Should I State or Should I Show? Aligning AI with Human Preferences

Keaton Ellis* Wanying Huang†

March 31, 2026

Abstract

As AI agents become more autonomous, properly aligning their objectives with human preferences becomes increasingly important. We study how effectively an AI agent learns a human principal's preference in choice under risk via *stated* versus *revealed* preferences. We conduct an online experiment in which subjects state their preferences through written instructions ("prompts") and reveal them through choices in a series of binary lottery questions ("data"). We find that on average, an AI agent given revealed-preference data predicts subjects' choices more accurately than an AI agent given stated-preference prompts. Further analysis suggests that the gap is driven by subjects' difficulty in translating their own preferences into written instructions. When given a choice between which information source to give to an AI agent, a large portion of subjects fail to select the more informative one. Moreover, when predictions from the two sources conflict, we find that the AI agent aligns more frequently with the prompt, despite its lower accuracy. Overall, these results highlight the revealed preference approach as a powerful mechanism for communicating human preferences to AI agents, but its success depends on careful implementation.

*Department of Economics, Monash University. Email:keaton.ellis@monash.edu
†Department of Economics, Monash University. Email: kate.huang@monash.edu

1 Introduction

The emergence of agentic artificial intelligence (AI) has prompted widespread discussion of a future in which AI agents perform tasks autonomously on behalf of humans, with little or no direct oversight. Although views may differ on the exact form this AI-assisted future will take, there is a scholarly consensus that AI will fundamentally transform economic interactions and productivity.[1] At the same time, the proliferation of AI agents raises new types of principal-agent problems: misalignment may arise not because AI agents have preferences of their own, but because human principals cannot fully articulate their preferences—a phenomenon known as *specification hazard* (Imas et al., 2025; Shahidi et al., 2025).

As a simple example, consider a traveler (the principal) who tasks an AI agent with "buying the cheapest flight from Melbourne, Australia to Washington, D.C." The agent may select an itinerary with an extremely long layover—an outcome the principal might implicitly wish to avoid but did not explicitly rule out in their instructions. More generally, the challenge is that people often cannot fully specify their preferences across a wide range of scenarios. Thus, in practice, AI agents will need to first infer the principal's underlying preferences from coarse or incomplete guidelines and then select the alternative that best aligns with these inferred preferences. Because such incompleteness in preference specification is likely to persist, a first-order challenge is to design mechanisms that mitigate the resulting misalignment.

One natural way to address this challenge is to use the principal's past choices as an alternative source of information. In this paper, we adopt such a *revealed preference* approach, in which observed choice data provides a direct channel through which human principals communicate their preferences to an AI agent. We compare this approach with principals' *stated preferences*, elicited through incentivized prompts that they write to guide the (AI) agent's decisions on their behalf. A priori, it is unclear which approach yields a better alignment. On the one hand, if human principals can articulate their preferences, then an AI agent given those preferences can implement them across a wide range of decision problems. On the other hand, human principals may be unable to articulate their preferences as precisely as their preferences are revealed through their choices.

To compare the efficacy of these approaches, we conduct an incentivized online experiment that elicits both revealed and stated preferences in a choice under risk

[1] See some recent work, e.g., Brynjolfsson et al. (2025); Acemoglu (2025); Gans (2026).

setting. The experiment consists of two main parts. In Part I, subjects answer a series of lottery choice problems that vary in difficulty along several dimensions. They are then asked to write a prompt stating their preferences to an AI agent that will act on their behalf in a new series of similar lottery choice problems. In Part II, subjects answer this new series of lottery choice problems, and their choices are used to benchmark the performance of the AI agents. For each subject, we then instantiate two AI agents using Anthropic's Claude Opus 4.5: one given the subject's earlier choices ("Data-AI") and one given the subject's stated preferences ("Prompt-AI").[2] We evaluate the performance of these agents by comparing their predictions with subjects' actual choices, using out-of-sample prediction accuracy ("match rate") as our metric.

Overall, we find that AI agents perform significantly better on average when given revealed preferences than when given stated preferences. In particular, Data-AI outperforms Prompt-AI across different levels of question difficulty, and its mean match rate is comparable to that achieved by standard economic models of choice under risk (i.e., expected utility theory). At the same time, there is marked heterogeneity in the agent performance gap across subjects: subjects who exhibit more behavioral biases are harder to predict for both types of AI agents. This pattern is especially pronounced for Prompt-AI, which performs as well as Data-AI for subjects who exhibit no such biases, but performs 10% worse for the most biased subjects. Thus, the subjects whose choices are least consistent with the canonical framework are precisely those who benefit most from providing revealed-preference information rather than stated preferences.

We next investigate the source of this performance gap: whether it reflects subjects' difficulty in writing informative prompts or AI agents' difficulty in interpreting them. Our evidence suggests that the gap is largely driven by the former. In particular, we find that subjects' written instructions are more predictive of their Part II choices when they align more closely with their Part I choices. Moreover, when we instead use AI to generate synthetic prompts under the same information and instructions provided to subjects, the performance of the resulting agents, which we refer to as AutoPrompt-AI, is on par with that of Data-AI. In other words, more informative prompts do exist, and subjects' own prompts perform better when they

[2] At the time of writing, this is a frontier large language model (LLM). In Section 4.4, we replicate our analysis using another frontier model, GPT-5.4 from OpenAI, and find that most of our results hold both qualitatively and quantitatively.

better reflect their Part I choices; this suggests that subjects' difficulty in writing informative prompts is the main driver of the performance gap between AI agents.

We then study whether human principals recognize this gap when choosing between the two agentic regimes—Data-AI and Prompt-AI—and the welfare consequences of those choices. In the experiment, after subjects have made their choices and written their prompt, we also ask them to choose exactly one of the two agents (Data-AI or Prompt-AI) to delegate to. The chosen agent then makes choices on a subject's behalf in the second part. We find that subjects are more likely to delegate to Data-AI than to Prompt-AI, with 59% choosing the former. When we later elicit their beliefs about each agent's match rate, we find that their delegation choices reflect perceived AI performance: overall, 85% of subjects choose the agentic regime they believe to be weakly better. However, subjects substantially overestimate the true absolute performance gap and are often mistaken about which agent performs better. Consequently, a substantial share (35%) of subjects fail to choose the better-performing agent.

Finally, we investigate how an AI agent behaves when given both stated and revealed preference information, and we refer to this agent as "Both-AI".[3] Intuitively, if stated and revealed preferences are complementary, this agent should perform strictly better than agents given only one source of information. Moreover, if stated preferences are simply noisier than revealed preferences, an AI agent could, in principle, disregard the noisier stated preferences, rely only on revealed preference information, and achieve a similar match rate to Data-AI. However, we find that Both-AI performs significantly *worse* than Data-AI and only marginally better than Prompt-AI. Interestingly, we find that this performance reduction is driven by the 25% of subject-question pairs in our sample for which Prompt-AI and Data-AI make *conflicting* predictions: within this subset of conflicting questions, Both-AI largely falls back to the prediction of Prompt-AI, despite its lower accuracy.

Taken together, our findings suggest that revealed preference can be an effective approach for human principals to communicate their preferences to an AI agent. However, this approach may also require careful implementation: humans may fail to opt into using this information on their own, and overloading AI agents with information can reduce their predictive performance.

[3]Specifically, for each subject, we instantiate a third AI agent endowed with both the subject's choices and written prompt in the first part of the experiment.

2 Related Literature

Our paper contributes to a burgeoning literature on AI agents as economic actors.[4] A central concern in this literature is the misalignment between an AI agent's objective function and that of its human principal (Gabriel, 2020). Such misalignment can arise from the technical limitations of AI agents, such as inaccurate inference or hallucination (Huang et al., 2025b), but it can also arise from a human principal's inability to fully articulate their preferences to the AI agent. We focus on the second channel of misalignment, which we view as the first-order issue in our environment.[5] As is well known, prompt writing is hard, and there is no reliably positive universal strategy (Zamfirescu-Pereira et al., 2023; Meincke et al., 2025a,b,c). As shown recently in Imas et al. (2025), there is also substantial heterogeneity in human principals' ability to state their preferences, and these differences carry through to economic outcomes when decisions are implemented by AI agents. This source of misalignment may become more severe in high-dimensional decision problems, where it becomes harder for human principals to articulate their preferences precisely (Shahidi et al., 2025; Liang, 2026). Our work shows that revealed preference information, as an alternative way of communicating human principals' preferences, could mitigate such misalignment.

Our paper is related to the line of work that uses LLMs to extract preferences from natural language for economic applications. For example, Rusak et al. (2025) show that LLMs can convert free-text taste descriptions over job roles into cardinal utilities that capture human subjects' preferences, thereby improving allocation mechanisms in a labor-market matching experiment. In an auction setting, Huang et al. (2025a) show that LLM-powered proxies can help determine which revealed-preference information to query from humans. Furthermore, Li et al. (2023) show that such stated preference elicitation can be improved with a framework that allows active feedback and interaction between the language models and humans. Rather than studying how to improve preference elicitation using language, we compare stated and revealed preferences as alternative inputs for aligning AI agents in a controlled and incentivized

[4] See recent surveys, e.g., Immorlica et al. (2024) and Hadfield and Koh (2025).

[5] A growing literature supports this view, showing that LLMs exhibit consistent decision-making in choice under risk contexts (Chen et al., 2023). Kim et al. (2024) show that GPT's recommendations are consistent with expected utility maximization and can be aligned with subjects' risk aversion when provided with simulated choice data.

environment.[6]

Our work also relates to a broad literature studying human-AI interactions.[7] Within this, we are most closely related to studies that consider human beliefs about AI ability and their effects on human delegation decisions to AI. Vafa et al. (2024) show that overestimation of AI performance can lead to worse outcomes because of overdelegation. Similarly, He et al. (2025) find that humans believe AI agent behavior is far closer to their own than it actually is, while Dreyfuss and Raux (2025) find that these beliefs update more sharply in tasks where AI performance does not conform to human expectations (e.g., by performing poorly in a "human-easy task"). We contribute to this literature by examining the question of human beliefs about the performance of different types of AI agents arising from different sources of information, and by showing that the resulting misperceptions also matter for effective human-AI interaction.

3 Experimental Design

Structure of the Experiment. The experiment consists of three parts: Part I and II are the main components of the experiment, in which subjects make choices under risk, write instructions for an AI agent, and decide what type of information they would like to provide to an AI agent that will later act on their behalf.

In Part I of the experiment, subjects complete three tasks. They are informed that their decisions in each task will be used to instantiate an AI agent. This agent will then make decisions on their behalf in a new but structurally similar set of lottery problems that may be used to determine their bonus payment. In the first task, subjects choose between 13 pairs of binary lotteries. These lottery problems are classified into three different categories—"easy", "hard", and "behavioral" (more detail below)—and all are presented in random order. Subjects also have access to five summary statistics about each lottery to assist their decision-making.[8] We refer

[6]Other papers have also collected stated and revealed preferences in an unincentivized manner, such as Fedyk et al. (2024) and Lai et al. (2026). However, they elicit stated preferences only in the more limited form of Likert scales rather than richer free-text responses, and they do not examine LLM responses to the two information sources separately. They also study different settings: the former augments LLMs with demographic information, while the latter focuses on writing assistance.

[7]For a survey, see Jackson et al. (2025).

[8]These statistics are the expected value (displayed as "Average Payment"), the variance (displayed as "Payment Variability"), the minimum payment, the maximum payment, and the probability of receiving the maximum payment. See Arrieta and Nielsen (2025) for a similar implementation.

to the resulting AI agent endowed with this choice information as **Data-AI**.

In the second task, after completing their choices, subjects write a free-form prompt describing their preferences for choosing between lotteries, which will then be given to an AI agent. Subjects are encouraged and incentivized to provide clear, flexible guidance that can be used across different situations, but do not explicitly provide prompt examples to prevent anchoring (Furnham and Boo, 2011). While writing their prompt, subjects have access to their previous choices, including the summary statistics of the lotteries. Finally, to limit the opportunity cost of time spent on the prompt, a minimum of 60 seconds is enforced on the prompt-writing screen before subjects can progress (Spiliopoulos and Ortmann, 2018). We refer to the resulting AI agent endowed with the prompt as **Prompt-AI**.

In the third task, subjects decide which type of information to provide to the AI agent instantiated from this task. They must select exactly one of two options: their *written instructions* from task 2 or their *past choice data* from task 1. We refer to this as their "delegation decision."

In Part II of the experiment, subjects now face the new set of 13 binary lottery pairs that are designed to be structurally similar to those in Part I (see more details on the lottery specifics below). Subjects' choices in these problems then serve as the ground truth for evaluating AI agent predictions. However, subjects are not informed, at the time they make their Part II decisions, that these choices will be used to benchmark AI accuracy. This avoids subjects potentially choosing in accordance with their written prompts or past choices, which may not always correspond to their true preferences. In both Part I and Part II, a brief brain break is implemented after every six binary lottery questions to reduce fatigue. During these breaks, subjects are asked to locate a camouflaged animal in two distinct images.

Part III of the experiment consists of supplemental measures. We elicit subjects' beliefs about the prediction accuracy of Prompt-AI and Data-AI agents. Subjects are incentivized to correctly guess how many questions each type of AI agent predicts correctly. We also ask subjects why they made their delegation choice. Subjects also complete a series of control tasks, containing a short IQ test (ICAR, Condon and Revelle, 2014), risk attitude elicitation using two investment tasks (Gneezy and Potters, 1997), and measures of overconfidence. At the end of the experiment, subjects answer a brief, unincentivized survey covering self-reported comfort with AI tools, literacy, impatience, frequency of AI use, education level, and management experience,

as well as a personality survey (TIPI; Gosling et al., 2003). We additionally collect demographic variables such as age and gender.

Incentives. The payment of subjects consists of several parts. First, subjects receive a $5 completion payment. Second, to incentivize subjects to write their prompt carefully and to select the information they prefer to send to the AI to act on their behalf, one of four sets of 13 binary lottery choices is randomly selected for the bonus payment. Three of these sets consist of AI predictions for the Part II lottery problems, generated using different information sources from Part I: one from Data-AI based on the subject's past choices, one from Prompt-AI based on the subject's written prompt, and one from either based on the subject's delegation decision. The fourth set consists of the subject's own 13 choices from Part II.

From the selected set, *one* decision is drawn at random, with equal probability across the 13 problems, and the corresponding lottery is implemented for payment. This procedure ensures that subjects have incentives to make genuine choices, write informative prompts, and delegate to the AI agent so that it can best act on their behalf. Finally, subjects are paid for almost all decisions made in Part III, except the brief, unincentivized summary survey at the end. The experiment lasted on average 29 minutes. The average total earnings per subject were just over $11, implying an hourly rate of approximately $23 per hour.

Implementation. The experiment was programmed in oTree (Chen et al., 2016) and administered via Prolific in February and March 2026, with recruitment restricted to U.S.-based adults. A total of 147 subjects passed all comprehension quizzes at the beginning and completed the experiment.

All predictions made by the AI agent were generated using Claude Opus 4.5 (`claude-opus-4-5-20251101`), with extended thinking enabled. Each agent (Prompt-AI or Data-AI) is prompted to reason through each lottery pair and submit its choices in a standardized format. The full system prompts used for each agent are reproduced in Appendix B. The study was approved by the Monash Human Research Ethics Committee (ID: 50731), and was pre-registered on AsPredicted (# 268985). Instructions and screenshots of the interface are presented in the Online Appendix.

The Lottery Choices. Human subjects face two sets of 13 lottery pairs, one in Part I and one in Part II. Within each set, the 13 pairs span three conceptually distinct

types of decisions: "easy", "hard", and "behavioral", which are shown in Table A.1 in Appendix A. Easy questions involve either first-order stochastic dominance or large differences in expected value between the two lotteries. Hard questions, in contrast, involve lotteries with smaller expected value differences, more outcomes on average, no obvious heuristic solution, or require reasoning based on more subtle dominance concepts such as second-order stochastic dominance. Finally, behavioral questions are designed to test for (reverse) common ratio and (reverse) common consequence effects—two canonical environments in which violations of expected utility theory are frequently observed empirically (Blavatskyy et al., 2022, 2023).

We use two easy questions and five hard questions in both parts. In Part I, we source one easy question and four hard questions from Agranov and Ortoleva (2017).[9] The remaining easy and hard questions compare first- and second-order dominated lotteries, respectively. In Part II, we increase the stakes tenfold for one easy question and one hard question; for the remaining questions we introduce slight perturbations while keeping expected value differences largely unchanged.

In each part, we also generate two sets of three behavioral questions using the framework of McGranaghan et al. (2026). Specifically, for fixed prizes $H > M > 0$, consider the following three binary choice questions that are parameterized by a vector (p, r) where $p, r \in (0, 1)$:[10]

(i) AB choice: choose lottery $A = (M, 1)$ or lottery $B = (H, p)$.

(ii) AB' choice: choose lottery $A = (M, 1)$ or lottery $B' = (H, pr; M, 1 - r)$

(iii) CD choice: choose lottery $C = (M, r)$ or lottery $D = (H, pr)$.

Under expected utility theory, multiplying the probabilities of non-zero outcomes by a common factor r, or replacing a shared consequence of a $1 - r$ probability of $\$M$ with a $1 - r$ probability of \$0, should not change preferences. However, these factors are routinely observed to affect choice. The common ratio effect is identified by a preference reversal between the AB choice and CD choices: individuals prefer A to B in the AB task but then prefer D to C in the CD task. The common consequence

[9] In Agranov and Ortoleva (2017), lottery outcomes are reported in tokens rather than dollar amounts. We convert those token values into dollars and present all outcomes to subjects directly in monetary terms.

[10] Here, lotteries are represented as $(x_1, p_1; \ldots, x_n, p_n)$ where each outcome x_i occurs with probability $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. To simplify notation, we omit the null outcome of 0 and use (x, p) to denote the prospect $(x, p; 0, 1 - p)$.

effect is identified by an analogous preference reversal between AB' and CD. The reverse of these effects, intuitively, reverse choice in each question.

For the behavioral problems in Part I, we choose $(p, r) = (8/10, 1/4)$ and $(p, r) = (1/2, 1/2)$. For the behavioral problems in Part II, we choose $(p, r) = (10/11, 11/100)$ and $(p, r) = (3/10, 1/2)$. We additionally eliminate certainty for option A in Part II, instead using $A = (M, 0.95)$ to eliminate the certainty effect. These parameterizations are chosen to generate a variety of common ratio, reverse common ratio, common consequence, and reverse common consequence patterns (Allais, 1953; Kahneman and Tversky, 1979; McGranaghan et al., 2026).[11]

4 Results

Our primary outcome of interest is *AI match rate*, defined as the fraction of Part II decisions in which the AI's prediction matches the subject's actual choice. Note that match rate is an out-of-sample measure. We first examine the match rates of Prompt-AI and Data-AI, AI agents instantiated with stated and revealed preference, respectively. We next investigate the source of the discrepancy in match rates and whether subjects identify it. Finally, we turn to how AI agents handle conflicting information coming from the two sources.

4.1 Revealed versus Stated Preferences

Figure 1 presents our first result. Panel (a) shows the cumulative distribution of per-subject match rates for Prompt-AI and Data-AI. The dashed line at 50% indicates the expected match rate of random guessing. First, we see that, regardless of model, the subject-level match rates are nearly all to the right of the dashed line, indicating that AI agents effectively incorporate both revealed preference and stated preference information. Second, we also see vast heterogeneity in the match rate at the subject level, ranging from 31% to 100%. Note that the match rates of these two AI agents are correlated across subjects (Pearson's correlation coefficient $\rho = 0.5$, $p < 0.001$), indicating that some heterogeneity in match rate is driven by heterogeneity in subject

[11] In Part I, the former set of parameters is expected to generate a common ratio effect but no common consequence effect, while the latter is expected to generate strong reverse common ratio and strong reverse common consequence effects. Likewise, in Part II, the former is expected to generate strong common ratio and common consequence effects, while the latter is expected to generate only a reverse common consequence effect.

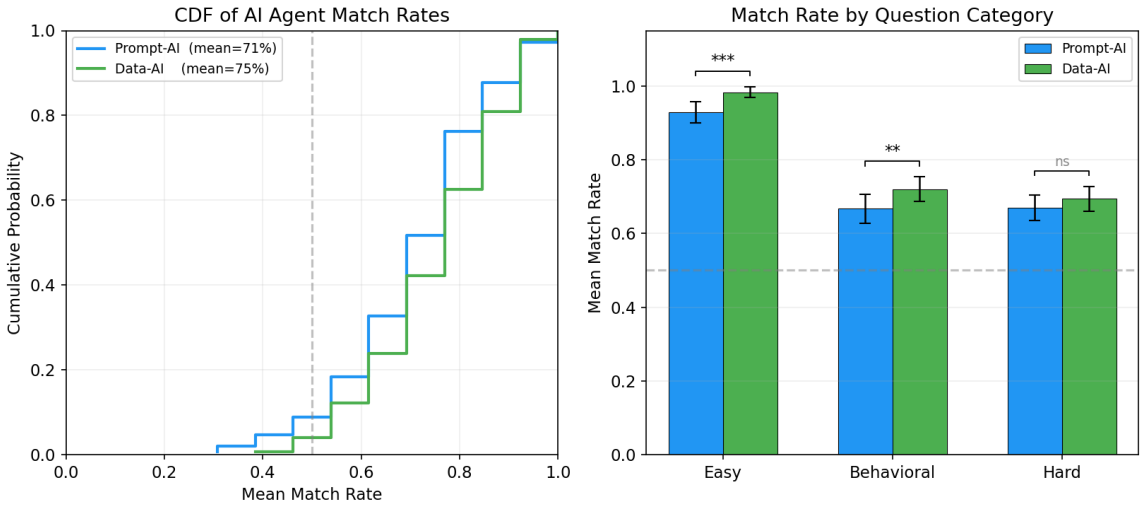

FIGURE 1: Comparison of Prompt-AI and Data-AI match rates ($N = 147$). *Panel (a)*: Empirical CDF of per-subject match rates for Prompt-AI (blue) and Data-AI (green); the dashed vertical line marks 50%. *Panel (b)*: Mean match rates by question category (Easy, Behavioral, Hard); the dashed horizontal line marks 50%. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing Data-AI to Prompt-AI within each category. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

ability. However, when comparing AI agents, the match rate from Data-AI first-order stochastically dominates that from Prompt-AI, indicating systematically higher predictive accuracy across subjects. The average match rate is 75% for Data-AI and 71% for Prompt-AI, resulting in a difference of 4 percentage points (paired t-test: $t = 3.58$, $p < 0.001$; Wilcoxon signed-rank: $p < 0.001$; $N = 147$). Thus, on average AI agents given revealed preferences predict better than those given stated preferences.[12]

Panel (b) of Figure 1 compares match rates across AI agents by the question categories discussed in Section 3. We see that, for both types of AI agents, easy questions exhibit significantly higher match rates than either behavioral or hard questions. In contrast, within each agent, match rates do not differ significantly between behavioral and hard questions (paired t-tests: Prompt-AI, $p = 0.910$; Data-AI, $p = 0.242$), sug-

[12]To benchmark the performance of these AI agents in our setting, we also empirically estimated a standard structural model of risk preferences using subjects' choices in Part I (see more details for its implementation in Appendix C). We find that on average, the EUT model achieves a match rate of 76%, which is statistically indistinguishable from that of Data-AI (paired t-test: $t = 0.94$, $p = 0.348$), suggesting that the predictive performance of Data-AI is comparable to that of a standard structural model of risk preferences. This is supportive of our view that technical limitations of AI agents are not a first-order issue in our environment.

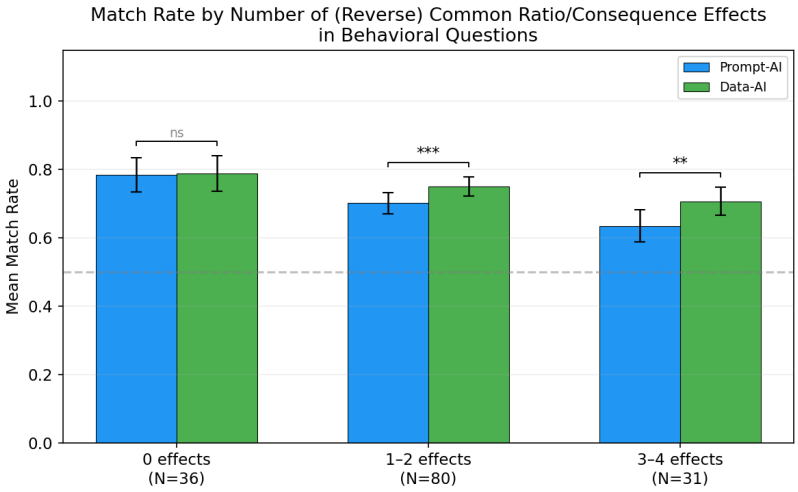

FIGURE 2: Mean match rates by a subject's number of behavioral effects observed in the Behavioral questions; the dashed horizontal line marks 50%. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing Data-AI to Prompt-AI within each group of subjects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

gesting that these two categories are similarly difficult for both agents. Within each question category, we also find that Data-AI exhibits higher match rates than Prompt-AI. These differences are statistically significant for easy and behavioral questions, but not for hard questions (paired t-tests; Easy: $p < 0.001$; Behavioral: $p = 0.011$; Hard: $p = 0.200$). Thus, these results suggest that the benefits of revealed-preference information are heterogeneous across question categories.

Subject-level Heterogeneity. To examine who benefits more from providing the AI agent with revealed-preference information, we classify subjects based on their answers to the two sets of three behavioral questions. Specifically, subjects are classified based on the number of times they exhibit the (reverse) common ratio or (reverse) common consequence effects: in total, 24% of subjects exhibit none of these patterns, while the remaining 76% exhibit at least one. Note that each set of behavioral questions contains the possibility of a (reverse) common ratio effect and (reverse) common consequence effect. Since we have two sets of behavioral questions, in sum there are four possible observable effects.

In Figure 2, we then compare the average match rates across AI agents for subjects who exhibit none of these behavioral patterns, those who exhibit one or two, and those

who exhibit three or four. We find that match rates decline for both AI agents as subjects more frequently exhibit these patterns, suggesting that as people become more "behavioral", their choices become less predictable from past information, regardless of its source. At the same time, among subjects who never exhibit such patterns, the performance of Prompt-AI and Data-AI is statistically indistinguishable. In contrast, among subjects who exhibit at least one such pattern, Data-AI performs significantly better than Prompt-AI (paired t-tests: 1–2 violations, $p = 0.008$; 3–4 violations, $p = 0.011$).[13] In sum, these results suggest that revealed-preference information is more valuable for subjects with stronger behavioral patterns, as their choices are harder to predict from stated preferences alone.

Mechanisms. To examine whether the performance gap between the AI agents reflects subjects' difficulty in writing prompts or an AI-level inability to process stated-preference information, we conduct two analyses. First, we use subjects' written prompts from Part I to predict their choices in Part I, which reflects to some extent their abilities to write informative prompts.[14] If subjects' ability to write informative prompts is the primary driver of Prompt-AI's Part II performance, then Prompt-AI's match rates should be correlated across parts. This is indeed the case: the Pearson's correlation coefficient is 0.468 ($p < 0.001$). The relationship is robust after controlling for Part III demographic and survey characteristics, as shown in Table D.2 in Appendix D.

Second, for each subject, we use Claude Opus 4.5 to generate a preference description from the subject's 13 choices in Part I, using exactly the same information and instructions provided to subjects. We then use this AI-generated description to predict the subject's choices in Part II. We refer to this agent as **AutoPrompt-AI**. If AutoPrompt-AI performs better than Prompt-AI, this would suggest that the performance gap is driven mainly by the information available to the AI.

[13] These results are not driven by observable differences across groups. In Table D.1 of Appendix D, we regress match rate on an AI-model indicator (Prompt-AI vs. Data-AI), the number of behavioral patterns exhibited, and their interaction, controlling for Part III demographics and survey characteristics while clustering standard errors at the subject level. Consistent with the stylized facts above, a higher number of behavioral patterns is associated with a lower match rate, and this negative relationship is larger in magnitude for Prompt-AI than for Data-AI. The only other controls that are significant in the regression are age and IQ measures, both of which are positively correlated with higher match rates.

[14] On average, the resulting match rate is 69.4%, which is statistically indistinguishable from the 70.8% achieved by **Prompt-AI** in Part II (paired t-test, $t = 0.93$, $p = 0.35$).

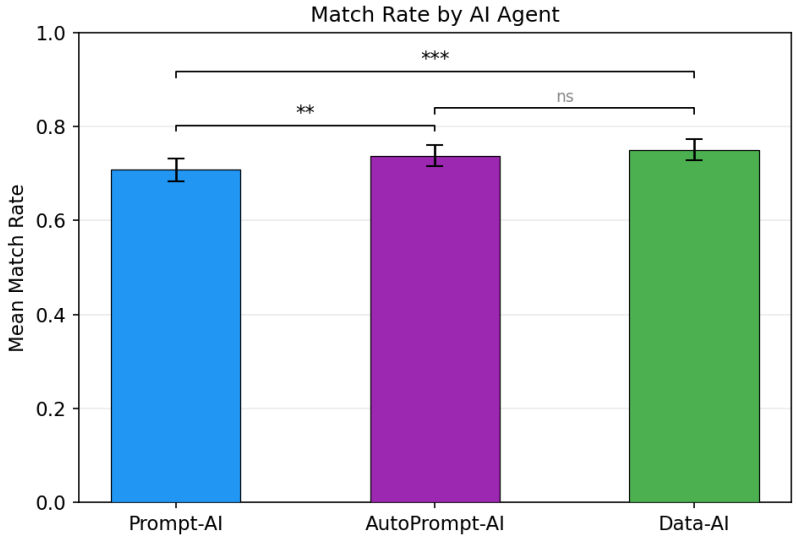

FIGURE 3: Comparison of Prompt-AI, Data-AI, and AutoPrompt-AI match rates ($N = 147$). AutoPrompt-AI (purple) uses a preference description auto-generated by Claude from the subject's Part I choices. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing AI agents. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The results are consistent with this hypothesis. In Figure 3, we compare the mean match rates of Prompt-AI, AutoPrompt-AI, and Data-AI. Overall, AutoPrompt-AI achieves an average match rate of 74%, which is significantly higher than that of Prompt-AI (paired t-test: $t = 2.443$, $p = 0.016$; Wilcoxon signed-rank: $p = 0.010$; $N = 147$) and statistically indistinguishable from that of Data-AI (paired t-test: $t = -1.381$, $p = 0.169$; Wilcoxon signed-rank: $p = 0.274$; $N = 147$). Taken together, our analyses suggest that much of the gap reflects subjects' difficulty in articulating informative prompts.

Result 1: *On average, AI agents perform better when given revealed preferences than when given stated preferences. This gap reflects humans' difficulty in articulating preferences in text, despite the fact that their choices are informative.*

4.2 Delegation

We now examine subjects' delegation decisions. Of the 147 subjects, 87 (59%) delegated to Data-AI. That is, they chose to send their past choice data to the AI rather

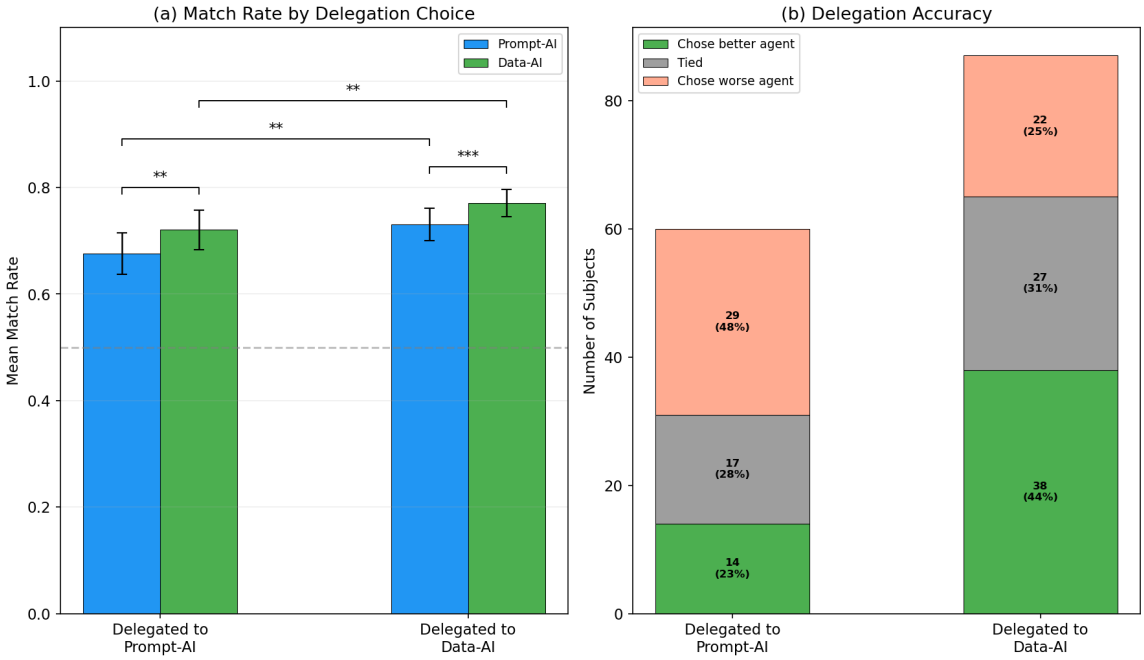

FIGURE 4: Delegation choice and accuracy ($N = 147$). *Panel (a)*: Mean Prompt-AI (blue) and Data-AI (green) match rates conditional on each subject's delegation choice. *Panel (b)*: Fraction of subjects in each delegation group who chose the ex-post better agent (green), were tied (grey), or chose the ex-post worse agent (salmon). Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing Prompt-AI to Data-AI within each delegation group or t-tests comparing an AI agent's match rate across groups. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

than their written prompts. This suggests an aggregate, though heterogeneous, preference for Data-AI.

Figure 4 assesses whether subjects correctly select the AI type that ultimately performs better on their behalf. Panel (a) plots the AI match rate conditional on subjects' delegation choices. Overall, Data-AI always achieves a higher match rate than Prompt-AI regardless of subjects' delegation decisions (paired t-test, $t = 3.580$, $p < 0.001$). This implies that, on average, those who delegate to Prompt-AI are delegating to the wrong type of AI agent. In addition, regardless of the AI type, subjects who delegate to Data-AI have a higher match rate than those who delegate to Prompt-AI (Data-AI: t-test, $p = 0.026$; Prompt-AI: t-test, $p = 0.030$). This suggests that there is heterogeneity in subjects' response quality, that is correlated with their delegation decisions.

Panel (b) shows ex-post alignment between delegation choices and realized performance for each subject. Conditional on the subject's delegation choice, we classify the chosen agent as better if it achieves a strictly higher match rate than its counterpart, and as worse if it achieves a strictly lower match rate. The remaining subjects have equivalent match rates for both AI agents. Overall, approximately 65% of subjects make a weakly better delegation decision. However, subjects who delegate to Data-AI choose the better agent more often than those who delegate to Prompt-AI. This is primarily because conditional on choosing Prompt-AI, whether Prompt-AI is actually a weakly better agent is approximately equivalent to a coin flip; in contrast, conditional on choosing Data-AI, Data-AI is the weakly better agent 75% of the time. Thus, while some subjects could potentially gain from delegating to Prompt-AI instead of Data-AI, a larger proportion—nearly half of those who delegated to Prompt-AI—would be strictly better off delegating to Data-AI instead (Fisher's exact test, $p = 0.001$).

Result 2: *More subjects prefer to provide revealed preference than stated preference to an AI agent. For the subjects who choose to provide stated preference information, nearly half would be strictly better off by providing revealed preference information instead.*

Beliefs. Next, we investigate how subjects' delegation decisions correspond to their beliefs about the match rate of different types of AI. Recall that each subject is asked to guess how many of the 13 lottery choices in Part II each AI agent would correctly predict. We begin by examining whether subjects' delegation choices are consistent with these beliefs. Overall, 85% of subjects act in a manner weakly consistent with their expressed beliefs, indicating that delegation decisions broadly reflect perceived AI performance. However, as Figure 5 shows, these beliefs are often too extreme and, for subjects who delegate to Prompt-AI, can even be incorrect in direction. Conditional on the delegation decision, on average, subjects substantially overestimate the true absolute difference in match rate between the two AI agents, and these discrepancies are statistically significant in both groups (paired t-tests; both $p < 0.001$). In particular, subjects who delegate to Prompt-AI believe it outperforms Data-AI by 7.3 percentage points, when in fact Data-AI outperforms Prompt-AI by 4.1 percentage points for this group. Moreover, this actual Data-AI advantage is similar among

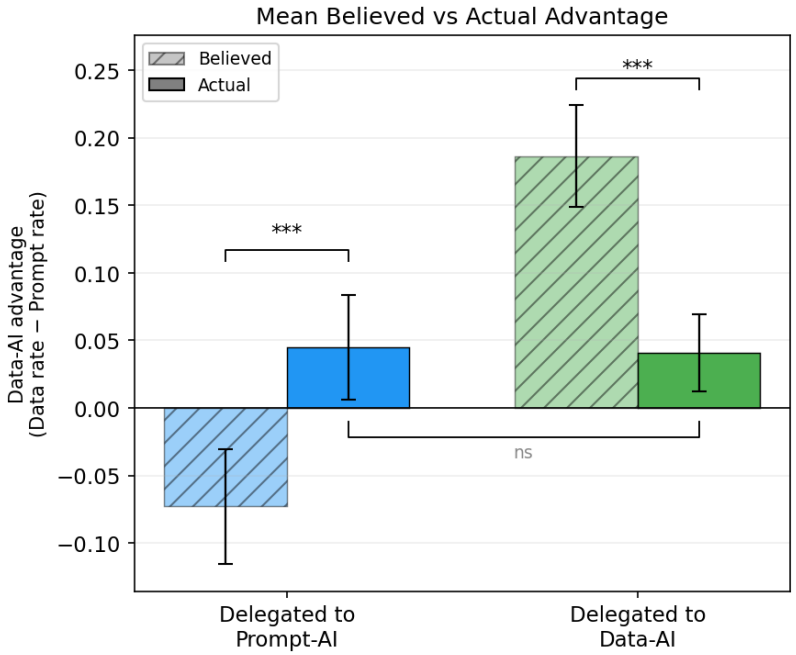

FIGURE 5: Mean believed and actual Data-AI advantage by delegation group ($N = 147$). Lighter bars show the mean believed advantage (guessed Data-AI match rate minus guessed Prompt-AI match rate); darker bars show the actual realized advantage. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Within-group brackets report paired t-tests comparing believed to actual advantage; the lower bracket compares the actual Data-AI advantage across delegation groups. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

subjects who delegate to Data-AI (4.5 percentage points; t-test: $p = 0.39$). Thus, while delegation decisions largely reflect beliefs, those beliefs are often wrong in both direction and magnitude.

Finally, as suggestive evidence, we also examine subjects' free-response justifications for their delegation choices. We use Claude Haiku 4.5 to label three dimensions of these responses. First, we ask whether the response mentions reasons other than optimization; only 12% of subjects do so, suggesting that non-optimization concerns play a minor role in their delegation decisions. Second, we check whether a response refers to either AI ability or the subject's own ability to communicate. Overall, 77% of subjects mention AI agent ability, while only 50% mention their own ability. This suggests that subjects focus more on the ability of the AI agent rather than their own, which in turn could contribute to their skewed beliefs regarding AI agents' predictive performance.

4.3 Combining Revealed and Stated Preferences

In this section we examine what happens when an AI agent is simultaneously given *both* subjects' written prompt and their Part I choice data—an agent we call **Both-AI**. This is analogous to "few-shot" prompting where the AI receives both a natural-language description of preferences and a few concrete choice examples. If the two sources of information are complementary, Both-AI should outperform agents given either source alone. If, instead, the two sources are substitutable, then our earlier results imply that revealed preference should serve as the primary source of information, since it delivers higher predictive accuracy.

In fact, Both-AI does not outperform both Data-AI and Prompt-AI alone: its mean match rate is 72%, which is significantly lower than that of Data-AI alone (75%; paired t-test: $t = -2.58$, $p = 0.011$) and only marginally higher than that of Prompt-AI alone (71%; paired t-test: $t = 1.81$, $p = 0.072$). We then investigate the source of this gap and find that it largely arises from conflicts between these two information sources. Figure 6 shows the outcome of each subject-question pair for Both-AI relative to the predictions of Prompt-AI and Data-AI. Prompt-AI and Data-AI make the same prediction in 75% of subject-question pairs throughout our sample. In these cases, Both-AI follows the common prediction 98% of the time, leaving little room for this subset to explain Both-AI's middling overall match rate. In the remaining 25% of subject-question pairs, where Prompt-AI and Data-AI predict *different* choices, Both-AI follows Prompt-AI 66% of the time and Data-AI only 34% of the time.

The above pattern stands in stark contrast to the asymmetric accuracy of stated and revealed preferences: in both contingencies, Data-AI is more accurate, as shown in the two terminal nodes on the right-hand side of Figure 6. The performance gap therefore arises primarily when stated and revealed preferences conflict. More broadly, the two sources do not appear to be complementary: giving the AI both does not improve predictive accuracy, and when they conflict, it places more weight on the less accurate stated-preference information. From an AI safety perspective, this may be normatively appealing, as the agent prioritizes explicitly stated preferences over implicitly revealed preferences. At the same time, because choices better reflect human preferences in our setting, this priority may also come at the cost of

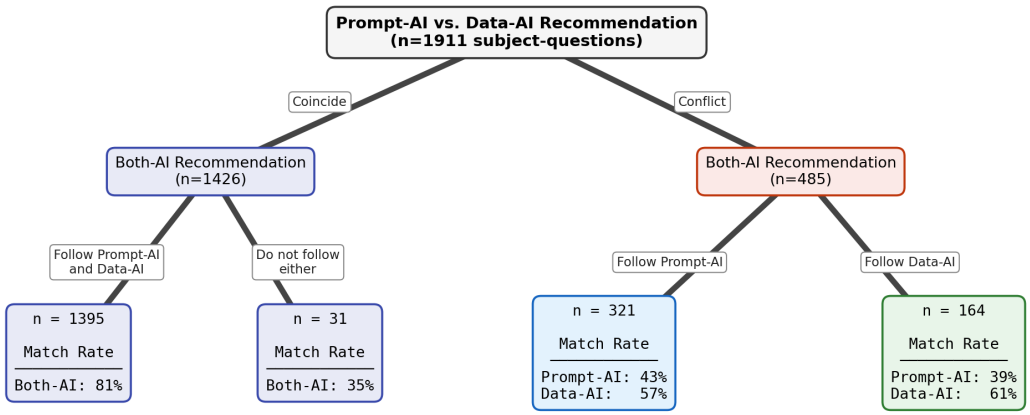

FIGURE 6: Decision tree partitioning all subject-question pairs by whether Prompt-AI and Data-AI agree (*Coincide*) or disagree (*Conflict*), and by which agent Both-AI follows in each case. Terminal nodes report the number of subject-question pairs, and the match rates of Prompt-AI and Data-AI with the subject's own choice ($N = 147$).

greater misalignment.

Result 3: *When stated and revealed preferences provide conflicting standalone predictions, AI agents who have access to both information sources largely listen to stated preference despite its lower accuracy.*

4.4 Robustness: GPT-5.4

As a robustness check, we additionally examine whether our results depend on the choice of Claude Opus 4.5 as the AI agent. While Claude Opus 4.5 is a frontier model at the time of writing, different AI models are known to behave differently in part because they vary in architecture, training data, and post-training or alignment methods (Zhao et al., 2023; Ouyang et al., 2025). We therefore replicate our analysis using GPT-5.4 (`gpt-5.4`), another frontier model. A natural hypothesis is that differences in post-training methods—for example, Anthropic's emphasis on constitutional-style alignment (Bai et al., 2022) and OpenAI's use of reinforcement learning from human feedback (Ouyang et al., 2022) may affect how models interpret instructions and infer preferences from observed choices. This analysis thus allows us to assess whether our main findings capture a broader feature of AI alignment with human principals, rather than an artifact of one particular model.

Appendix E shows the replications of the analysis using GPT-5.4.[15] The results are broadly similar. Overall, Data-AI still outperforms Prompt-AI, although the gap between the two agents becomes smaller. We highlight two main differences in the results. First, under GPT, AutoPrompt-AI no longer outperforms Prompt-AI. This appears to be driven by the substantially lower quality of the prompts generated by GPT. Consistent with this evidence, when GPT is asked to make predictions using Claude generated prompts, the results are the same as before (see Figure E.4). Second, under GPT, Both-AI performs approximately the same as Data-AI. This differs from our results with Claude, where Both-AI performs significantly worse than Data-AI. To understand this improvement across different AI models, we again examine questions in which stated and revealed preferences conflict and find that GPT follows Data-AI more often than Claude does (48% versus 34%). This pattern is consistent with differences in training methods, as Claude's constitutional-AI training may induce stricter adherence to stated preferences (Bai et al., 2022). Despite this improvement, GPT's Both-AI match rate remains far from perfect on conflicting questions, and performs only marginally better than random guessing. Overall, our results remain both qualitatively and quantitatively robust to GPT-5.4, suggesting that they extend beyond a singular AI model.

5 Conclusion

In this paper, we investigate whether a stated or revealed preference approach better mitigates the principal-agent problem caused by human principals' difficulty in articulating their preferences to AI agents. Using an incentivized online experiment, we find that human subjects communicate their preferences more effectively through revealed preference than through stated preference: on average, AI agents better predict future human choices when given revealed-preference data. At the same time, we also identify two important limitations that impact the benefits of revealed-preference information. First, subjects often fail to use revealed-preference-based agents even when doing so would benefit them; instead, they opt for less predictive AI agents based on stated preference. Second, when AI agents are given both stated and revealed preferences, they largely default to stated preferences whenever the two conflict. This

[15]Note that, in the experimental design, subjects were explicitly informed that the AI model used for payment was Claude. This may have influenced the instructions they wrote, their delegation decisions, and their beliefs. Accordingly, the GPT analysis should be viewed as exploratory.

means that merely adding revealed preference information on top of stated preference does not necessarily improve alignment. Thus, any mechanism for instantiating economic AI agents must address both of these issues.

Looking forward, we note that our environment of choice under risk explicitly removes known issues between stated and revealed preference, such as present bias (O'Donoghue and Rabin, 2015) and social desirability bias (Norwood and Lusk, 2011). In these domains, stated and revealed preferences are clearly not aligned, and this misalignment has implications for how AI agents should be instantiated (Kleinberg et al., 2024). We also assume that AI agents can effectively implement preferences they are given, which may not always hold. For example, Liang (2026) theoretically shows that in higher-dimensional matching problems, an infinite number of AI agent-recommended draws is less effective than a finite number of human-curated draws. Understanding the impact of stated versus revealed preferences under these types of conditions would provide for interesting avenues of future research.

References

Acemoglu, D. (2025): "The simple macroeconomics of AI," *Economic Policy*, 40, 13–58.

Agranov, M. and P. Ortoleva (2017): "Stochastic choice and preferences for randomization," *Journal of Political Economy*, 125, 40–68.

Allais, M. (1953): "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine," *Econometrica: journal of the Econometric Society*, 503–546.

Arrieta, G. and K. Nielsen (2025): "Procedural decision-making in the face of complexity," Tech. rep., Working Paper.

Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. (2022): "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*.

Blavatskyy, P., A. Ortmann, and V. Panchenko (2022): "On the experimental robustness of the Allais paradox," *American Economic Journal: Microeconomics*, 14, 143–163.

Blavatskyy, P., V. Panchenko, and A. Ortmann (2023): "How common is the common-ratio effect?" *Experimental Economics*, 26, 253–272.

Brynjolfsson, E., D. Li, and L. Raymond (2025): "Generative AI at work," *The Quarterly Journal of Economics*, 140, 889–942.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Chen, Y., T. X. Liu, Y. Shan, and S. Zhong (2023): "The emergence of economic rationality of GPT," *Proceedings of the National Academy of Sciences*, 120, e2316205120.

Condon, D. M. and W. Revelle (2014): "The international cognitive ability resource: Development and initial validation of a public-domain measure," *Intelligence*, 43, 52–64.

Dreyfuss, B. and R. Raux (2025): "Human Learning about AI," .

Fedyk, A., A. Kakhbod, P. Li, and U. Malmendier (2024): "Ai and perception biases in investments: An experimental study," *Available at SSRN*, 4787249.

Furnham, A. and H. C. Boo (2011): "A literature review of the anchoring effect," *The journal of socio-economics*, 40, 35–42.

Gabriel, I. (2020): "Artificial Intelligence, Values, and Alignment: I. Gabriel," *Minds and machines*, 30, 411–437.

Gans, J. S. (2026): "Optimal Use of Preferences in Artificial Intelligence Algorithms," *Working paper*.

Gneezy, U. and J. Potters (1997): "An experiment on risk taking and evaluation periods," *The quarterly journal of economics*, 112, 631–645.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann Jr (2003): "A very brief measure of the Big-Five personality domains," *Journal of Research in personality*, 37, 504–528.

Hadfield, G. K. and A. Koh (2025): "An Economy of AI Agents," Tech. rep., National Bureau of Economic Research, prepared for the NBER Handbook on the Economics of Transformative AI.

He, K., R. Shorrer, and M. Xia (2025): "Human Misperception of Generative-AI Alignment: A Laboratory Experiment," *arXiv preprint arXiv:2502.14708*.

Huang, D., F. Marmolejo-Cossío, E. Lock, and D. Parkes (2025a): "Accelerated preference elicitation with LLM-based proxies," *arXiv preprint arXiv:2501.14625*.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. (2025b): "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, 43, 1–55.

Imas, A., K. Lee, and S. Misra (2025): "Agentic Interactions," *Available at SSRN 5875162*.

Immorlica, N., B. Lucier, and A. Slivkins (2024): "Generative ai as economic agents," *ACM SIGecom Exchanges*, 22, 93–109.

Jackson, M. O., Q. Mei, S. W. Wang, Y. Xie, W. Yuan, S. Benzell, E. Brynjolfsson, C. F. Camerer, J. Evans, B. Jabarian, J. Kleinberg, J. Meng, S. Mullainathan, A. Ozdaglar, T. Pfeiffer, M. Tennenholtz, R. Willer, D. Yang, and T. Ye (2025): "AI Behavioral Science," *arXiv preprint arXiv:2509.13323*.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 363–391.

Kim, J., M. Kovach, K.-M. Lee, E. Shin, and H. Tzavellas (2024): "Learning to be homo economicus: Can an LLM learn preferences from choice," *arXiv preprint arXiv:2401.07345*.

Kleinberg, J., J. Ludwig, S. Mullainathan, and M. Raghavan (2024): "The inversion problem: Why algorithms should infer mental state and not just predict behavior," *Perspectives on Psychological Science*, 19, 827–838.

Lai, V., Z. Buçinca, N.-J. Akpinar, M. Houtti, H. B. Kang, K. Chian, N. Suh, and A. C. Williams (2026): "Users Mispredict Their Own Preferences for AI Writing Assistance," *arXiv preprint arXiv:2601.04461*.

Li, B. Z., A. Tamkin, N. Goodman, and J. Andreas (2023): "Eliciting human preferences with language models," *arXiv preprint arXiv:2310.11589*.

Liang, A. (2026): "Artificial Intelligence Clones," *arXiv preprint arXiv:2501.16996*.

McGranaghan, C., K. Nielsen, T. O'Donoghue, J. Somerville, and C. D. Sprenger (2026): "Connecting common ratio and common consequence preferences," *Journal of Political Economy*.

Meincke, L., E. Mollick, L. Mollick, and D. Shapiro (2025a): "Prompting Science Report 1: Prompt Engineering is Complicated and Contingent," Tech. rep., Generative AI Labs, The Wharton School of Business.

——— (2025b): "Prompting Science Report 2: The Decreasing Value of Chain of Thought in Prompting," Tech. rep., Generative AI Labs, The Wharton School of Business.

——— (2025c): "Prompting Science Report 3: I'll Pay You or I'll Kill You—But Will You Care?" Tech. rep., Generative AI Labs, The Wharton School of Business.

Norwood, F. B. and J. L. Lusk (2011): "Social desirability bias in real, hypothetical, and inferred valuation experiments," *American Journal of Agricultural Economics*, 93, 528–534.

O'Donoghue, T. and M. Rabin (2015): "Present bias: Lessons learned and to be learned," *American Economic Review*, 105, 273–279.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022): "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, 35, 27730–27744.

Ouyang, S., H. Yun, and X. Zheng (2025): "AI as decision-maker: ethics and risk preferences of LLMs," *Preprint at http://arxiv. org/abs/2406.01168*.

Rusak, G., B. S. Manning, and J. J. Horton (2025): "AI Agents Can Enable Superior Market Designs," *arXiv preprint*.

Shahidi, P., G. Rusak, B. S. Manning, A. Fradkin, and J. J. Horton (2025): "The Coasean Singularity? Demand, Supply, and Market Design with AI Agents," Tech. rep., National Bureau of Economic Research.

Spiliopoulos, L. and A. Ortmann (2018): "The BCD of response time analysis in experimental economics," *Experimental economics*, 21, 383–433.

Vafa, K., A. Rambachan, and S. Mullainathan (2024): "Do large language models perform the way people expect? measuring the human generalization function," in *Proceedings of the 41st International Conference on Machine Learning*, JMLR.org, ICML'24.

Zamfirescu-Pereira, J., R. Y. Wong, B. Hartmann, and Q. Yang (2023): "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. (2023): "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 1, 1–124.

Appendix A Lottery Pairs

Table A.1 summarizes all lottery questions and their corresponding parameters used in the experiment. Panel A shows all 13 lotteries used in Part I of the experiment, while Panel B shows all 13 lottery pairs used in Part II. Out of each 13 lottery pairs, there are 2 Easy questions, 5 Hard Questions, and 6 Behavioral questions.

TABLE A.1: Lottery Pairs by Category: Part I and Part II

Panel A: Part I	Lottery A	Lottery B	$\mathbb{E}[B] - \mathbb{E}[A]$
Easy			
FOSD	($3.00, 0.60; $1.00, 0.40)	($4.00, 0.70; $2.00, 0.30)	+1.20
Easy	($1.50, 0.50; $1.15, 0.50)	($1.55, 0.25; $0.25, 0.75)	−0.75
Hard			
Hard 1	($3.85, 0.25; $1.90, 0.75)	($4.70, 0.50; $0.80, 0.50)	+0.36
Hard 2	($4.50, 0.50; $0.50, 0.50)	($2.80, 0.25; $2.25, 0.50; $1.60, 0.25)	−0.28
Hard 3	($10.00, 0.25; $5.25, 0.25; $4.20, 0.25; $0.30, 0.25)	($6.75, 0.25; $5.85, 0.25; $3.00, 0.25; $2.70, 0.25)	−0.36
Hard 4	($4.05, 0.25; $2.55, 0.25; $1.50, 0.25; $0.65, 0.25)	($4.30, 0.25; $1.90, 0.25; $1.60, 0.25; $0.95, 0.25)	0.00
SOSD	($2.00, 1)	($4.00, 0.50)	0.00
Behavioral			
AB (p=0.8, r=0.25)	($4.00, 0.80)	($3.00, 1)	−0.20
AB′ (p=0.8, r=0.25)	($4.00, 0.20; $3.00, 0.75)	($3.00, 1)	−0.05
CD (p=0.8, r=0.25)	($4.00, 0.20)	($3.00, 0.25)	−0.05
AB (p=0.5, r=0.5)	($5.00, 0.50)	($2.00, 1)	−0.50
AB′ (p=0.5, r=0.5)	($5.00, 0.25; $2.00, 0.50)	($2.00, 1)	−0.25
CD (p=0.5, r=0.5)	($5.00, 0.25)	($2.00, 0.50)	−0.25
Panel B: Part II			
Easy			
FOSD	($2.00, 0.30; $1.00, 0.50)	($3.00, 0.50; $2.00, 0.20; $1.00, 0.30)	+1.10
Easy	($15.00, 0.50; $11.50, 0.50)	($15.50, 0.25; $2.50, 0.75)	−7.50
Hard			
Hard 1	($38.50, 0.25; $19.00, 0.75)	($47.00, 0.50; $8.00, 0.50)	+3.63
Hard 2	($5.00, 0.50)	($3.00, 0.25; $2.75, 0.50; $1.00, 0.25)	−0.13
Hard 3	($9.25, 0.25; $6.00, 0.25; $3.50, 0.25; $1.00, 0.25)	($7.30, 0.25; $5.50, 0.25; $3.50, 0.25; $2.00, 0.25)	−0.36
Hard 4	($3.75, 0.25; $3.00, 0.25; $2.00, 0.25)	($4.50, 0.25; $2.50, 0.25; $1.25, 0.25; $0.50, 0.25)	0.00
SOSD	($3.00, 0.25; $2.00, 0.50; $1.00, 0.25)	($3.00, 0.50; $1.00, 0.50)	0.00
Behavioral			
AB (p=10/11, r=0.11)	($3.00, 0.95)	($4.00, 0.91)	+0.79
AB′ (p=10/11, r=0.11)	($3.00, 0.95)	($4.00, 0.10; $3.00, 0.89)	+0.22
CD (p=10/11, r=0.11)	($3.00, 0.11)	($4.00, 0.10)	+0.07
AB (p=0.3, r=0.5)	($1.00, 0.95)	($5.00, 0.30)	+0.55
AB′ (p=0.3, r=0.5)	($1.00, 0.95)	($5.00, 0.15; $1.00, 0.50)	+0.30
CD (p=0.3, r=0.5)	($1.00, 0.50)	($5.00, 0.15)	+0.25

Notes: lotteries are represented as $(x_1, p_1; \ldots, x_n, p_n)$ where each outcome x_i occurs with probability $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. To simplify notation, we omit the null outcome of 0 and use (x, p) to denote the prospect $(x, p; 0, 1-p)$.

Appendix B AI System Prompts

This section reproduces the system prompts used for the Prompt-AI and Data-AI agents verbatim. Both agents additionally receive the 13 post-prompt lottery choices as a user message (see Section 3).

Prompt-AI System Prompt

```
You are helping a participant in an economics experiment make decisions
between lottery pairs. The participant has provided you with instructions
about their preferences.

Based on the participant's instructions below, you will make choices
between pairs of lotteries on their behalf. For each choice, respond
with ONLY "A" or "B" to indicate your selection.

PARTICIPANT'S INSTRUCTIONS:
"""
[PARTICIPANT'S WRITTEN PROMPT -- inserted verbatim]
"""

You will now be presented with lottery choices. After reasoning through
each choice, you MUST end your response with a JSON object containing
your choices in this exact format:

{"post_1": "X", "post_2": "X", "post_3": "X", "post_4": "X",
 "post_5": "X", "post_6": "X", "post_7": "X", "post_8": "X",
 "post_9": "X", "post_10": "X", "post_11": "X", "post_12": "X",
 "post_13": "X"}

Replace X with your actual choices. The JSON must be the last thing
in your response.
```

Data-AI System Prompt

```
You are an AI assistant helping a participant in an economics experiment
make decisions between lottery pairs.

Based on the participant's previous choices shown below, infer their
preferences and make similar choices for new lottery pairs.
```

```
You will now be presented with lottery choices. After reasoning through
each choice, you MUST end your response with a JSON object containing
your choices in this exact format:

{"post_1": "A", "post_2": "B", ..., "post_13": "A"}

Replace A/B with your actual choices. The JSON must be the last thing
in your response.

PARTICIPANT'S PREVIOUS CHOICES (Questions 1-13):
- Q1: Chose Lottery [X] (A: [description], B: [description])
- Q2: Chose Lottery [X] (A: [description], B: [description])
...
[All 13 pre-prompt choices with full lottery descriptions]
```

Appendix C EUT Structural Estimation

In this section, we describe our structural exercise. For each subject, we estimate a constant relative risk aversion (CRRA) expected utility model with a logit choice error using the 13 lottery choices in Part I. Given a lottery $L = (x_i, p_i)_{i=1}^n$, the expected utility is

$$EU(L;\,\rho) = \sum_{i=1}^{n} p_i\, u(x_i;\,\rho), \text{ where } u(x;\,\rho) = \begin{cases} \dfrac{(x+\omega)^{1-\rho}}{1-\rho} & \text{if } \rho \neq 1, \\ \ln(x+\omega) & \text{if } \rho = 1. \end{cases}$$

Note that ρ is the coefficient of relative risk aversion and $\omega = 0.01$ is a small wealth offset that ensures the utility function is defined at zero payoffs. For a lottery pair (A, B), the probability of lottery A being chosen by the agent with CRRA parameter ρ is thus

$$\mathbb{P}(A;\rho,\mu) = \frac{1}{1 + e^{-\mu[EU(A;\rho) - EU(B;\rho)]}},$$

where $\mu \geq 0$ is a precision (inverse noise) parameter. As $\mu \to \infty$, the agent deterministically chooses the higher-EU lottery. As $\mu \to 0$, their choices approach uniform random choice.

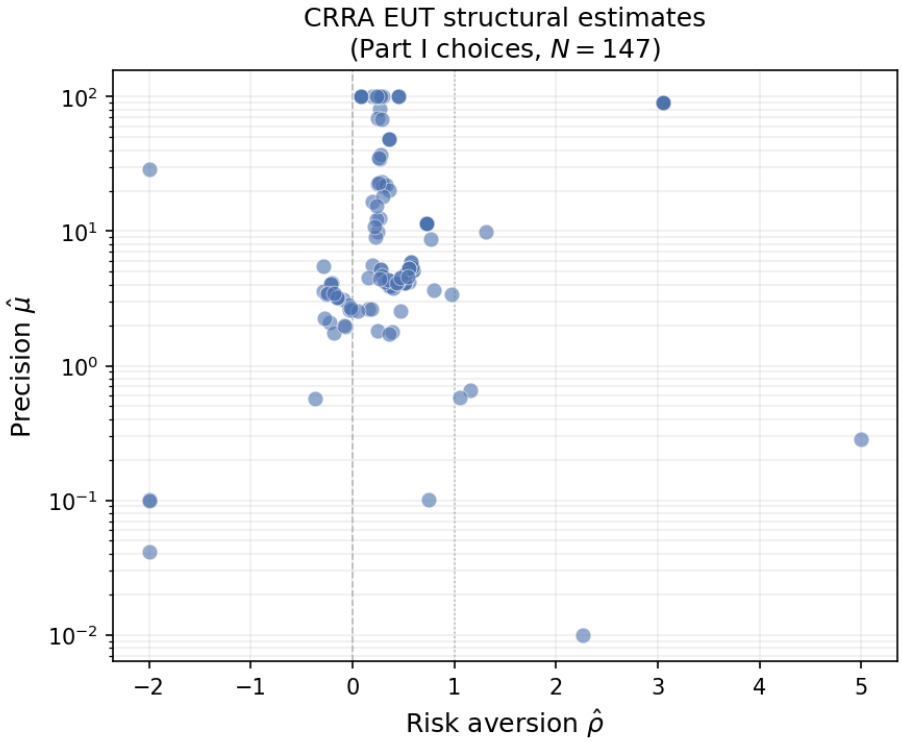

FIGURE C.1: Scatter plot of estimated $\hat{\mu}$ and $\hat{\rho}$ values per subject, based on their Part I choices. Dashed lines at $\rho = 0$ and $\rho = 1$ indicate risk neutrality and log utility, respectively.

For each subject, we observe 13 choices in Part I and use maximum likelihood to estimate their individual (ρ, μ). The log-likelihood is

$$\ell(\rho, \mu) = \sum_{k=1}^{13} \left[\mathbb{1}_{c_k = A_k} \ln \mathbb{P}(A_k; \rho, \mu) + \mathbb{1}_{c_k = B_k} \ln \mathbb{P}(B_k; \rho, \mu) \right],$$

where $c_k \in \{A_k, B_k\}$ is the observed choice from binary lottery menu k. We then optimize over the bounded parameter space $\rho \in [-2, 5]$, $\mu \in [0.01, 100]$ using L-BFGS-B, with a grid of 24 starting points ($\rho_0 \in \{-0.5, 0, 0.5, 1, 2, 3\}$, $\mu_0 \in \{0.1, 1, 5, 20\}$) to mitigate local optima. Finally, we retain the solution with the largest log-likelihood.

Figure C.1 shows a scatter plot of our estimated parameters for each subject. Most of our subjects (82%) are moderately risk averse, with estimated CRRA parameters ρ between zero and one. Among these subjects, the median ρ is 0.35, and the median precision parameter is 5.9. To put the consistency of our subject pool into perspective,

given their estimated parameters, the median number of questions (out of 13) they would choose with probability greater than 75% is eight. The questions for which the EUT model is less predictive are the Hard and Behavioral questions.

Appendix D Additional Heterogeneity Analysis

In this section, we include two regressions where we regress the match rate on (i) an indicator of whether the agentic regime is Data-AI or not, (ii) the number of behavioral effects observed in Part I, and (iii) their interaction term, while controlling for the following variables that have all been standardized to z-scores prior to estimation:

- *AI Comfort* and *Writing Comfort* are single 1–7 Likert items elicited at the end of the experiment: "How comfortable are you with using AI tools (e.g., ChatGPT, Claude, Copilot)?" and "How comfortable are you with writing instructions or explanations for others?", respectively, with endpoints labeled "Not at all comfortable" and "Very comfortable".

- *Impatience* is a single 1–7 Likert item: "How would you rate your level of impatience in general?", with endpoints labeled "Very patient" and "Very impatient".

- *IQ (Raven's)* is the number of correct answers on a 6-item matrix reasoning test from the International Cognitive Ability Resource (ICAR; Condon and Revelle, 2014) (score 0–6). Subjects receive 20¢ for each correct answer.

- *Measures of Overconfidence* consist of two types. The first is the difference between the subject's self-reported guess of their own score and their actual IQ test score, which captures absolute overconfidence. The second is the difference between the subject's self-reported guess of their own score and their performance relative to others, which captures relative overconfidence. Subjects receive 25¢ for each correct answer.

- *Risk Inv. (Simple)* and *Risk Inv. (Compound)* are the amounts (in tokens, from 0 to 100) invested in two variants of the investment task in Gneezy and Potters (1997). Each token is worth 0.5¢. Subjects can choose to invest their tokens into a safe asset, which leaves the invested amount unchanged, or in a risky asset, which multiplies the invested amount by 2.5 if the investment succeeds.

The simple task resolves uncertainty in a single coin flip whereas the compound task involves the same reduced distribution over payoffs, implemented by a compounded lottery: in stage one, there is a 25% chance of success, a 25% chance of failure, and a 50% chance of proceeding to the second stage. In the second stage, uncertainty is again resolved by a coin flip. Subjects are paid for each task based on their remaining balance of tokens.

- The "Big Five" personality traits (*Extraversion*, *Agreeableness*, *Conscientiousness*, *Emotional Stability*, *Openness*) are measured using the Ten-Item Personality Inventory (TIPI; Gosling et al., 2003), with each trait based on two 1–7 Likert items, one of which is reverse-coded, with endpoints labeled "Disagree strongly" and "Agree strongly". For example, *Extraversion* is measured as the average of "I see myself as extraverted, enthusiastic" and "I see myself as reserved, quiet" (reverse-coded).

As shown in Table D.1, for subjects who exhibit no behavioral effects, the performance of Data-AI and Prompt-AI is similar. Each additional behavioral effect is associated with a 2.2 percentage point decline in match rate for Data-AI, compared with a 4.4 percentage point decline for Prompt-AI. Subjects with higher IQ scores and older age also have higher match rates: for example, a one standard deviation increase in IQ score or age increases the match rate by four or two percentage points, respectively.

Table D.2 regresses the Prompt-AI match rate in Part II on its match rate in Part I, controlling for the variables listed above. A one percentage point increase in the Part I match rate is associated with a 0.27 percentage point increase in the Part II match rate. As before, Part II match rate is negatively correlated with behavioral effects and positively correlated with IQ and age. Relative overconfidence is also negatively correlated with the Part II match rate.

TABLE D.1: OLS: Match Rate (Prompt-AI and Data-AI)

	Match rate
Data-AI (vs Prompt-AI)	0.008
	(0.016)
Behavioral Effects	-0.044***
	(0.010)
Data-AI × Behavioral Effects	0.022**
	(0.010)
AI Comfort	-0.011
	(0.011)
Writing Comfort	0.002
	(0.012)
Impatience	0.013
	(0.011)
IQ (Raven's)	0.040***
	(0.013)
Risk Inv. (Simple)	-0.002
	(0.017)
Risk Inv. (Compound)	0.018
	(0.015)
Overconfidence (abs.)	-0.011
	(0.013)
Overconfidence (rel.)	-0.015
	(0.013)
Age	0.021**
	(0.010)
Female	-0.007
	(0.011)
Constant	0.775
	(0.019)
Observations	292
Subjects	146
R^2	0.243

Notes: Standard errors clustered by subject are reported in parentheses. All demographic regressors are standardized as z-scores. Behavioral Effects is included as a count variable ranging from 0 to 4, indicating the number of behavioral patterns exhibited by a subject. Agreeableness, Conscientiousness, Emotional Stability, Extraversion, and Openness are included as controls (none of them are significant), but are not reported. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table D.2: OLS: Prompt-AI Match Rate Across Parts

	Prompt-AI Match Rate (Part II)
Prompt-AI Match Rate (Part I)	0.268***
	(0.069)
Behavioral Effects	-0.028**
	(0.012)
AI Comfort	-0.009
	(0.014)
Writing Comfort	0.001
	(0.015)
Impatience	-0.001
	(0.016)
IQ (Raven's)	0.044**
	(0.017)
Risk Inv. (Simple)	-0.008
	(0.027)
Risk Inv. (Compound)	0.033
	(0.024)
Overconfidence (abs.)	-0.014
	(0.019)
Overconfidence (rel.)	-0.032**
	(0.016)
Age	0.027**
	(0.013)
Female	-0.004
	(0.014)
Constant	0.563***
	(0.060)
Observations	146
R^2	0.411

Notes: Standard errors clustered by subject are reported in parentheses. All demographic regressors are standardized as z-scores. Behavioral Effects is included as a count variable ranging from 0 to 4, indicating the number of behavioral patterns exhibited by a subject. Agreeableness, Conscientiousness, Emotional Stability, Extraversion, and Openness are included as controls (only Agreeableness is significant, at the 10% level), but are not reported. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix E GPT-5.4 Analysis

In this section, we replicate the main analysis using GPT-5.4 (`gpt-5.4`) as the decision model instead of Claude Opus 4.5. All experimental conditions and subject samples are identical to those in the main text ($N = 147$).

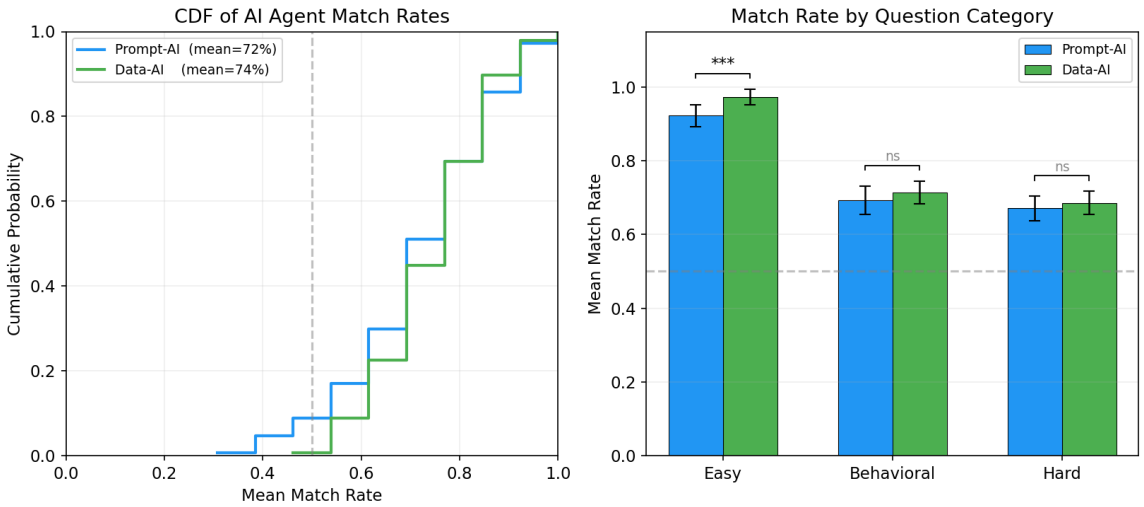

FIGURE E.1: Comparison of Prompt-AI and Data-AI match rates using GPT-5.4 ($N = 147$). *Panel (a)*: Empirical CDF of per-subject match rates for Prompt-AI (blue) and Data-AI (green); the dashed vertical line marks 50%. *Panel (b)*: Mean match rates by question category (Easy, Behavioral, Hard); the dashed horizontal line marks 50%. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing Data-AI to Prompt-AI within each category. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

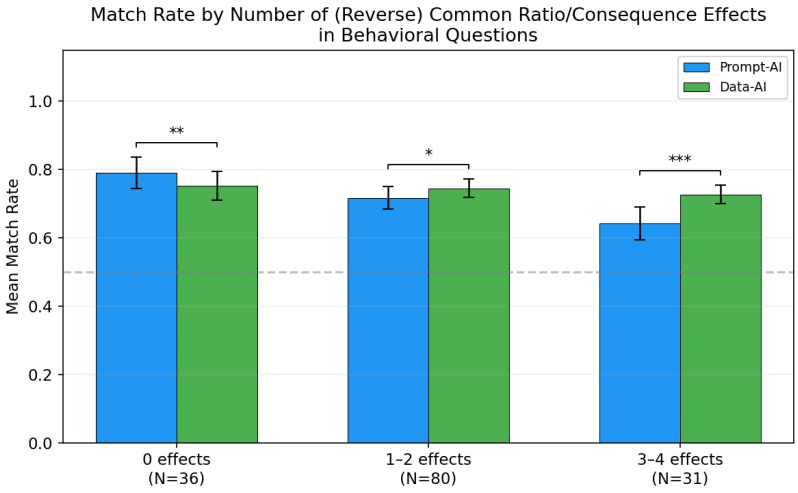

FIGURE E.2: Mean GPT-5.4 match rates by a subject's number of behavioral effects observed in the Behavioral questions; the dashed horizontal line marks 50%. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing Data-AI to Prompt-AI within each group of subjects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

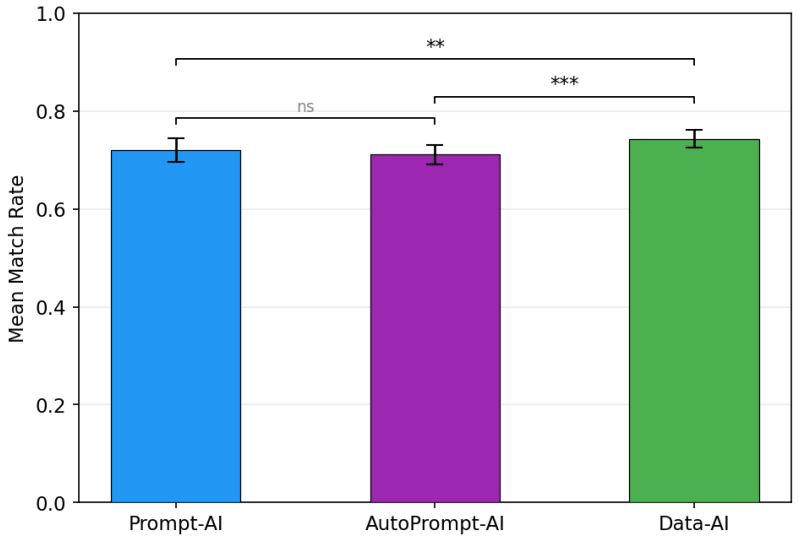

FIGURE E.3: Comparison of Prompt-AI, Data-AI, and AutoPrompt-AI match rates using GPT-5.4 ($N = 147$). AutoPrompt-AI (purple) uses a preference description auto-generated by GPT-5.4 from the subject's Part I choices. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing AI agents. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

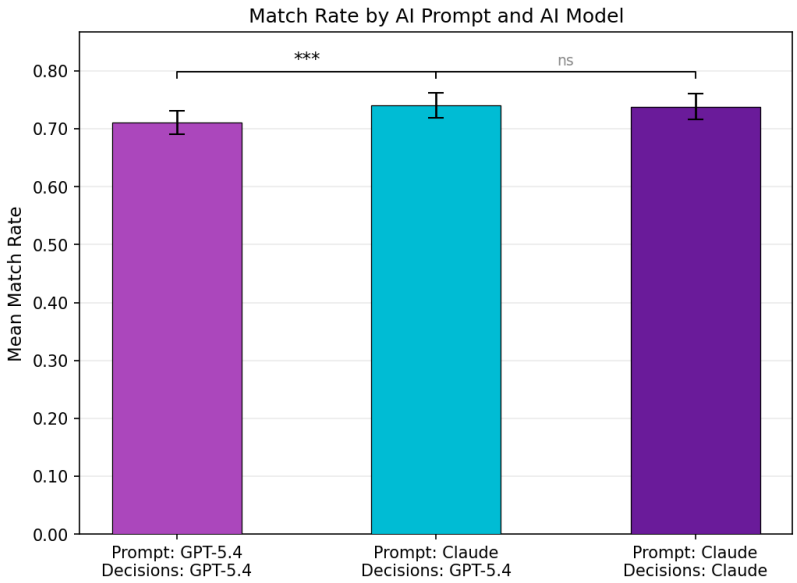

FIGURE E.4: Comparison of AutoPrompt-AI match rates using prompts generated by Claude and GPT-5.4, passed through Claude or GPT-5.4 ($N = 147$). Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing AI agents. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

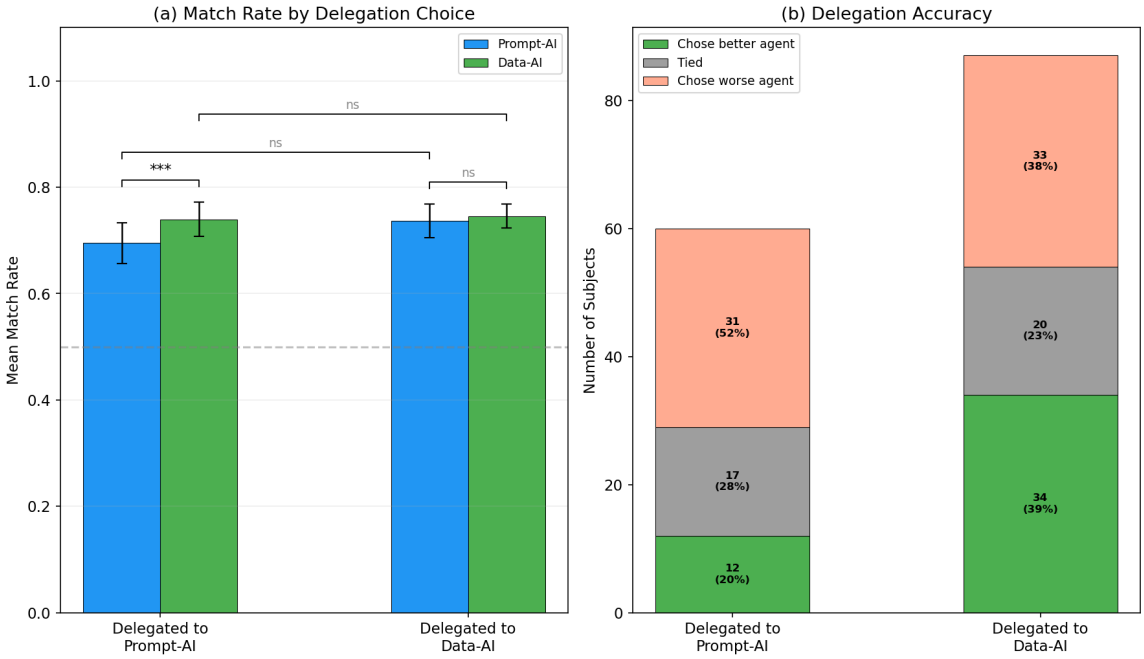

FIGURE E.5: Delegation choice and GPT-5.4 accuracy ($N = 147$). *Panel (a)*: Mean Prompt-AI (blue) and Data-AI (green) match rates conditional on each subject's delegation choice. *Panel (b)*: Fraction of subjects in each delegation group who chose the ex-post better agent (green), were tied (grey), or chose the ex-post worse agent (salmon). Error bars show 95% confidence intervals with standard errors clustered at the subject level. Stars denote paired t-tests comparing Prompt-AI to Data-AI within each delegation group or t-tests comparing an AI agent's match rate across groups. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

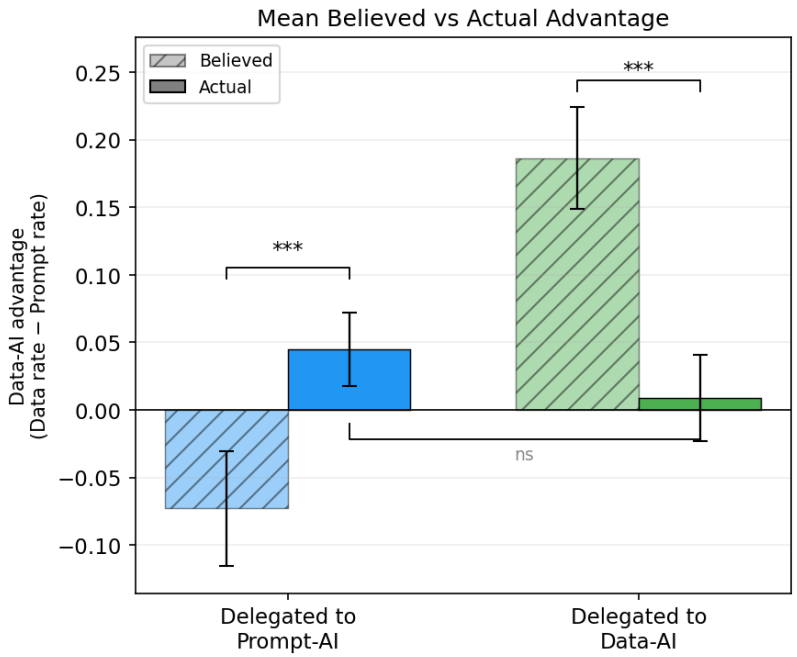

FIGURE E.6: Mean believed and actual Data-AI advantage by delegation group, GPT-5.4 ($N = 147$). Lighter bars show the mean believed advantage (guessed Data-AI match rate minus guessed Prompt-AI match rate); darker bars show the actual realized advantage. Error bars show 95% confidence intervals with standard errors clustered at the subject level. Within-group brackets report paired t-tests comparing believed to actual advantage; the lower bracket compares the actual Data-AI advantage across delegation groups. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

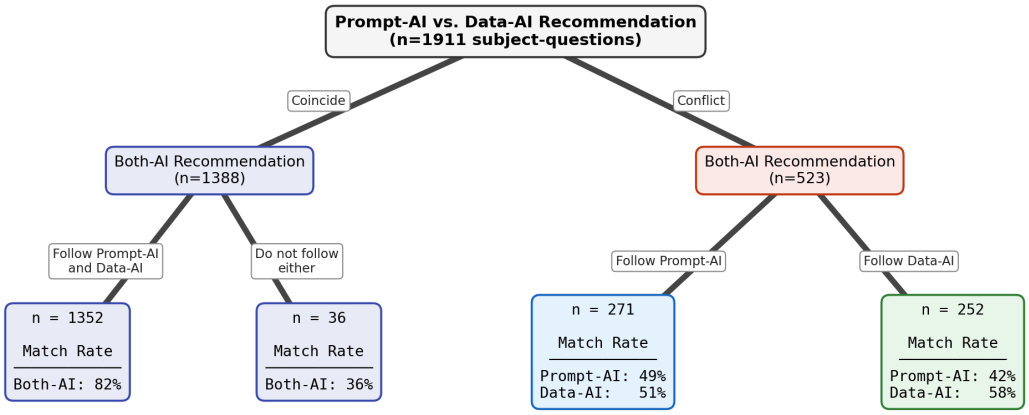

FIGURE E.7: Decision tree partitioning all subject-question pairs by whether Prompt-AI and Data-AI agree (*Coincide*) or disagree (*Conflict*), and by which agent Both-AI follows in each case, using GPT-5.4. Terminal nodes report the number of subject-question pairs, and the match rates of Prompt-AI and Data-AI with the subject's own choice ($N = 147$).