isre.2023.1217. URL: https://pubsonline.informs.org/doi/10.1287/isre.2023.1217 (visited on 12/16/2023).

[179] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. *Runaway Feedback Loops in Predictive Policing*. arXiv:1706.09847 [cs, stat]. Dec. 2017. DOI: 10.48550/arXiv.1706.09847. URL: http://arxiv.org/abs/1706.09847 (visited on 08/14/2024).

[180] D. Petropoulos Petalas, H. Van Schie, and P. Hendriks Vettehen. "Forecasted economic change and the self-fulfilling prophecy in economic decision-making". In: *PLOS ONE* 12.3 (Mar. 23, 2017). Ed. by H. Rao, e0174353. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0174353. URL: https://dx.plos.org/10.1371/journal.pone.0174353 (visited on 12/03/2023).

[181] J. Juergensen, J. S. Weaver, C. N. May, and H. A. Demaree. "More Than Money: Experienced Positive Affect Reduces Risk-Taking Behavior on a Real-World Gambling Task". English. In: *Frontiers in Psychology* 9 (Nov. 2018). Publisher: Frontiers. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.02116. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.02116/full (visited on 08/14/2024).

[182] A. M. Isen, T. E. Nygren, and F. G. Ashby. "Influence of positive affect on the subjective utility of gains and losses: it is just not worth the risk". eng. In: *Journal of Personality and Social Psychology* 55.5 (Nov. 1988), pp. 710–717. ISSN: 0022-3514. DOI: 10.1037//0022-3514.55.5.710.

[183] T. E. Nygren, A. M. Isen, P. J. Taylor, and J. Dulin. "The Influence of Positive Affect on the Decision Rule in Risk Situations: Focus on Outcome (and Especially Avoidance of Loss) Rather Than Probability". In: *Organizational Behavior and Human Decision Processes* 66.1 (Apr. 1996), pp. 59–72. ISSN: 0749-5978. DOI: 10.1006/obhd.1996.0038. URL: https://www.sciencedirect.com/science/article/pii/S0749597896900382 (visited on 08/15/2024).

[184] B. E. Hermalin and A. M. Isen. *The Effect of Affect on Economic and Strategic Decision Making*. en. SSRN Scholarly Paper. Rochester, NY, July 2000. DOI: 10.2139/ssrn.200295. URL: https://papers.ssrn.com/abstract=200295 (visited on 08/15/2024).

[185] J. S. Lerner and D. Keltner. "Beyond valence: Toward a model of emotion-specific influences on judgement and choice". en. In: *Cognition & Emotion* 14.4 (July 2000), pp. 473–493. ISSN: 0269-9931, 1464-0600. DOI: 10.1080/026999300402763. URL: http://www.tandfonline.com/doi/abs/10.1080/026999300402763 (visited on 07/25/2024).

[186] R. Raghunathan and M. T. Pham. "All Negative Moods Are Not Equal: Motivational Influences of Anxiety and Sadness on Decision Making". In: *Organizational Behavior and Human Decision Processes* 79.1 (July 1999), pp. 56–77. ISSN: 0749-5978. DOI: 10.1006/obhd.1999.2838. URL: https://www.sciencedirect.com/science/article/pii/S0749597899928388 (visited on 08/15/2024).

[187] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam. "Emotion and Decision Making". In: *Annual Review of Psychology* 66.1 (Jan. 3, 2015), pp. 799–823. ISSN: 0066-4308, 1545-2085. DOI: 10.1146/annurev-psych-010213-115043. URL: https://www.annualreviews.org/doi/10.1146/annurev-psych-010213-115043 (visited on 04/11/2024).

[188] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. Welch. "Risk as feelings". eng. In: *Psychological Bulletin* 127.2 (Mar. 2001), pp. 267–286. ISSN: 0033-2909. DOI: 10.1037/0033-2909.127.2.267.

[189] T. Schlösser, D. Dunning, and D. Fetchenhauer. "What a feeling: The role of immediate and anticipated emotions in risky decisions". In: *Journal of Behavioral Decision Making* 26.1 (2013). Place: US Publisher: John Wiley & Sons, pp. 13–30. ISSN: 1099-0771. DOI: 10.1002/bdm.757.

[190] N. A. Young, M. M. Shuster, and J. A. Mikels. "The Sure Thing: The Role of Integral Affect in Risky Choice Framing". In: *Emotion (Washington, D.C.)* 19.6 (Sept. 2019), pp. 1035–1043. ISSN: 1528-3542. DOI: 10.1037/emo0000505. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6387652/ (visited on 07/26/2024).

[191] J. J. Gross. "Emotion Regulation: Current Status and Future Prospects". In: *Psychological Inquiry* 26.1 (Jan. 2015). Publisher: Routledge _eprint: https://doi.org/10.1080/1047840X.2014.940781, pp. 1–26. ISSN: 1047-840X. DOI: 10.1080/1047840X.2014.940781. URL: https://doi.org/10.1080/1047840X.2014.940781 (visited on 08/15/2024).

[192] M. Matthews, T. L. Webb, R. Shafir, M. Snow, and G. Sheppes. "Identifying the determinants of emotion regulation choice: a systematic review with meta-analysis". In: *Cognition and Emotion* 35.6 (Aug. 2021). Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2021.1945538, pp. 1056–1084. ISSN: 0269-9931. DOI: 10.1080/02699931.2021.1945538. URL: https://doi.org/10.1080/02699931.2021.1945538 (visited on 08/15/2024).

[193] P. Rekar, M. Pahor, and M. Perat. "Effect of emotion regulation difficulties on financial decision-making." en. In: *Journal of Neuroscience, Psychology, and Economics* 16.2 (June 2023), pp. 80–93. ISSN: 2151-318X, 1937-321X. DOI: 10.1037/npe0000172. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/npe0000172 (visited on 07/25/2024).

[194] L. N. Martin and M. R. Delgado. "The Influence of Emotion Regulation on Decision-making under Risk". In: *Journal of Cognitive Neuroscience* 23.9 (Sept. 2011), pp. 2569–2581. ISSN: 0898-929X. DOI: 10.1162/jocn.2011.21618. URL: https://doi.org/10.1162/jocn.2011.21618 (visited on 08/15/2024).

[195] M. Im and J. Oh. "Effect of emotion regulation as a de-biasing mechanism on overconfidence in investment behavior". en. In: *Journal of Financial Services Marketing* 21.3 (Sept. 2016), pp. 209–225. ISSN: 1479-1846. DOI: 10.1057/s41264-016-0003-4. URL: https://doi.org/10.1057/s41264-016-0003-4 (visited on 07/25/2024).

[196] J. Cohen, M. T. Pham, and E. B. Andrade. *The Nature and Role of Affect in Consumer Behavior*. en. SSRN Scholarly Paper. Rochester, NY, Apr. 2008. URL: https://papers.ssrn.com/abstract=1115521 (visited on 08/20/2024).

[197] W. Chu, M. Im, and H. Jang. "Overconfidence and emotion regulation failure: How overconfidence leads to the disposition effect in consumer investment behaviour". en. In: *Journal of Financial Services Marketing* 17.1 (Apr. 2012), pp. 96–116. ISSN: 1479-1846. DOI: 10.1057/fsm.2012.7. URL: https://doi.org/10.1057/fsm.2012.7 (visited on 07/25/2024).