Human Learning about AI

Bnaya Dreyfuss Raphaël Raux

Harvard University MIT Sloan

First version: June 8, 2024

This version: February 3, 2026

Abstract

We study how people form expectations about the performance of artificial intelligence (AI) and consequences for AI adoption. Our main hypothesis is that people rely on human-relevant task features when evaluating AI, treating AI failures on human-easy tasks, and successes on human-difficult tasks, as highly informative of its overall performance. In lab experiments, we show that projection of human difficulty onto AI predictably distorts subjects' beliefs and can lead to suboptimal adoption, as failing human-easy tasks need not imply poor overall performance for AI. We find evidence for projection in a field experiment with an AI giving parenting advice. Potential users strongly infer from answers that are equally uninformative but less humanly-similar to expected answers, significantly reducing trust and future engagement. Our results suggest AI "anthropomorphism" can backfire by increasing projection and de-aligning people's expectations and AI performance.

Corresponding author: raphraux@mit.edu. We thank Katie Coffman, Benjamin Enke, Jesse Shapiro, and David Yang for precious guidance throughout this project. We benefited from helpful discussions with Isaiah Andrews, Leonardo D'Amico, David Laibson, Dev Patel, Matthew Rabin, Josh Schwartzstein, as well as seminar participants at Harvard. Finally, we are particularly thankful to Emily Oster and Alex Tryon from ParentData.org and Dewey Labs for access to chatbot conversation data. All experiments were approved by Harvard IRB (Protocol ID 23-0588), and separately pre-registered at AEA RCT Registry (ID 0012253; 0012622; 0014039; 0014086). Mathematical tasks were obtained from TIMSS: © 2024 International Association for the Evaluation of Educational Achievement (IEA); TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. This project was financed by support from both the Chae Family Economics Fund and the Institute for Quantitative Social Science at Harvard.

> The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard.
>
> Steven Pinker, *The Language Instinct*

1 Introduction

Recent AI systems are remarkably versatile, but their performance is often "jagged": they can excel on some tasks while failing on closely related ones (Gans, 2026). Misjudging where AI is reliable has important economic consequences: highly trained consultants incur net productivity losses when they trust AI on tasks outside its competence (Dell'Acqua et al., 2023), and incorrect clinical recommendations can steer radiologists toward erroneous diagnoses or follow-up decisions (Bernstein et al., 2023; Yu et al., 2024).

But we increasingly encounter this jagged technology in *anthropomorphic* form: user-facing AIs are named "assistants" that converse in the first person, use human turn-taking cues, and add human modalities like voice and personalization.[1] If we encounter a jagged technology through a human interface, are we likely to evaluate its performance with human yardsticks—and to be systematically misled?

This paper studies a simple hypothesis: people evaluate AI using the same inferential shortcuts they use for humans—a phenomenon we call *Human Projection* (HP). They take as informative for AI the cues that diagnose human competence, such as task difficulty or whether a mistake appears reasonable to humans. They use these cues to infer an overall notion of "ability" from sparse feedback and generalize their assessment across tasks, shaping adoption choices. This structure is powerful but becomes misspecified when what is hard, easy, or reasonable differs between humans and machines.

We develop a formal framework for HP that yields sharp predictions for priors and belief updating. We then embed HP in a Berk–Nash adoption model in which delegation

[1] Examples of new anthropomorphic features include OpenAIs ChatGPT voice and "personalities", Anthropic's Claude voice and personalization, xAI's Grok companions, Microsoft's Copilot personalization, or Google's Gemini personal intelligence.

choices select which performance signals are observed. We show HP induces *all-or-nothing adoption* in equilibrium. Partial adoption is not sustained even when AI is only superior on some tasks, because feedback in one task shifts beliefs about all other tasks. As a result, the same projection mechanism can generate *both* over- and under-adoption relative to the task-by-task optimum.

We test these predictions in a series of pre-registered experiments. In a lab study on beliefs in performance, we find strong projection of human task difficulty: priors fall steeply with difficulty, and updating is stronger after easy failures and hard successes. Because difficulty is only weakly informative about AI performance in our setting, this yields systematic miscalibration—overestimating AI on human-easy tasks and underestimating it on human-hard tasks. In a repeated-delegation experiment mirroring the adoption model, we vary HP by stripping human-like cues and presenting the same AI as a "black box." Over-adoption falls sharply, consistent with weaker projection of a single-index ability, limiting generalizations across tasks. In a framed field experiment with a parenting-advice chatbot, we find that *conditional on observing a (rare) mistake*, continued engagement drops sharply when the mistake is less humanly "reasonable"—i.e., less similar to a helpful reply. This pattern, consistent with the projection of human similarity onto AI outputs, leads users to disengage from an otherwise helpful tool.

Human Projection framework. We begin in Section 2 by formalizing Human Projection in a setting where a principal forecasts an agent's binary performance across a domain of tasks and learns from perfectly observed feedback. Human Projection has two components. *Feature projection* means treating human-relevant task features as informative for AI performance—namely, projecting human difficulty onto AI difficulty. *Ability–Model projection* means interpreting outcomes through a human "ability model": success is increasing in a unidimensional latent ability and decreasing in difficulty, and signals satisfy a diagnosticity property (MLRP) so that hard successes and easy failures are especially informative about ability. HP is not a bias *per se*, but rather a disciplined way to map features into priors and to extrapolate from limited feedback. It yields systematic misspecification when either projection component is misguided—when human difficulty is weakly informative for AI, or when AI performance is "jagged" such that it

does not admit the same single-index ability structure that people implicitly apply. The framework yields two testable predictions for beliefs when people project human difficulty onto AI. First, prior beliefs in success should decline with human difficulty, even for AI. Second, because outcomes are interpreted through the Ability–Model, "hard successes" and "easy failures" are more informative of a respectively high and low ability level, moving beliefs more.

Effects on beliefs. We test these predictions in a pre-registered beliefs experiment ($n = 911$ for AI, $n = 222$ for humans; Section 4) on a new domain of 414 multiple-choice math items drawn from international standardized assessments. We first obtain an item-level population benchmark of *human difficulty* as the share correct in an incentivized online test with around 41 answers per item. We then evaluate ChatGPT 3.5 on the same items and find that item-level difficulty is only weakly informative about AI success (slope -0.07, s.e. 0.08, $R^2 = 0.002$). We document similarly weak human–AI difficulty alignment across frontier models and prompt variants (Section 3), consistent with recent findings on AI's "jagged" capability patterns (Xie et al., 2024; Li et al., 2025).

Yet, our experiment finds subjects' beliefs are strongly anthropomorphic. Priors about AI exhibit a steep human-difficulty gradient (OLS slope -0.31, s.e. 0.01, $R^2 = 0.12$; about half the human slope). Updating is significantly stronger after easy failures (vs. hard failures) and hard successes (vs. easy successes), and this effect increases with the subject's difficulty projection. Beliefs are systematically misspecified in our setting, overstating AI performance on human-easy tasks while understating it on human-hard tasks despite the model's jagged, non-human performance gradient.

Consequences for user field engagement. In Section 5 we take Human Projection to a field setting and study its behavioral consequences for engagement with a parenting-advice chatbot. Because "correctness" and "difficulty" are inherently subjective in advice settings, we focus on a human-relevant feature of *failures*: the *reasonableness* of a misunderstanding—whether the useless answer is a plausible, human-like miss (similar to what a helpful answer would have said) or an implausible one. We adapt the Human Projection framework to this setting by treating reasonableness (human similarity to a

useful answer) as the projected feature, yielding an analogue of "easy" versus "hard" failures: conditional on a failure (useless answer), less reasonable misunderstandings should trigger more negative updating and lower engagement. In a framed field experiment ($n = 905$) built from real conversations, we use matched failure pairs that are equally unhelpful and differ only in reasonableness. Among a sample of potential users, less reasonable failures generate substantially larger drops in beliefs and trust (about 80% and 75% larger, respectively). This maps into reduced subsequent engagement, measured both by a revealed-preference link choice (9.4 p.p. drop, $p = 0.005$) and by actual follow-through in chatbot logs (2.0 p.p. drop, $p = 0.045$). Interpreting these patterns through the lens of HP, users appear to treat the human reasonableness of its (rare) misunderstandings as highly diagnostic of overall performance, reducing engagement with an otherwise useful tool.

Effects on equilibrium adoption. We end in Section 6 by studying AI adoption in equilibrium, when AI signals are endogenous to delegation decisions. We model a principal observing feedback only on tasks they choose to delegate to AI, and interpreting it through the HP model. Projecting a single-index ability leads to a coarse conclusion: the principal infers that AI is either "better" or "worse" than humans in the domain, leading to *all-or-nothing* adoption in Berk–Nash equilibrium (Esponda and Pouzo, 2016). In this benchmark result, partial adoption is not sustained even when optimal under task-by-task evaluation. HP can thus generate distortions towards under-adoption (no adoption despite AI being superior on some tasks) or over-adoption (full adoption despite AI being inferior on some tasks). Because human-like cues are known to induce anthropomorphic reasoning (Chugunova and Sele, 2022), we predict that removing these cues weakens Human Projection and reduces all-or-nothing adoption—shifting choices toward partial adoption when AI's advantage is task-specific.

We test this prediction in an adoption experiment presented in Section 6. We construct two pools of problems, drawn from our domain of math items, such that humans outperform AI on the human-easy pool while AI outperforms humans on the human-hard pool. Thus, partial adoption (AI only on the hard pool) is optimal under task-by-task evaluation. Participants repeatedly delegate 30 problems from each pool to a human

or to AI, observing only the realized outcome of their chosen agent. They then make a one-shot, high-stakes adoption decision. We randomize whether the same underlying technology is framed anthropomorphically (a named assistant with a human-like icon, $n = 208$) or as a non-anthropomorphic "black box" ($n = 218$). The black-box framing significantly reduces the share of all-or-nothing adoption, primarily by reducing over-adoption ($\Delta = -0.146$, $p = 0.001$). Mechanism evidence points to weaker Ability–Model projection: across several proxies of cross-task generalization, the Black Box framing sharply reduces the coupling of AI beliefs across pools (e.g., easy–hard belief comovement falls from 0.69 to 0.32, $p < 0.001$), while effects on difficulty projection are small and less stable. Subjects appear to rely less on a single-index ability to extrapolate across task types and to treat performance as pool-specific.

Taken together, we formalize and document Human Projection: a tractable inference rule that maps sparse performance feedback and salient task features into broad generalizations about an agents competence. This heuristic becomes systematically misleading when the projected features are weakly informative for the target agent. AI is a natural setting: models are jagged and context-dependent, yet are presented in human-like ways that invite human yardsticks. Our framework clarifies when such cues distort beliefs and learning, and how those distortions translate into delegation and adoption.

Related literature. A growing behavioral-economics literature studies *mental models* as simplified—and often misspecified—representations that discipline inference and choice. Models, even when wrong, tend to persist as they shape data acquisition and interpretation (Gagnon-Bartsch et al., 2023; Fudenberg et al., 2024). Berk–Nash equilibrium (Esponda and Pouzo, 2016) provides a natural solution concept when decision makers optimize within a restricted—and potentially misspecified—model class (Lanzani, 2025; Samuelson and Steiner, 2025), consistent with evidence that people often apply simple models over richer but more complex ones (Kendall and Oprea, 2024). Important channels of misspecification include biased recall (Bordalo et al., 2023; Bohren et al., 2025), selective attention (Graeber, 2023; Bastianello and Imas, 2026), or improper coarsening (Graeber et al., 2025). They help rationalize under- and over-reaction to signals (Fan et

al., 2025; Augenblick et al., 2025; Bai et al., 2025) and understand persistent errors in settings such as risky forecasting (Esponda et al., 2024), financial decisions (Andre et al., 2024; Enke et al., 2024), or as technology adoption (Hanna et al., 2014). We add to this agenda by showing how a simple mental model—drawing on IRT principles (Lord and Novick, 2008)—affects beliefs and usage of new technology. Among mental models are "projection" heuristics, which distort perceptions of oneself (Loewenstein et al., 2003) and of others, shaping social and market learning (Gagnon-Bartsch and Rosato, 2024) and strategic behavior through second-order beliefs (Madarász et al., 2023). We bring projection to technology and show direct links to adoption and usage choices. Closest to our study is evidence that people generalize across tasks using human intuitions rather than true capability structure (Vafa et al., 2024), and systematically overstate the degree of alignment with human behavior (He et al., 2025).

AI's productivity effects have been found to be heterogeneous and context-dependent (Noy and Zhang, 2023; Brynjolfsson et al., 2025). A central reason is that effective adoption hinges on users' beliefs about *task-specific* capability. When performance is jagged across tasks, users misperceive where AI is reliable, generating misreliance and foregone gains in both lab settings (Chevrier et al., 2024) and high-stakes deployments (Dell'Acqua et al., 2023; Bernstein et al., 2023; Yu et al., 2024). Who benefits also depends on behavioral features of the user such as self-assessments of competence (Caplin et al., 2024), motivating designs that account for predictable biases and offer tailored assistance (Dreyfuss and Hoong, 2025). We offer a unified mechanism for misreliance and for the coexistence of algorithm *aversion* (Dietvorst et al., 2015; Dietvorst and Bharti, 2020) and *appreciation* (Logg et al., 2019): HP can induce either under- or over-adoption, depending on how AI's comparative advantage differs from what users project.

Finally, we connect to work on machine *anthropomorphism* (Reeves and Nass, 1996; Chugunova and Sele, 2022). One strand studies *determinants*: human-like design cues (e.g., voice, first-person language) increase anthropomorphic inferences and perceived competence (Cohn et al., 2024) and system trust (De Visser et al., 2016), with context-dependent effects (Schimmelpfennig et al., 2025). A second strand studies *consequences*: treating systems as social agents can raise customer spending (Igarashi et al., 2026),

human-like interaction (Schanke et al., 2021), and offset trust costs of AI disclosure (Xu et al., 2024), but also backfire via inflated expectations or anger (Crolic et al., 2022). We contribute by providing a simple structure to clarify how human-like cues may affect user behavior through systematic distortions in beliefs and inferences.

2 A Theoretical Framework for Human Learning

In this section we develop the Human Projection (HP) framework: a structure for human inferences on performance within a domain of tasks, projected on potentially non-human agents. HP decomposes into two distinct, complementary components: Ability–Model projection, in which the DM summarizes outcomes with a unidimensional latent ability, and difficulty projection, where they project the human-difficulty of tasks onto the agent. Resulting beliefs may be misspecified misspecified either because the projected human partition mismatches AIs performance-relevant partition, or because projected and true AI difficulty differ. Appendix A provides the full formal exposition.

2.1 Setup

We consider a principal trying to assess the performance of an agent i within a domain $\mathcal{T} = \{t_1, \dots t_K\}$ composed of K tasks or problems. Performance is stochastic: for each task $t_k \in \mathcal{T}$, the agent has a success rate $s_k \in [0,1]$. Overall performance is represented by the vector $\mathbf{s} = (s_1, \dots, s_K) \in [0,1]^K$, which is unknown to the principal.

We assume that the principal entertains the following mental model: the agent has a unidimensional type denoted $\theta \in \Theta \subseteq \mathbb{R}$, unknown to the principal. This type θ represents a latent variable of *ability* within the domain. Each task has a level of *difficulty* denoted by $\delta \in \Delta \subseteq \mathbb{R}$, which is known to the principal for humans. We denote the (human) difficulty of task t by $\delta^H(t)$, and the (true) AI difficulty by $\delta^A(t)$; these may differ (a human-difficult task need not be AI-difficult). A task with difficulty δ is denoted by t^δ. Unless specified otherwise, "difficulty" refers to human difficulty δ^H. The probability that agent $i \in \{H, A\}$ with ability θ succeeds in solving a given task t^δ is given by

$p^i : \Theta \times \Delta \to (0,1)$. Assumption 1 imposes structure on the mapping from ability and difficulty to success rates:

Assumption 1 (Ability–Model Projection)**.** *For each agent $i \in \{H, A\}$, $p^i(\theta, \delta)$ satisfies:*

1. ***Ability:*** *If $\theta' > \theta$, then $p^i(\theta', \delta) \geq p^i(\theta, \delta)$.*
2. ***Difficulty:*** *If $\delta' > \delta$, then $p^i(\theta, \delta) \geq p^i(\theta, \delta')$.*
3. ***MLRP:*** *If $\theta' > \theta$ and $\delta' > \delta$ then $\frac{p^i(\theta',\delta')}{p^i(\theta,\delta')} \geq \frac{p^i(\theta',\delta)}{p^i(\theta,\delta)}$ and $\frac{1-p^i(\theta',\delta')}{1-p^i(\theta,\delta')} \leq \frac{1-p^i(\theta',\delta)}{1-p^i(\theta,\delta)}$.*

The first two parts significantly reduce the complexity of the principal's updating problem by placing all features on a unidimensional scale: agents are ordered by ability, and tasks are ordered by difficulty. For MLRP, consider two agents of unequal ability levels, and take the ratio (high over low) of their success rates on tasks. This ratio increases with task difficulty under MLRP: as tasks become harder, both success rates decrease, but that of the high-ability agent decreases *slower*, thereby increasing the ratio. Conversely for failure rates: as difficulty increases, they increase for both agents, but slower for the high-ability one, thereby reducing the ratio. Intuitively, MLRP implies that a success on a hard task is more "diagnostic" of high ability than a success on an easier task, while a failure on an easy task is more diagnostic of low ability than a failure on a harder task.

These restrictions constitute the "Ability Model:" a mental model used to assess human performance within a domain. It is inspired by Item Response Theory (Lord and Novick, 2008), used to estimate students' proficiency on the basis of standardized test performance. Most modern tests, including TIMSS (source of our mathematical tasks in Section 3), estimate student proficiency using models which satisfy the above assumptions.[2]

The principal is trying to learn about agent ability θ^i, where agent $i \in \{H, A\}$ can be a human or an AI. They receive binary signals of agent performance on tasks—successes or failures—within domain $\mathcal{T}$, and update their prior G^i according to Bayes' rule. When

[2]In Item Response Theory, commonly used logistic and ogive functions, mapping ability and difficulty to a probability of a correct answer, satisfy all parts of Assumption 1. For an example see the TIMSS Technical Report.

evaluating AI, we assume the principal interprets AI outcomes through the same human-relevant domain structure and the associated unidimensional A–M within $\mathcal{T}$. We refer to this as *Ability–Model projection*. It is misspecified when AI performance within $\mathcal{T}$ is not well summarized by a single latent ability, or when the human task partition is a poor approximation to the AI-relevant partition.

We now introduce *difficulty projection*, a source of belief misspecification about AI. We assume the principal perfectly observes human difficulty δ^H but only imperfectly perceives AI difficulty:

Assumption 2 (Difficulty Projection)**.** *For a given problem $t \in \mathcal{T}$, the perceived AI difficulty is given by $\tilde{\delta}^A(t) = \lambda\delta^H(t) + (1-\lambda)\delta^A(t)$, where $\lambda \in [0,1]$ is the degree of projection.*

This assumption states that people evaluate AI difficulty based on human difficulty. We micro-found it in Appendix A with a simple cognitive imprecision model following Woodford (2020), where the principal receives a noisy signal of (true) AI difficulty, which they combine with their prior anchored around (true) human difficulty, leading to this "shrinkage" formula.

Human Projection. Together, Assumptions 1 and 2 constitute what we call *Human Projection* (HP): relying on human-relevant structure when forming beliefs about AI performance. HP has two components. *Ability–Model projection* applies the human A–M (and the human-relevant task partition) to AI within $\mathcal{T}$. *Difficulty projection* anchors perceived AI difficulty on δ^H, as captured by $\lambda > 0$. Misspecification can arise through either component, and is increasing in the mismatch between the projected human structure and AI's true performance structure.

The Ability Model in itself is not particularly restrictive: it assumes that instead of updating on success rates task by task, the principal updates on a single variable of ability, which influences all success rates. When the domain contains many tasks that are perceived as similar, this assumption is relatively weak. For example, it is natural to think of a single "math ability" that shapes performance across a broad set of math problems, from arithmetic to algebra. Within the model, both the mappings (p^A and p^H) and priors over ability (G^A and G^H) are allowed to be agent-specific. Priors may differ if

e.g., one has higher uncertainty about AIs, or thinks machines are generally better than humans in the relevant domain.

2.2 Experimental Predictions

We provide the heuristic statements of two predictions (one for priors and one for belief updating) which we experimentally test in Section 4.1. Formal statements and proofs are in Appendix A.

Prediction 1. *Under difficulty projection ($\lambda > 0$), expected success rates are decreasing in human difficulty for both humans and AI.*

In other words, prior beliefs in both human and AI performance are expected to be "sloping down" with the human difficulty of tasks. Prediction 1 follows directly from the assumptions that $p^i(.,.)$ is decreasing everywhere with difficulty, and that $\tilde{\delta}^A$ is increasing in δ^H.

Prediction 2. *Consider any two tasks, one human-easier and one human-harder: $\delta^H_{easy} < \delta^H_{hard}$. The principal observes performance on one of these tasks and forms beliefs about performance on other tasks. Under Human Projection, for any prior G and prediction task t:*

*(i) After observing a **failure**, posterior beliefs in performance are higher if the failure occurred on the human-harder task (relative to a failure on the human-easier task)*

*(ii) After observing a **success**, posterior beliefs in performance are higher if the success occurred on the human-harder task (relative to a success on the human-easier task)*

Prediction 2 states that after observing one binary signal of performance, the mean posterior success rate is *larger* if the task on which performance was observed was *harder*. To help build intuition for this result, consider first the case of an observed failure. The principal lowers their posterior on agent ability and this decrease is larger in the case of an "easy failure," because for easy tasks a failure is more diagnostic of a low ability level (compared to failures on hard tasks). Lower ability then reduces perceived success rates on all tasks within the domain: the principal's expectations of performance after an easy failure are relatively lower than after a hard failure. The same logic applies for successes:

succeeding on the harder task is more diagnostic of a high ability level than succeeding on an easy task, therefore further increases posterior beliefs. In Section 4.1 we test this hypothesis in terms of belief movement: (i) after a failure, beliefs in performance *decrease less* if the signal task was harder; (ii) after a success, beliefs in performance *increase more* if the signal task was harder.

3 Human vs GenAI Difficulty in Mathematics

Human Projection generates belief misspecification precisely when *actual* GenAI performance departs from the human-relevant structure that people use to form expectations. In this section, we document such departures in the domain of mathematics. We first introduce a benchmark of standardized tasks and a high-precision measure of *human* difficulty. We then present a case study (ChatGPT 3.5) where AI performance is uncorrelated with human difficulty in our domain, contrasting with the anthropomorphic expectations we document in our experiments (Sections 4 and 6). Finally, we discuss the generality of the human-AI difficulty misalignment, using evidence from more recent models on our benchmark and from the difficulty-alignment literature. Appendix B provides details on benchmark construction and performance data.

3.1 Benchmark of Tasks and Human Difficulty

We collect and manually re-transcribe released items from the *Trends in International Mathematics and Science Study* (TIMSS). These international standardized tests assess student proficiency at the 4^{th}-grade, 8^{th}-grade, and High School level. We focus on multiple-choice items, which have either 4 or 5 possible answers (A–E), yielding a binary and objectively graded outcome. We obtain a final dataset of 414 items (29% from 4^{th} grade, 58% from 8^{th} grade, and 13% from High School) spanning the usual range of mathematical topics.

This benchmark is well-suited for comparing human difficulty to AI performance. First, items are designed to measure mathematical proficiency using test-scaling method-

ologies consistent with the Ability Model. Second, question format is stable across items, and tasks span a broad range of topics and human difficulty. Third, many items are not easily accessible online and were obtained through direct request to the IEA, reducing the risk of benchmark contamination.

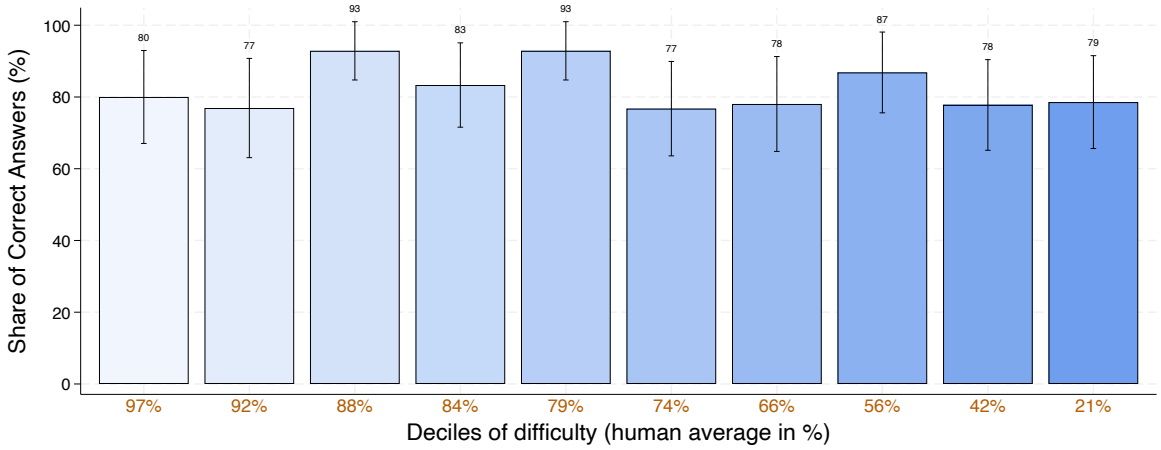
3.2 A Case Study: ChatGPT and the Human Difficulty Gradient

Human performance data and difficulty measure. We collect incentivized human performance data on our benchmark using adult participants from Prolific. The data come from two samples collected under highly similar test-like conditions (October 2023 and December 2023), which we pool. Participants are instructed to approach the questions as they would a real math test and to remain time-conscious.

We obtain high-precision item-level human performance: the average item is answered 41.5 times allowing us to measure "human difficulty" with low sampling noise at the item level. Following the standardized testing literature (Bachman, 1990), we define human task difficulty by the share incorrect: $\text{Difficulty}_t = 100 \times (1 - \text{ShareCorrect}_t)$, reported in percentage points. Hereafter, "difficulty" refers to this empirical measure of *human* difficulty. The average item difficulty is 30.3 and the median is 24; the most difficult item has difficulty 92, and the easiest has difficulty 0 (7 items).

ChatGPT performance. The GenAI performance used in our experiments comes from ChatGPT (3.5, August 2023 version), evaluated on all items under zero-shot text prompting. We classify an answer as correct if the model selects the correct answer key, and as incorrect otherwise. Figure 1 reports ChatGPT accuracy by deciles of human difficulty. Unlike human performance, which decreases sharply with difficulty by construction, ChatGPT accuracy is approximately flat across the human difficulty gradient: errors occur throughout the distribution rather than concentrating on items humans find hard. Human difficulty has negligible predictive power for ChatGPT correctness. Regressing the scaled indicator for a correct answer (scaled by 100) on the 0–100 human-difficulty index yields an estimated slope of -0.07 (s.e. 0.08, $p = 0.37$, $R^2 = 0.002$). This finding is robust to a more recent API-base replication (May 2025), which yields an even flatter

Figure 1: ChatGPT Performance by Human Difficulty Deciles



Notes: The figure plots ChatGPT (3.5) share of correct answers within each decile of human difficulty. Each decile contains between 40 and 42 items. Numbers under each bar report the average human success rate in the corresponding decile. Human difficulty is computed from an average of 41.5 answers per item.

mapping (slope -0.03, s.e. 0.09, $p = 0.73$, $R^2 = 0.0003$), with slightly lower average accuracy; Appendix Figure 8 reports the corresponding decile-accuracy plot.

3.3 Broader Evidence on Human–AI Difficulty Misalignment

Two complementary bodies of evidence suggest that the difficulty mismatch we document is not an idiosyncrasy of our setting: even highly capable models can exhibit difficulty and error patterns that are only weakly aligned with humans.

Evidence from our benchmark across models. We evaluate multiple advanced models on the same TIMSS item set and quantify "human-likeness" using item-level measures of difficulty alignment and answer-choice overlap (Appendix B, Table 5). The central pattern is stable across model families and prompt variants: *high mean accuracy does not imply a human-like difficulty ordering*. Rank and linear associations between human item easiness and model correctness are typically small, and human difficulty explains little of the item-level variation in model success (low R^2). At the same time, answer-choice overlap is substantial in the aggregate (models often select the human modal option), but it is much less consistent on mistakes: conditional on error, models do not reliably concentrate their probability mass on the same distractors that humans choose. Moreover,

"human-likeness" is sensitive to presentation and instructions: suppressing chain-of-thought (Wei et al., 2022) or switching from text to image prompts significantly affects both accuracy and the mapping between human and model success across items, implying that difficulty alignment is context-dependent rather than a monotone consequence of model capability.

External evidence of systematic misalignment. Recent computer-science work reaches similar conclusions through different approaches. One theme is *non-monotonicity and inconsistency*: LLMs can solve questions that appear complex while failing nearby variants that humans find straightforward, and scaling does not necessarily yield more *consistent* performance across related questions (Yang et al., 2024). Complementary benchmarks document substantial failure modes on tasks that are "easy" by human linguistic and reasoning standards across multiple domains beyond mathematics (Williams and Huckle, 2024). A second theme is *brittle generalizations*: changes to problems that appear trivial to humans (e.g., reversals) can sharply affect model correctness, and mitigating these effects is not straightforward (Berglund et al., 2023; Wang and Sun, 2025). A third theme is *difficulty miscalibration and inverse scaling*: stronger models can underperform on specific task classes (McKenzie et al., 2023), and large-scale, multi-domain evidence shows that (i) machine and human difficulty rankings are only weakly correlated and (ii) models' explicit difficulty judgments are largely decoupled from their realized correctness, indicating a lack of "self-awareness" (Li et al., 2025). Related gaps also appear in agentic benchmarks, where tasks that are intuitive for humans can be challenging for agents and vice versa (Mialon et al., 2023; Xie et al., 2024).

Takeaways. Taken together, our results and related evidence imply that human difficulty is an imperfect indicator for GenAI performance: models can be highly accurate on average while exhibiting weak and context-sensitive alignment with the human difficulty gradient and with human error modes. When users nonetheless rely on human difficulty as informative towards AI ability, as HP predicts, their beliefs can be systematically misspecified, with consequences for when and how they delegate to or trust these systems—we study this in the next sections.

4 Documenting Human Projection: Beliefs in Performance

We design an experiment to elicit beliefs in performance over the domain of mathematical tasks presented in Section 3. We test whether beliefs are consistent with the predictions of our Human Projection framework (Section 2), i.e., whether people see *human* task difficulty as a relevant feature in predicting AI performance.

4.1 Experimental Design

Overview and randomization. Participants predict an agent's performance on math items sampled from our task domain. We randomize the identity of the agent between subjects: in *Human*, beliefs concern a randomly selected participant from our initial human benchmark; in *AI*, beliefs concern ChatGPT (GPT-3.5). To minimize demand effects, we never reference "difficulty" in the task instructions, and each treatment describes only the relevant agent (AI is not mentioned in *Human*, and human test-takers are not mentioned in *AI*). The experiment has two stages: (i) elicitation of prior beliefs across a set of items spanning the difficulty distribution, and (ii) a belief-updating task where respondents update beliefs about performance on one *hard* prediction item after observing performance on a separate signal item. Within each agent treatment, we additionally randomize the updating signal condition between subjects: Easy Fail, Hard Fail, Easy Success, or Hard Success. Overall, each respondent reports 11 prior beliefs (10 in Stage 1 and 1 in Stage 2) and 1 posterior belief (Stage 2).

Introducing the agent. We begin by providing brief background information about the agent to increase realism and reduce noise in elicited beliefs. In *AI*, we introduce large language models and illustrate ChatGPT responding to a small set of generic prompts unrelated to the math items. In *Human*, we describe the human benchmark sample and testing conditions from the initial data collection, and provide basic aggregate demographics of the test-taker pool.

Stage 1: prior beliefs. Each participant is shown 10 items on the same page and provides two incentivized responses for each item: (i) their own answer, and (ii) a prob-

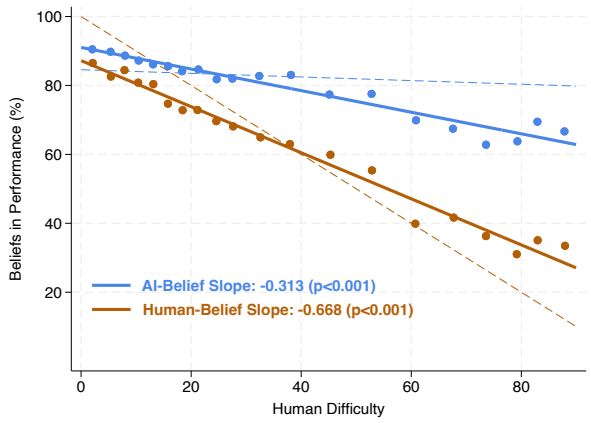
abilistic belief about whether the agent answered correctly. Beliefs are elicited with the same wording across treatments except for the agent label: "What do you think is the % chance that [a random participant / ChatGPT] answered correctly?" Item draws are stratified to expose each respondent to a broad range of difficulties: 2 items are sampled from the high-school (HS) pool (excluding the subset reserved for the updating stage), and 8 items are sampled from the combined 4th- and 8th-grade pool. Eliciting respondents' own answers before beliefs provides them with a private signal of task difficulty and generates additional performance data that we use to refine the human-difficulty index.

Stage 2: belief updating. The updating stage elicits one prior and one posterior belief for a single (hard) prediction item. First, respondents draw one prediction item from a set of 10 High-School items reserved for this stage, answer it, and report their prior belief that the agent answered it correctly. Second, respondents observe the agent's performance on a separate *signal* item (distinct from the prediction item) drawn from one of four pools corresponding to the randomized condition (easy vs. hard $\times$ success vs. failure). "Easy" and "Hard" signal pools are defined using the human-difficulty index (lowest and highest deciles), and success/failure is implemented by drawing from pools where the agent's realized answer is respectively correct or incorrect. Respondents see the signal item, the agent's answer, and whether that answer is correct. Third, the prediction item is shown again and respondents report their posterior belief: "Given what you saw, what do you think is the % chance that [the same participant / ChatGPT] answered this question correctly?" We define belief movement as posterior minus prior for the prediction item, and test Prediction 2 by comparing movement across signal-difficulty conditions within successes and within failures.

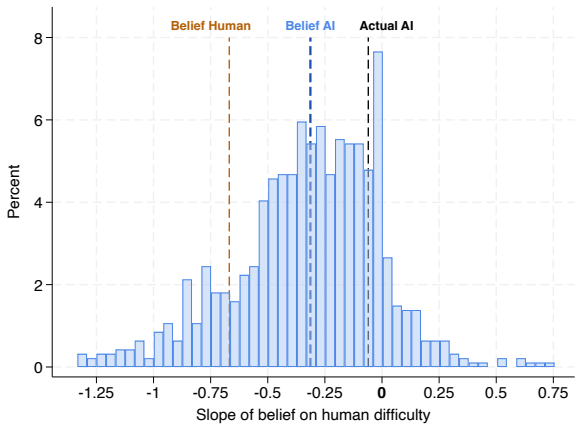
Incentives and logistics. Correct answers to math items are rewarded at \$0.05 each. We keep those stakes low to be comparable to the initial benchmark and to limit incentives to use outside help. Belief elicitation is incentivized using a binarized scoring rule (Hossain and Okui, 2013), described in intuitive terms; respondents can earn up to \$0.10 per prior belief report (10 reports) and up to \$0.30 for the updating belief report. The study was

Figure 2: Prior Beliefs in Performance and Implied Difficulty Projection

(a) Binned Beliefs in Performance



(b) Histogram of Belief Slopes



Notes: Panel (a) presents binned scatter plots of prior beliefs about human and ChatGPT performance. The sample excludes subjects who failed the comprehension question related to performance prediction. $n = 2442$ for humans, and $n = 10021$ for ChatGPT. Dashed lines represent actual human and AI performance. We report coefficients from the basic OLS regression with standard errors clustered at the subject level. Panel (b) plots the distribution of subject-specific slopes from regressing ChatGPT-belief on human difficulty (11 elicitation items per subject); vertical lines denote the true ChatGPT performance slope, the pooled ChatGPT-belief slope, and the pooled human-belief slope.

implemented in Qualtrics and run on Prolific in December 2023, following an initial human benchmark collection in October 2023. We recruited 244 subjects in *Human* and 971 in *AI*; after excluding subjects who fail attention checks and stage-specific comprehension questions, the analysis samples are 222 (*Human*) and 911 (*AI*) for the priors stage, and 232 (*Human*) and 824 (*AI*) for the updating stage.[3] As stated in the IRB submission and instructions, the experiment does not rely on deception: beliefs are incentivized, and performance signals are generated using real human or model answers.

4.2 Results

Interpretive baseline. Predictions 1–2 concern how beliefs vary with *human* task difficulty. For beliefs about humans, this index is relevant by construction. For beliefs about AI, however, there is no strong *a priori* reason for the human-defined difficulty ranking to

[3]We include separate comprehension questions for each stage which can yield different sample sizes across stages and conditions. In *AI*, failure conditions for updating exhibited higher check-failure rates, and we recruited additional participants to maintain cell sizes.

Table 1: Correlations of Beliefs in Performance with Item Difficulty

	Beliefs in Human Perf.			**Beliefs in AI Perf.**		
	(1)	(2)	(3)	(4)	(5)	(6)
Task difficulty	-0.668***	-0.665***	-0.664***	-0.313***	-0.314***	-0.317***
	(0.020)	(0.021)	(0.022)	(0.010)	(0.010)	(0.011)
Controls		✓	✓		✓	✓
Subject FE			✓			✓
R-squared	0.391	0.389	0.602	0.116	0.123	0.440
Observations	2541	2497	2497	10340	10153	10153

Notes: The table reports OLS coefficients for task difficulty. Controls include demographics of age, gender, income, education, and prior AI familiarity (for *AI*). Standard errors clustered at the subject level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

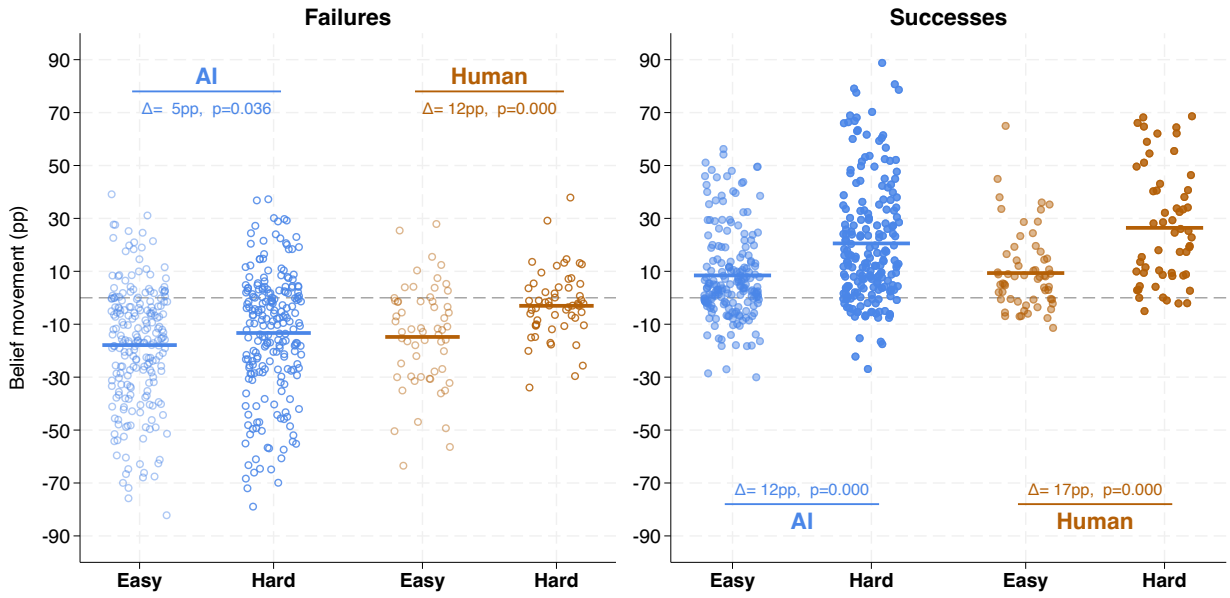
be equally informative. As evidence in Section 3 shows, models can differ from humans in which items they find easy or hard, and this misalignment can vary across model families and over time. We thus treat the dependence of AI beliefs on human difficulty as an empirical object, directly testing Human Projection.

Result 1: difficulty gradients in prior beliefs. Consistent with Prediction 1, prior beliefs decline sharply with human difficulty for both humans and AI (Figure 2; Table 1). The AI gradient is about half as steep as the human gradient: the pooled OLS slopes in Figure 2 imply a reduced-form $\hat{\lambda} \equiv \hat{\beta}^{AI}/\hat{\beta}^{H} \approx (-0.313)/(-0.668) = 0.47$, and this ratio is similar when adding covariates and estimating both slopes in a pooled interaction specification.[4] Beliefs are also higher on average for AI (mean 80%) than for humans (mean 64%). Overall, subjects treat the human difficulty index as highly informative for AI success probabilities—consistent with difficulty projection (difficulty alone explains about 12% of the variation in AI priors; Table 1).

In our GPT-3.5 case study, this induces systematic misspecification: relative to realized performance, subjects overestimate AI success on easier items and underestimate it on harder items (solid vs. dashed lines in Figure 2, panel A), and elicited beliefs have very limited predictive power for correctness across items (bivariate $R^2 \simeq 0.002$). Con-

[4]Implied projection is not systematically related to self-reported AI familiarity (joint test of familiarity dummies: $p = 0.98$), but appears lower among older respondents.

Figure 3: Belief Updating after Exogenous Signals of Performance



Notes: The figure shows jittered dot plots of belief movement (posterior minus prior), winsorized at the 2.5th and 97.5th percentiles. Sample sizes from left to right are $n = 221$, $n = 216$, $n = 59$, $n = 56$, $n = 192$, $n = 195$, $n = 60$, $n = 57$. Colored horizontal lines denote condition means; reported Δ are mean differences (hard vs easy) within failures and within successes, with p-values from two-sided tests of equality.

sistent with this aggregate pattern, the distribution of subject-specific AI belief slopes is centered above the true AI slope, so a majority of subjects impose an overly steep difficulty gradient (Figure 2, panel B).

Result 2: difficulty affects belief updating. Consistent with Prediction 2, the same performance outcome is interpreted differently depending on the signal task's difficulty. After a *success*, posterior beliefs increase more when the success occurs on a harder task; after a *failure*, posterior beliefs decrease less when the failure occurs on a harder task (Figure 3). Table 2 corroborates this pattern using a continuous difficulty measure and remains qualitatively unchanged across specifications with fixed effects and controls. Across both agents, the difficulty effect is larger after successes than after failures, a pattern that holds for both humans and AI (Figure 3; Table 2).[5]

[5]One mechanical contributor is that the realized separation in difficulty between the "easy" and "hard" signal pools is smaller for failures than for successes: using the refined difficulty index, the hard–easy gap is $77.0 - 9.1 = 67.9$ for failures versus $87.3 - 7.1 = 80.2$ for successes. This discrepancy arises because pools were constructed using preliminary performance data, while the difficulty measure is refined using

Table 2: Belief Updating - Effect of Signal Difficulty on Beliefs

	Belief change after a *success*				**Belief change after a *failure***			
	Human		**AI**		**Human**		**AI**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Task difficulty	0.222***	0.218***	0.157***	0.162***	0.158***	0.164***	0.067**	0.076**
	(0.046)	(0.050)	(0.025)	(0.026)	(0.046)	(0.047)	(0.030)	(0.031)
Controls		✓		✓		✓		✓
Prediction task FE	✓	✓	✓	✓	✓	✓	✓	✓
R-squared	0.221	0.240	0.110	0.123	0.148	0.162	0.060	0.074
Observations	117	113	387	371	115	110	437	425

Notes: The table reports OLS estimates of the effect of the signal task's human difficulty on belief movement (posterior minus prior), estimated separately for successes and failures and separately by agent. All specifications include prediction-task fixed effects and (in the indicated columns) respondent controls; robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Additional results. We next separate two reasons why AI priors may differ from human priors: (i) different *levels* due to different perceived average ability, and (ii) different *gradients* due to weaker difficulty projection.[6] As a simple diagnostic, we compare AI beliefs to beliefs about humans among respondents with high average beliefs about human performance (top decile). This restricted sample yields a human belief curve close in level *and* slope to the AI curve (Figure 11, Appendix C). Interpreted descriptively, the human–AI gap in priors appears to partly reflect higher baseline beliefs about AI ability, not only differences in the difficulty gradient.

Then, Prediction 2 also implies that respondents with stronger difficulty projection should exhibit a larger effect (within failures and within successes) of signal difficulty on updating. Consistent with this implication, when we split respondents by their estimated degree of difficulty projection, the hard–easy updating gap is systematically larger among high-projection subjects (Appendix C, Figure 12). Here as well the effect is more pronounced after successes: the hard–easy gap rises from about 5pp among low-projection subjects to about 21pp among high-projection subjects, with smaller but directionally similar differences after failures.

Takeaways. Taken together, the results support Human Projection in two distinct senses.

the larger beliefs-experiment sample (see Section 4.1).

[6]Because we elicit probabilistic beliefs rather than the subjective difficulty-performance mappings, differences in average beliefs can reflect differences in perceived ability even assuming the same mappings.

First, AI priors inherit a human-difficulty gradient (Prediction 1), consistent with *difficulty projection* alone. Second, subjects generalize from one observed outcome to other tasks in a way that depends on signal difficulty (Prediction 2), consistent with combining difficulty projection with *Ability–Model projection*—i.e., interpreting outcomes as informative about a latent, cross-task ability. In our GPT-3.5 case study, this anthropomorphic structure yields systematically misspecified AI beliefs.

We now turn from diagnosing Human Projection in beliefs to its behavioral consequences for AI adoption, in two distinct contexts: (i) continued engagement in the field with an advice chatbot (Section 5), and (ii) delegation decisions in a controlled environment with endogenous AI signals (Section 6).

5 Consequences of HP for Engagement with AI Chatbots

In this section we examine a concrete implication of Human Projection for real-world *engagement* with AI advice chatbots. While our earlier analysis uses objective task difficulty as the projected feature, the same logic applies to other human-relevant cues in deployed systems; here the key feature is the *reasonableness* of a (wrong) answer. Conditional on being wrong (defined as unhelpful), some answers look like plausible human misunderstandings ("close" to what a reasonable person might have said), while others appear less reasonable; such differences often affect inferences about human competence. HP may lead users to apply the same logic to AI failures, potentially leading them to disengage after rare but un-human-like mistakes. We vary the human reasonableness of AI failures in a framed field experiment, and measure resulting beliefs, trust, and engagement.

5.1 Theoretical Framework in Context

We adapt the framework to the ParentData chatbot setting, a context in which "difficulty" and correctness are significantly more subjective than for mathematics. The chatbot only returns human-vetted, high-quality answers, but it can "fail" by misunderstanding the query and returning an answer that is useful for a different question, but

useless for the user's question. We summarize the setup and the key prediction here and relegate formal details to Appendix A.

Setup. A principal (a user) cares about whether an agent (the chatbot) provides a *useful* answer to some questions. We model performance as binary usefulness, mirroring our experimental focus on inference from failures. That is, for every question q_i there is a non-empty set of useful answers A_i. Upon receiving answer a_j, the principal receives strictly positive utility if a_j belongs to A_i, and 0 otherwise.

Similarity projection. Assume each pair of answers (a, a') has a human–similarity index S capturing semantic overlap, contextual proximity, shared meaning, etc. Define the *reasonableness* of answer a_j to question q_i as $r_{ij} = \max_{a \in A_i} S(a_j, a)$. In words, the reasonableness of a wrong answer is its (human) similarity to a useful answer. The chatbot matches a user query to pre-made questions using a score highly correlated with textual similarity measures, and displays the highest-score answer.[7] Because the system returns the answer paired with the nearest-neighbor query under its similarity metric, the displayed response is (by construction) among the highest-scoring pre-written matches to the users prompt. Absent projection, conditional on observing a failure (i.e., useless answer), human reasonableness need not affect beliefs *a priori*: no alternative, lower-score answer should be interpreted as "more reasonable" than the top-score answer. If instead users project human similarity onto AI, then holding fixed the question and the uselessness of the answer, a less humanly-reasonable failure should trigger more negative inferences, lowering posterior beliefs in performance.

To see the intuition, consider the following *real* conversation used in the experiment. The user asks: "Which is the best car seat brand?" One failure answers about where to install a car seat (front vs. back seat), while another discusses the best baby food brand. The former is typically judged more reasonable because it shares the relevant context and thus appears more humanly similar to a useful answer (which would compare brands and recommend one).

[7] More precisely, it matches the user's question to all pre-made questions and displays the answer corresponding to the highest-score match. Since there is a one-to-one mapping between pre-made questions and answers, we work directly with answer similarity.

In this context, Human Projection thus involves the projection of human similarity and of the modified Ability Model, where the principal perceives higher-ability agents as relatively more likely to give more reasonable answers. It yields the following prediction, analog of Prediction 2 for failures:

Prediction 3. *When the user observes an AI failure, if the answer is less reasonable from a human perspective: (i) Beliefs in AI performance decrease more; (ii) Trust in AI decreases more; (iii) Likelihood of continued engagement is lower.*

5.2 Engagement Experiment

We test Prediction 3 in a framed field experiment (Harrison and List, 2004) with a sample designed to mirror ParentData's user base. We recruit online participants who match the platform's core user demographics and show them a sequence of real parent-chatbot conversations in an ecological survey interface, varying between subjects the *reasonableness* of the chatbot's misunderstandings. We then elicit beliefs in performance, trust, and measure engagement both through a revealed-preference link choice and actual subsequent usage.

5.2.1 Design and procedures

Selecting matched failure pairs. To isolate the role of *human reasonableness* in reactions to AI failures, we constructed matched *pairs* of real ParentData conversations in which the chatbot misunderstands the user's question. We pre-registered a three-step procedure. First, we manually coded human-AI conversations for (i) the user's intent and (ii) whether the answer reflects a misunderstanding (hence, is unhelpful), and formed candidate pairs of same-intent conversations where both answers were misunderstandings. Second, in two separate rating studies with the target population (parents/currently expecting), we elicited (a) perceived *reasonableness* of the specific misunderstanding (defined as the perceived probability that a reasonable human would make the same misunderstanding); and (b) answer *usefulness*, using instructions designed to separate the

two constructs. Finally, we retained five pairs for which the two sides are (i) same-intent, (ii) equally unhelpful (same median usefulness and no significant difference in means at the 10% level), and (iii) sharply different in reasonableness (large median gaps). Appendix D provides the full procedure, instructions, and the selected conversations with their ratings.

Design. The experiment is between-subject and varies only the *reasonableness* of AI failures. After instructions, subjects view five real human-chatbot conversations. The first three are "successes" (median usefulness 4–5) and are identical across treatments; the last two are "failures" (median usefulness 1–2, held fixed within each pair). Each failure is drawn from the five matched pairs described above, and treatment assignment determines which side is shown: subjects in *Reasonable* see the more reasonable misunderstanding, while those in *Unreasonable* see the less reasonable one.

Each conversation page initially displays the user query with a *Generate* button. Clicking it reveals the AI answer via a typewriter animation mimicking the actual chatbot. After each conversation, we elicit beliefs about AI performance and trust in the chatbot.[8] Beliefs are incentivized by paying for accuracy against an external benchmark: we draw 100 conversations at random and elicit usefulness ratings from the same target population.

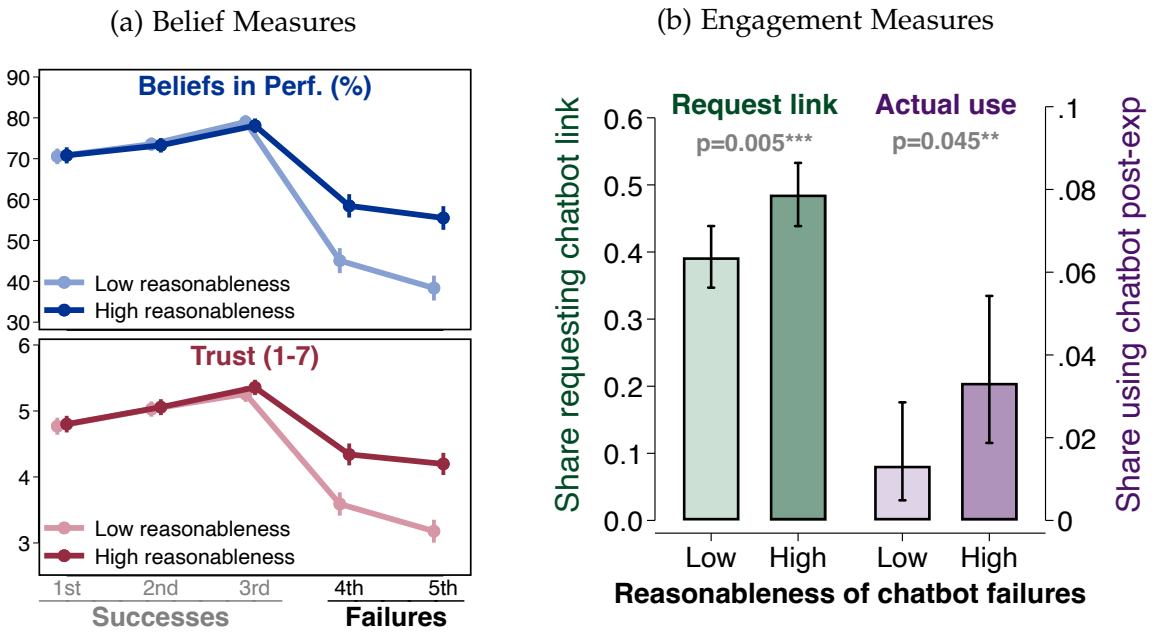
Immediately after the last conversation, subjects make the main willingness-to-engage choice: receive a link either to the chatbot or to a "large list (500+) of good-quality parenting articles." The selected link opens in a new tab upon survey completion; both options are free and non-binding. As an actual engagement measure, we track subsequent chatbot usage by matching the experiment IP addresses to those appearing in ParentData chatbot logs.[9] We conclude with demographics and prior familiarity with AI and ParentData.

Logistics. All participants (rating studies and experiment) were recruited on Prolific

[8]For beliefs: "What is the % chance the chatbot answers a random parenting question correctly?"; 0-100 scale. For trust: "How much do you trust the chatbot?"; 1–7 scale.

[9]For confidentiality, we do not observe the content of post-experiment questions. The link to the chatbot is available here, and to parenting articles here. The chatbot was trained on a corpus that includes the full list of parenting articles.

Figure 4: Effect of Reasonableness of Chatbot Failures



Notes: The figure plots average beliefs (0-100% scale) and trust (1-7 scale) in the chatbot. The first three points are posterior beliefs after each successful conversation (median usefulness of 4 or 5), while the last two are posteriors after failures (median usefulness of 1 or 2). Sample sizes are $n = 451$ for *Reasonable* and $n = 454$ for *Unreasonable*. Confidence intervals at the 95% level are included.

in July–August 2024. We targeted the "potential users" sample in two steps: a Prolific pre-screen (85% female; U.S.; age 18–45; in a relationship) followed by in-survey screening for current or expecting parents (excluding those with only adult children and those not parents/not trying). The pre-registration specifies the pair-selection procedure and elicitation designs, the engagement experiment, planned sample sizes, and primary hypotheses (engagement, beliefs, trust). The final sample is $n = 905$ (451 *Reasonable*, 454 *Unreasonable*). The survey lasted slightly over 5 minutes on average; mean earnings were \$0.87 (base \$0.70 plus up to \$0.50 in performance bonuses, $5 \times$ \$0.10, for belief accuracy).

5.2.2 Results

We present the main results here and relegate additional details to Appendix D.3. We first examine beliefs about performance and trust. Figure 4, Panel A, aligns with Prediction 3: beliefs and trust decline after failures in both treatments, but the drop is significantly larger under *Unreasonable* (about 80% larger for beliefs and 75% larger for

Table 3: Effect of Reasonableness – OLS Estimates

	Belief measures				Engagement measures			
	Belief in perf.		Trust		Request link		Actual use	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Unreasonable	-15.271***	-15.658***	-0.888***	-0.906***	-0.094***	-0.095***	-0.020**	-0.022**
	(1.864)	(1.833)	(0.108)	(0.106)	(0.033)	(0.033)	(0.010)	(0.010)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Pair FE		✓		✓		✓		✓
R-squared	0.096	0.138	0.097	0.130	0.023	0.043	0.016	0.025
Observations	905	905	905	905	905	905	905	905

Notes: This table reports OLS estimates of the effect of assignment to the *Unreasonable* (low-reasonableness) failure condition. For beliefs and trust, the dependent variables are the average of the post-failure outcomes after the 4th and 5th conversations. Engagement outcomes are indicators for requesting the chatbot link and subsequent chatbot usage within three weeks. Controls include age, gender, education, race indicator, and prior familiarity with AI and ParentData. Columns marked Pair FE include fixed effects for the matched failure pairs shown in the 4th and 5th conversations. Robust standard errors in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

trust, relative to *Reasonable*). OLS estimates in Table 3 confirm these patterns, and treatment effects are robust to adding demographic and familiarity controls and fixed effects for the matched failure pairs shown in the 4th and 5th conversations.

Engagement moves in the same direction. Figure 4, Panel B, shows that low reasonableness decreases both (i) the probability of requesting the chatbot link (vs. parenting articles) and (ii) subsequent usage (asking at least one question within three weeks; unconditional rate). We further find that even conditional on requesting the chatbot link, follow-through is lower under *Unreasonable* than under *Reasonable* (2.3% vs. 6.5%; $p = 0.048$). Excluding subjects who report prior familiarity with the website yields qualitatively similar patterns, with larger level gaps in engagement (link choice: 49% vs. 39%, $p = 0.005$; usage: 3.4% vs. 0.9%, $p = 0.010$; *Reasonable* vs. *Unreasonable*).

Taken together, the results suggest users engage in HP when interacting with the chatbot: the human reasonableness of its errors strongly shapes inferences and subsequent engagement. While reasonableness is arguably informative when evaluating human mistakes, it may not be so for AI: in a separate rating exercise with a similar sample of potential users, the chatbot is typically rated about 3.5 out of 5 in usefulness, and only about 14% of conversations are judged not useful (median usefulness ≤ 2)—the events our framework treats as AI failures.

6 AI Adoption in Equilibrium under Human Projection

We study AI adoption in in Berk-Nash equilibrium when principals evaluate AI performance through HP, showing that HP can generate *both* over- and under-adoption compared to an optimal benchmark.[10] Then, we test this model in a lab experiment that experimentally varies HP across subjects.

6.1 Framework for Equilibrium AI Adoption

Setup. A principal manages n types of tasks indexed by $i \in \{1, \ldots, n\}$ with (human-relevant) difficulties $\delta_1 < \delta_2 < \cdots < \delta_n$. Each task can be performed by either a human agent or an AI. The human agent has known ability $\theta^H \in \Theta \subseteq \mathbb{R}_+$, which determines success rates via a performance mapping $p^H(\theta^H, \delta_i) \in [0,1]$ for each task i. The AI has unknown success rates denoted by the vector $\mathbf{p}^A = (p_1^A, \ldots, p_n^A) \in [0,1]^n$. The principal's profit takes a separably additive form $\Pi(a_1, \ldots, a_n) = \sum_{i=1}^n \left[R_i \cdot \tilde{p}_i^{a_i} - C_i^{a_i}\right]$, where $a_i \in \{\text{Human}, \text{AI}\}$ denotes the agent assigned to task i, $\tilde{p}_i^{a_i} \in [0,1]$ is the success probability of agent a_i on task i, $R_i > 0$ is the revenue from success on task i, and $C_i^{a_i} \geq 0$ is the fixed cost of deploying agent a_i for task i. Since $R_i > 0$, maximizing profit is equivalent to maximizing $\sum_{i=1}^n R_i \cdot p_i^{a_i}$ where $p_i^{a_i} \equiv \tilde{p}_i^{a_i} - C_i^{a_i}/R_i$ denotes the *cost-weighted success rate*. Without loss of generality, we thus interpret all success rates as these cost-weighted values and denote them simply as p_i^H and p_i^A.[11]

Beliefs, projection, and actions. We assume the principal engages in Human Projection, as defined in Section 2. They entertain a mental model of performance $p(\cdot, \cdot)$ satisfying Assumption 1: for tractability, we focus on full projection and additionally impose $p^A(\cdot, \cdot) = p^H(\cdot, \cdot) = p(\cdot, \cdot)$—the principal uses the same model for humans and AI, even though true AI performance vector $\mathbf{p}^A$ may depart from it. The principal holds a prior belief G^A over an AI "ability" parameter $\theta^A \in \Theta$, which may differ arbitrar-

[10] In Appendix A we consider an extension of the model where the AI technology frontier expands over time. We show that Human Projection delays initial adoption relative to the no-projection benchmark, but once the hyperplane $\mathcal{H}$ is crossed, AI is immediately over-adopted for all tasks including those where it performs worse than humans.

[11] We maintain that costs are not so large as to push $p_i^{a_i} = \tilde{p}_i^{a_i} - C_i^{a_i}/R_i$ outside $[0,1]$.

ily from the known human ability θ^H. Under these assumptions, adoption decisions generate performance data that discipline beliefs about θ^A.

Equilibrium. We model adoption and learning using Berk–Nash equilibrium (Esponda and Pouzo, 2016). The principal chooses an adoption set $x \subseteq \{1, \ldots, n\}$, where $i \in x$ means task i is assigned to AI. Given x, outcomes are an n-vector of independent successes/failures, with true success rates equal to p_i^A on adopted tasks and p_i^H otherwise. Under Human Projection, the principal evaluates AI performance through a single ability parameter θ and forecasts success on adopted task i as $p(\theta, \delta_i)$, while human performance is p_i^H. For a given action x and belief θ, let $Q(\cdot \mid x)$ denote the objective outcome distribution and $Q_\theta(\cdot \mid x)$ the subjective one. The "best-fitting" belief under x minimizes the Kullback–Leibler divergence

$$K(x, \theta) = \mathbb{E}_{Q(\cdot|x)}\left[\ln \frac{Q(y \mid x)}{Q_\theta(y \mid x)}\right] = \sum_{i \in x}\left[p_i^A \ln \frac{p_i^A}{p(\theta, \delta_i)} + (1 - p_i^A) \ln \frac{1 - p_i^A}{1 - p(\theta, \delta_i)}\right].$$

A (pure) Berk–Nash equilibrium is an action x^* and belief $\theta(x^*)$ such that (i) x^* maximizes the principal's subjective payoff given $\theta(x^*)$, and (ii) $\theta(x^*)$ minimizes $K(x^*, \theta)$.

Optimal adoption benchmark. If the principal does not engage in HP, the optimal adoption rule is task-by-task: assign task i to AI if and only if $p_i^A > p_i^H$. This can yield no adoption, partial adoption, or full adoption depending on how $\mathbf{p}^A$ compares to $\mathbf{p}^H$. In particular, partial adoption is optimal whenever AI has a comparative advantage in some but not all tasks, as our evidence suggests in Section 3.

Theorem 1 (All-or-Nothing Adoption)**.** *Under full Human Projection:*

(i) ***Partial adoption*** *is never a BkNE, even when it would be the optimal action.*

(ii) ***Full adoption*** *is the unique BkNE when* $\mathbf{p}^A$ *lies strictly above the hyperplane* $\mathcal{H}$ *given by*

$$\mathcal{H} = \left\{\mathbf{p} \in [0,1]^n : \sum_{i=1}^n w_i(\theta^H) p_i = \sum_{i=1}^n w_i(\theta^H) p_i^H\right\}, \qquad w_i(\theta^H) = \frac{\partial p(\theta^H, \delta_i)/\partial \theta}{p_i^H(1 - p_i^H)}.$$

(iii) ***No adoption*** *is the unique BkNE under exploration when* $\mathbf{p}^A$ *lies strictly below* $\mathcal{H}$*. Moreover, it can be sustained without exploration under sufficiently pessimistic priors.*[12]

[12]Specifically, when the prior G^A satisfies $\mathbb{E}_{G^A}[p(\theta^A, \delta_i)] \leq p_i^H$ for all i. A sufficient condition is

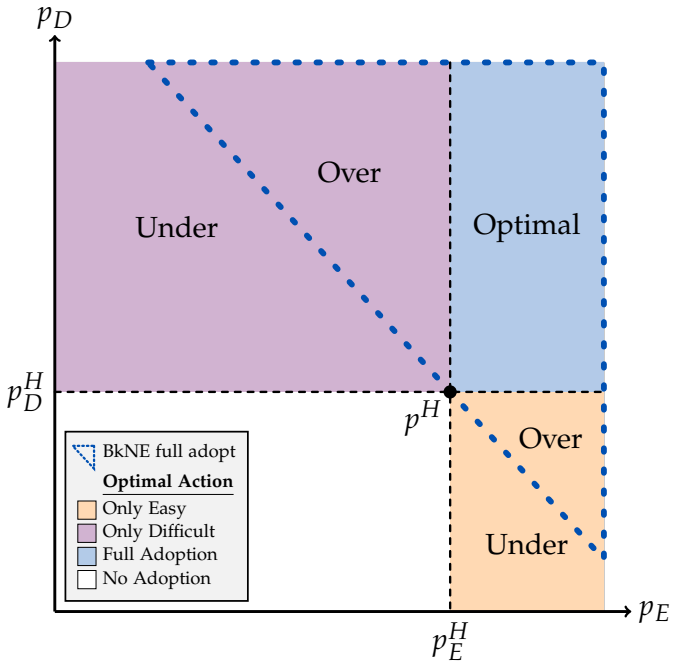
Intuition. The all-or-nothing pattern arises because the principal tries to infer a single ability parameter θ^A from the data generated by AI use and then applies the same mapping $p(\cdot, \delta_i)$ across tasks. Under partial adoption, the KL-minimizing belief θ^* typically ends up either above or below the human benchmark θ^H. If $\theta^* > \theta^H$, the principal concludes that AI should outperform the human *on every task* and switches to full adoption; if $\theta^* < \theta^H$, the principal concludes AI is worse *on every task* and switches to no adoption. Thus, partial adoption cannot be sustained: the principal is pushed toward the coarse conclusion that AI is either "better" or "worse" than humans in the domain, even when the true technology $\mathbf{p}^A$ features comparative advantage and calls for partial adoption.

The hyperplane $\mathcal{H}$ in Theorem 1 defines the boundary separating these regions in the AI technology space $[0,1]^n$. Figure 5 illustrates this boundary for the two-task case, where the hyperplane is a downward-sloping line passing through $\mathbf{p}^H$. It shows that Human Projection yields both regions of over-adoption (full adoption despite AI being inferior on some task) and under-adoption (no adoption despite AI being superior on some task) relative to the optimal benchmark. Note that the general n-task result becomes more ecologically plausible as the number of tasks increases: tracking task-specific AI success rates becomes cognitively burdensome, making it natural to rely on a heuristic of inferring a single ability parameter θ^A, informing all tasks.

Theorem 1 serves as a benchmark: it assumes *full* Ability–Model projection (a unidimensional θ^A) and that the principal forecasts AI performance using the human-relevant difficulty index (δ_i). In realistic AI adoption settings, however, either channel may be attenuated, and this is precisely where partial adoption arises in BkNE. For instance, if the principal perceives multiple AI-relevant subdomains (type-specific abilities), AI can rationally be viewed as better on some tasks but worse on others. Likewise, if tasks have AI-specific effective difficulties, comparisons to the human benchmark need not shift uniformly across tasks. This motivates an empirical question: what features of an AI system make Human Projection more (or less) likely, thereby shifting adoption patterns away from the benchmark? In what follows, we derive a testable prediction focusing on a salient and widely used AI design feature: "anthropomorphic" cues.

$\mathbb{E}_{G^A}[\theta^A] \leq \theta^H$ if $p(\theta, \delta)$ is concave in θ.

Figure 5: Adoption Equilibrium Regions Under Human Projection



Notes: Illustration for the two-task case ($n = 2$). The dashed blue line represents the hyperplane $\mathcal{H}$ from Theorem 1. Full adoption is sustained as BkNE above this line; no adoption is sustained below it. Colored regions indicate optimal task-by-task adoption. Over-adoption occurs when full adoption is sustained despite AI being inferior to humans on at least one task. Under-adoption occurs when no adoption is sustained despite AI being superior on some task.

Experimental prediction. Many consumer-facing AI systems are presented with human-like cues (e.g., dialogue interfaces, names, voices, etc). A natural hypothesis is that these cues increase reliance on HP when reasoning about AI performance—consistent with findings from experimental psychology (Chugunova and Sele, 2022).

Assumption 3. *Anthropomorphic presentation of AI increases the likelihood that principals engage in Human Projection.*

In our setting, we expect a non-anthropomorphic ("black box") presentation to weaken HP through either projection channel (Ability–Model or difficulty). Combined with Theorem 1, Assumption 3 yields the main experimental prediction:

Prediction 4. *A non-anthropomorphic presentation of AI decreases the share of all-or-nothing adoption decisions by reducing Human Projection.*

6.2 Adoption Experiment

Our experiment implements the endogenous, selected feedback from delegation assumed by the model, and tests whether reducing Human Projection shifts adoption away from all-or-nothing choices and toward partial adoption when AI holds a comparative advantage in only some task types.

6.2.1 Design

We test Prediction 4 using the benchmark of tasks described in Section 3. Participants are introduced to the tasks and AI, then go through a delegation phase where they acquire costly signals about human and AI performance, and make a one-shot high-stakes adoption decision at the end. Our experimental variation shifts the prevalence of Human Projection at the treatment level during instructions, followed by a delegation phase and a final adoption choice testing equilibrium predictions.

Tasks. We introduce the tasks as belonging to two pools, "blue" and "green", and familiarize participants with them during instructions: we show 6 examples (held fixed across treatments) from each pool, with human success rates and AI performance (a success) for one of these problems. Problems in the blue pool are human-easy (mostly 4^{th}- and 8^{th}-grade; 78% human success), while problems in the green pool are human-hard (mostly high-school; 23% human success).[13] AI success is held constant at 66% across both pools, so humans outperform the AI on blue tasks and the AI outperforms humans on green tasks. This design specifically tests for the treatment effect on the share of Full Adoption and makes Only Hard the optimal adoption choice given our incentive scheme.

Treatment variation. Our experimental variation shifts the prevalence of Human Projection at the treatment level by varying the human-likeness of AI's framing. In the

[13]The blue pool is relatively human-easy while the green one is relatively human-hard. In the baseline version of the experiment, we never explicitly refer to the difficulty of each pool, and find evidence of significant misperceptions among subjects. We run a replication with an identical design where we reduce misperceptions by explicitly calling blue tasks as "mostly of 8th-grade level" and green tasks as "mostly of High School level." See Section 6.2.2 for details.

Anthropomorphic condition, we use a standard consumer-facing AI framing: we present the AI as "Morgan," a neutral AI based on "currently available LLM-technology," describe its behavior in the active voice, and depict it with a human-like logo. This framing resembles interfaces used by leading LLM companies (e.g., Anthropic's *Claude* or OpenAI's *ChatGPT*). In the *Black Box* condition, we present the AI as a "black box": we use the passive voice when describing its behavior and omit any reference to a human being.

We invite participants to use information from the delegation phase to inform their final adoption decision. Using language inspired by Esponda et al. (2024), we induce participants to be more agnostic about the black box's performance: "100 green problems were put into the black box. A certain number (which you won't be told) came out as successes, and the rest as failures" (*Black Box*), compared to Morgan's: "Morgan attempted 100 green problems. It successfully solved a certain number (which you won't be told), and failed the remaining ones" (*Anthropomorphic*).

Delegation phase and final adoption. The delegation phase consists of 60 problems (30 from each pool) shown in random order. On each problem, the participant chooses whether to delegate to a random human or to the AI, and then observes whether their chosen agent succeeds or fails on the next screen.[14] Delegation is incentivized via a small bonus for each success observed. After delegating all 60 problems, participants make a one-shot adoption decision with larger incentives: they choose whether to adopt a human or the AI to solve 10 tasks from each pool. This yields four adoption patterns: No Adoption, Only Easy, Only Hard, and Full Adoption.

We elicit probabilistic beliefs in human and AI performance on each pool at three points: before the first delegation (initial beliefs), after 30 delegations (interim), and after the final adoption decision (final), using the following language: "What do you think is the success rate of [Morgan / the black box] on the [blue / green] problems?".

Logistics. We programmed the experiment in Qualtrics and recruited participants on Prolific in August 2024. The pre-registration specifies the design, sample sizes, and main hypotheses. We collected two samples. The first includes 150 subjects in *Anthropomorphic*

[14]Performance corresponds to ChatGPT 3.5's benchmark in Section 3. For humans, we draw outcomes from a Bernoulli distribution with the empirical success rate for that pool.

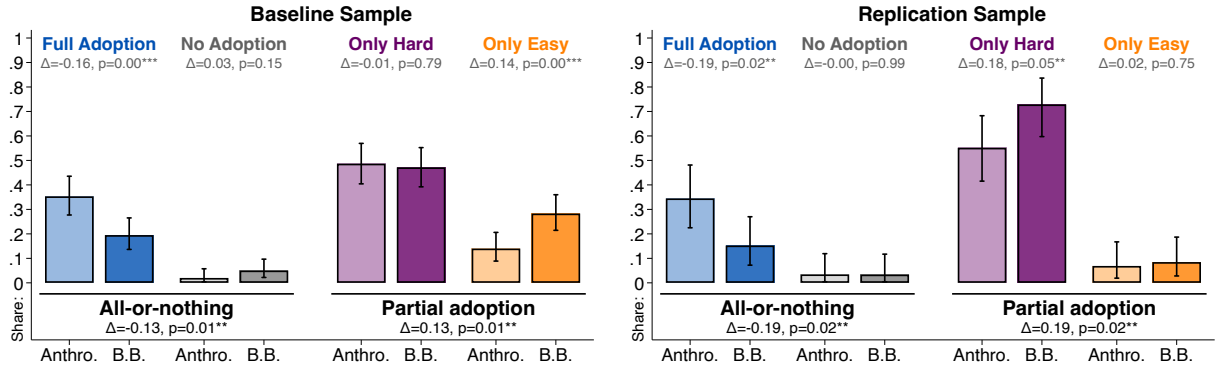
and 159 in *Black Box*. While this sample delivers the main pre-registered reduction in all-or-nothing adoption, it also exhibits non-optimal choices that were not pre-registered and appear consistent with misperceptions of the relative *human* difficulty of the two pools. We therefore collected a replication sample (58 and 59 subjects, respectively) with identical procedures except that human difficulty is made more salient: blue problems are (truthfully) labeled "grade school," and green problems "high school." We report results from both samples in the main text (Section 6.2.2) and present evidence of misperceptions in Appendix E. On average, subjects earned \$4.2 for a median completion time of around 15 minutes. The base fee was \$1.7 with a potential bonus of \$4.4 (\$0.03 per training-phase success observed, \$0.05 per accurate belief report, and up to \$2 for the adoption decision).

6.2.2 Results

Main effect: all-or-nothing adoption. Figure 6 reports the shares of participants selecting each final adoption choice, among the baseline (left) and replication samples (right). Consistent with Prediction 4 and our pre-registration, we find that all-or-nothing adoption is significantly lower under *Black Box* than under *Anthropomorphic* in both samples (baseline: $\Delta = -0.13$, $p = 0.01$; replication: $\Delta = -0.19$, $p = 0.02$), driven by a reduction in Full Adoption (baseline: $\Delta = -0.16$, $p < 0.01$; replication: $\Delta = -0.19$, $p = 0.02$). This shift is offset by an increase in partial adoption. In the baseline sample, the increase is concentrated in Only Easy rather than Only Hard (the ex post optimal choice given realized AI performance across pools); this unanticipated pattern motivated the replication and the analyses below.

(Mis)perceptions of difficulty and optimal adoption. In the baseline sample, much of the *Black Box* reallocation goes to Only Easy. This pattern is closely linked to a simple proxy for misperceived *human* difficulty: 56% report initial beliefs about *human* performance that are weakly lower on the easy pool than on the hard pool. We replicate the experiment with an identical design while making the human difficulty ordering more salient; the same reduction in Full Adoption reappears, but it shifts towards Only Hard

Figure 6: Treatment Effect on Adoption Decisions



Notes: The figure reports shares (0–1) of subjects choosing each final adoption choice, across treatments *Anthropomorphic* and *Black Box*. Sample sizes are respectively $n = 150$ and $n = 159$ (baseline) and $n = 58$ and $n = 59$ (replication). All-or-nothing adoption corresponds to either Full or No adoption, and Partial adoption to either Only Hard or Only Easy. p-values from two-sided proportions tests across treatments are reported.

rather than Only Easy (Figure 6, right panel). Consistent with this interpretation, the proxy falls sharply from 56.3% (baseline) to 17.1% (replication; $\Delta = -39.2$ pp, $p < 0.001$) and is unaffected by treatment in either wave (Table 8; Appendix E.1).

Pooled beliefs-conditional regressions (Table 9) confirm that misperceptions mainly shift the *direction* of partial adoption while leaving the all-or-nothing reduction unchanged: among subjects with correctly ordered initial beliefs, *Black Box* increases Only Hard without raising Only Easy, whereas the pattern reverses among those with misordered beliefs. Thus, the misperceptions proxy predicts *which* pool participants delegate to under partial adoption but not *whether* they partially adopt.

Mechanism: weakening of Human Projection. Table 4 reports belief-based proxies for the two components of Human Projection: Ability–Model projection (how strongly subjects summarize AI performance with a single "ability" index) and difficulty projection (how closely subjects map human difficulty differences onto AI). The *Black Box* framing reduces all three measures of Ability–Model projection, with the largest effects in initial beliefs (e.g., easy–hard belief comovement falls from 0.69 to 0.32, and the difficulty-gradient precision index from 0.95 to 0.85; both $p < 0.001$). Effects tend to weaken for interim and final beliefs but most remain statistically significant. By contrast, difficulty-projection effects are small and less stable: the rank-mapping proxy shows no detectable

Table 4: Adoption Experiment – Projection Mechanism

Channel	Measure	Beliefs	Treatment		Difference	p
			Anthropomorphic	*Black Box*		
Ability-Model Projection	*Single-index AI ability (belief comovement)*	Initial	0.69	0.32	**-0.37**	<0.001
		Interim	0.38	0.18	**-0.21**	0.020
		Final	0.41	0.17	**-0.24**	0.008
	Belief ordering consistent with human difficulty	Initial	0.803	0.661	**-0.142**	<0.001
		Interim	0.654	0.532	**-0.122**	0.010
		Final	0.562	0.505	**-0.058**	0.232
	Difficulty-gradient precision (IQR-based)	Initial	0.950	0.845	**-0.105**	<0.001
		Interim	0.885	0.855	**-0.030**	0.010
		Final	0.915	0.865	**-0.050**	0.014
Difficulty Projection	*Human-AI difficulty mapping (rank)*	Initial	0.54	0.49	**-0.05**	0.483
		Interim	0.14	0.23	**0.09**	0.314
		Final	0.08	-0.04	**-0.12**	0.202
	Human-AI difficulty mapping (level)	Initial	0.925	0.902	**-0.023**	0.054
		Interim	0.925	0.915	**-0.010**	0.115
		Final	0.940	0.900	**-0.040**	0.012
Observations			208	218		

Notes: The table reports belief-based proxies for Ability–Model projection and difficulty projection at the Initial, Interim, and Final phases (0–100 scales; higher values indicate more projection). Appendix E defines each proxy and reports the corresponding test procedures.

change in any phase, while the level-mapping proxy declines modestly (borderline in Initial, and significant in Final: 0.94 vs. 0.90, $p = 0.012$). Overall, the evidence indicates that *Black Box* primarily weakens respondents' tendency to extrapolate from signals of performance via a structured AI "ability", while having at most small effects on the human-to-AI difficulty mapping.

Delegation behavior. Table 11 in Appendix E reports that *Black Box* treatment reduces delegation to the AI during training, consistent with the lower initial beliefs induced by the *Black Box* framing. Conditional on delegating, realized AI failure rates are similar across framings, suggesting the treatment mainly shifts delegation propensity rather than the feedback received.

7 Conclusion

This paper documents *Human Projection*: a mental model through which people evaluate AI by extrapolating from observed performance using human-relevant features, such as task difficulty and the perceived reasonableness of errors. We develop a formal framework for Human Projection and validate its predictions experimentally, documenting clear anthropomorphic structure in beliefs about AI performance, both in priors and in belief updating from feedback. We then show this mental model has important behavioral consequences along two adoption margins. In the field, users penalize less-reasonable chatbot mistakes by reducing subsequent engagement with an otherwise useful tool. In a lab experiment testing an equilibrium adoption model, we find that de-anthropomorphizing AI framing reduces inefficient all-or-nothing reliance when AI has comparative advantage in some tasks but not others, highlighting a practical pitfall of anthropomorphism.

Our results have three practical implications for model design and human–AI interaction. First, beyond average accuracy, procedures such as RLHF or DPO should account for the features (e.g., task difficulty or error reasonableness) that influence users' generalization from observed performance and subsequent use. Second, our findings caution against treating benchmark scores as a sufficient guide to generalization across tasks or domains. Aggregate test accuracy can hide substantial item-level heterogeneity in performance and error type and induce unreliable inferences. This motivates the development of richer approaches to measuring model capability (Zhou et al., 2025). Third, anthropomorphic design presents a tradeoff: it can increase perceived legibility, yet also miscalibrate beliefs about capability relative to realized performance across tasks. As a result, the effects of human-like cues on engagement, trust, and reliance are likely to be context-dependent (Schimmelpfennig et al., 2025).

These dynamics could have serious consequences well beyond the settings we study. A large share of adolescents already engage with AI companions for emotionally sensitive conversations.[15] In contexts where engagement is monetized, providers may have

[15]Common Sense Media, *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions* (2025), na-

incentives to amplify anthropomorphic cues and deploy emotionally manipulative retention tactics (De Freitas et al., 2025). This can raise serious risks of dependency and boundary confusion precisely when the stakes are highest: understanding these inference and attachment channels is an urgent agenda for future research.

References

Andre, Peter, Philipp Schirmer, and Johannes Wohlfart, "Mental models of the stock market," 2024.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from weak signals and underinference from strong signals," *The Quarterly Journal of Economics*, 2025, *140* (1), 335–401.

Bachman, Lyle F, *Fundamental considerations in language testing*, Oxford university press, 1990.

Bai, Xuechunzi, Stefan Uddenberg, Brandon P Labbree, and Alexander Todorov, "Learning too much from too little: False face stereotypes emerge from a few exemplars and persist via insufficient sampling.," *Journal of personality and social psychology*, 2025, *128* (1), 61.

Bastianello, Francesca and Alex Imas, "Biases in belief updating within and across domains," Technical Report, National Bureau of Economic Research 2026.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans, "The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A"," *arXiv preprint arXiv:2309.12288*, 2023.

Bernstein, Michael H, Michael K Atalay, Elizabeth H Dibble, Aaron WP Maxwell, Adib R Karam, Saurabh Agarwal, Robert C Ward, Terrance T Healey, and Grayson L Baird, "Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography," *European radiology*, 2023, *33* (11), 8263–8269.

Bohren, J. Aislinn, Josh Hascher, Alex Imas, Michael Ungeheuer, and Martin Weber, "Mental Models of Information: The Role of Attention and Memory," *Becker Friedman Institute for Economics Working Paper*, 2025, (2024-19), 93. University of Chicago.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer, "Memory and probability," *The Quarterly Journal of Economics*, 2023, *138* (1), 265–311.

tionally representative survey. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

Caplin, Andrew, DJ Deming, Shangwen Li, DJ Martin, Philip Marx, Ben Weidmann, and KJ Ye, "The abcs of who benefits from working with ai: Ability, beliefs," Technical Report, and calibration. Technical report, National Bureau of Economic Research 2024.

Chevrier, Mathieu, Brice Corgnet, Eric Guerci, and Julie Rosaz, "Algorithm Credulity: Human and Algorithmic Advice in Prediction Experiments," *Available at SSRN 4828701*, 2024.

Chugunova, Marina and Daniela Sele, "We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines," *Journal of Behavioral and Experimental Economics*, 2022, *99*, 101897.

Cohn, Michelle, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth, "Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models," in "Extended Abstracts of the CHI Conference on Human Factors in Computing Systems" 2024, pp. 1–15.

Crolic, Cammy, Felipe Thomaz, Rhonda Hadi, and Andrew T Stephen, "Blame the bot: Anthropomorphism and anger in customer–chatbot interactions," *Journal of Marketing*, 2022, *86* (1), 132–148.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani, "Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality," *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 2023, (24-013).

Dietvorst, Berkeley J and Soaham Bharti, "People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error," *Psychological science*, 2020, *31* (10), 1302–1314.

_ , Joseph P Simmons, and Cade Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *Journal of experimental psychology: General*, 2015, *144* (1), 114.

Dreyfuss, Bnaya and Ruru Hoong, "Calibrated Coarsening: Designing Information for AI-Assisted Decisions," Technical Report, working paper 2025.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann, "Associative memory, beliefs and market interactions," *Journal of Financial Economics*, 2024, *157*, 103853.

Esponda, Ignacio and Demian Pouzo, "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 2016, *84* (3), 1093–1130.

_ , **Emanuel Vespa, and Sevgi Yuksel**, "Mental models and learning: The case of base-rate neglect," *American Economic Review*, 2024, *114* (3), 752–782.

Fan, Tony Q, Yucheng Liang, and Cameron Peng, "The inference-forecast gap in belief updating," Technical Report 2025.

Freitas, Julian De, Zeliha Oguz-Uguralp, and Ahmet Kaan-Uguralp, "Emotional manipulation by AI companions," *arXiv preprint arXiv:2508.19258*, 2025.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack, "Selective-Memory Equilibrium," *Journal of Political Economy*, 2024, *132* (12), 3978–4020.

Gagnon-Bartsch, Tristan and Antonio Rosato, "Quality is in the eye of the beholder: taste projection in markets with observational learning," *American Economic Review*, 2024, *114* (11), 3746–3787.

_ , **Matthew Rabin, and Joshua Schwartzstein**, "Channeled attention and stable errors," Technical Report 2023.

Gans, Joshua S, "A Model of Artificial Jagged Intelligence," Technical Report, National Bureau of Economic Research 2026.

Graeber, Thomas, "Inattentive inference," *Journal of the European Economic Association*, 2023, *21* (2), 560–592.

_ , **Christopher Roth, and Marco Sammon**, "Coarse Categories in a Complex World," Technical Report, University of Bonn and University of Cologne, Germany 2025.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Harrison, Glenn W and John A List, "Field experiments," *Journal of Economic literature*, 2004, *42* (4), 1009–1055.

He, Kevin, Ran Shorrer, and Mengjia Xia, "Human Misperception of Generative-AI Alignment: A Laboratory Experiment," *arXiv preprint arXiv:2502.14708*, 2025.

Hossain, Tanjim and Ryo Okui, "The binarized scoring rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

Igarashi, Mirai, Hiroshi Onishi, Osamu Sakai, and Shohei Hasegawa, "How chatbot anthropomorphism shapes customer purchase decisions: Evidence from a field experiment on a cosmetic E-commerce retailer," *Journal of Retailing and Consumer Services*, 2026, *91*, 104721.

Kendall, Chad and Ryan Oprea, "On the complexity of forming mental models," *Quantitative Economics*, 2024, *15* (1), 175–211.

Lanzani, Giacomo, "Dynamic concern for misspecification," *Econometrica*, 2025, *93* (4), 1333–1370.

Li, Ming, Han Chen, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou, "Can LLMs Estimate Student Struggles? Human-AI Difficulty Alignment with Proficiency Simulation for Item Difficulty Prediction," *arXiv preprint arXiv:2512.18880*, 2025.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin, "Projection bias in predicting future utility," *the Quarterly Journal of economics*, 2003, pp. 1209–1248.

Logg, Jennifer M, Julia A Minson, and Don A Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, 2019, *151*, 90–103.

Lord, Frederic M and Melvin R Novick, *Statistical theories of mental test scores*, IAP, 2008.

Madarász, Kristóf, David Danz, and Stephanie Wang, "Projective Thinking: Model, Evidence, and Applications," Technical Report 2023.

McKenzie, Ian R, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu et al., "Inverse scaling: When bigger isn't better," *arXiv preprint arXiv:2306.09479*, 2023.

Mialon, Grégoire, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom, "Gaia: a benchmark for general ai assistants," *arXiv preprint arXiv:2311.12983*, 2023.

Noy, Shakked and Whitney Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, 2023, *381* (6654), 187–192.

Pinker, Steven, "The language instinct (1994/2007)," 2007.

Reeves, Byron and Clifford Nass, "The media equation: How people treat computers, television, and new media like real people," *Cambridge, UK*, 1996, *10* (10), 19–36.

Samuelson, Larry and Jakub Steiner, "Robust Latent Representations," 2025.

Schanke, Scott, Gordon Burtch, and Gautam Ray, "Estimating the impact of humanizing customer service chatbots," *Information Systems Research*, 2021, *32* (3), 736–751.

Schimmelpfennig, Robin, Mark Díaz, Vinodkumar Prabhakaran, and Aida Davani, "Humanlike AI Design Increases Anthropomorphism but Yields Divergent Outcomes on Engagement and Trust Globally," *arXiv preprint arXiv:2512.17898*, 2025.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan, "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function," *arXiv preprint arXiv:2406.01382*, 2024.

Visser, Ewart J De, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman, "Almost human: Anthropomorphism increases trust resilience in cognitive agents.," *Journal of Experimental Psychology: Applied*, 2016, *22* (3), 331.

Wang, Boshi and Huan Sun, "Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure," *arXiv preprint arXiv:2504.01928*, 2025.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, 2022, *35*, 24824–24837.

Williams, Sean and James Huckle, "Easy problems that llms get wrong," *arXiv preprint arXiv:2405.19616*, 2024.

Woodford, Michael, "Modeling imprecision in perception, valuation, and choice," *Annual Review of Economics*, 2020, *12* (1), 579–601.

Xie, Tianbao, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei et al., "Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments," *arXiv preprint arXiv:2404.07972*, 2024.

Xu, Yuqian, Hongyan Dai, and Wanfeng Yan, "Identity disclosure and anthropomorphism in voice chatbot design: A field experiment," *Management Science*, 2024.

Yang, Zhe, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, and Zhifang Sui, "Can large language models always solve easy problems if they can solve harder ones?," *arXiv preprint arXiv:2406.12809*, 2024.

Yu, Feiyang, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar, "Heterogeneity and predictors of the effects of AI assistance on radiologists," *Nature Medicine*, 2024, *30* (3), 837–849.

Zhou, Hongli, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu et al., "Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory," *arXiv preprint arXiv:2505.15055*, 2025.

ONLINE APPENDIX

This appendix presents formal details and proofs of the framework and experimental predictions (Section A); details on the construction of the domain of mathematical tasks and procedures for measuring human and AI performance (Section B); separate sections for additional results and details about the beliefs experiment (Section C), engagement experiment (Section D), and adoption experiment (Section E). Further experimental instructions and materials are available at our Experimental Materials Appendix.

A Theoretical Appendix

This section provides formal details, statements and proofs of results and experimental predictions presented in the main text.

A.1 Details on Human Projection Framework

We consider a principal trying to assess the performance of an agent i within a *domain* $\mathcal{T}_j = \{t_1, \ldots, t_K\}$ composed of K tasks indexed by $k \in \{1, \ldots, K\}$ such that $(t_1, \ldots, t_K) \in \{0,1\}^K$ denote possible outcomes (success or failure) for each task of a random draw from some joint distribution over $\{0,1\}^K$. The vector $\boldsymbol{s} = (s_1, \ldots, s_K)$ collects the marginal distributions identified with the success rates $s_k \in [0,1]$, i.e., such that $\Pr(t_k = 1) = s_k$ for all $k \in \{1, \ldots, K\}$. These success rates are unknown to the principal.

Domains. A *domain* is any collection of tasks that the principal regards as sufficiently "similar" that performance can be summarized by a unidimensional notion of ability and an ordered difficulty index. This notion is deliberately broad: it may correspond to a subject area (e.g., math), a subfield (e.g., algebra), or any other grouping that supports stable inferences from limited observations.[16] In practice, the task partition rele-

[16] A natural extension allows multiple domains and cross-domain inference. Formally, one can posit domain-specific abilities $(\vartheta_1, \ldots, \vartheta_J)$ and a joint prior Π over Θ^J. Cross-domain updating is then governed by the dependence in Π: if Π factorizes across domains, within-domain signals do not update beliefs about other domains; if Π exhibits positive dependence, signals in one domain shift beliefs in others. Our analysis below fixes a single domain and requires no assumptions on Π.

vant for AI performance need not coincide with the human partition; the Ability–Model projection we introduce below corresponds to treating the human-relevant partition as informative for AI, thereby overstating the overlap between human and AI "domains."

Ability Model. Fix a human-relevant domain $\mathcal{T}_j$. The principal entertains an *Ability Model* for performance within $\mathcal{T}_j$: the agent has a latent ability $\theta_j \in \Theta \subseteq \mathbb{R}$, unknown to the principal. Each task $t \in \mathcal{T}_j$ has a level of *difficulty* denoted by $\delta_j(t) \in \Delta \subseteq \mathbb{R}$, and a problem with difficulty δ is denoted by t^δ. Conditional on $(\theta_j, \delta_j(t))$, the probability that agent $i \in \{H, A\}$ succeeds on task t is given by a mapping

$$p^i : \Theta \times \Delta \to [0,1], \qquad \Pr(t = 1 \mid \theta_j) = p^i(\theta_j, \delta_j(t)).$$

We assume that for each agent i, the mapping $p^i(\cdot,\cdot)$ satisfies the properties of Assumption 1 (in particular, monotonicity in θ and δ and the corresponding MLRP condition). To be consistent with our dataset of mathematical problems (spanning from 4^{th} grade to High School level questions), this MLRP assumption implicitly focuses on the lower and middle part of the human task difficulty distribution. It encompasses the very easiest tasks, which are solved with probability close to 1 even by agents of low ability levels, and progressively harder tasks, where success rates diverge due to ability. At the other extreme—tasks so difficult that success is arbitrarily close to 0 even for very high ability—the MLRP restrictions remain well-defined but become less empirically informative because all success rates are near zero.

Since all statements below are within a fixed domain, we drop the domain subscript j and simply write $\mathcal{T}$, θ, and $\delta(t)$.

Difficulty perception. We assume that the principal perfectly observes task difficulty for humans $\delta^H(t_k)$ but that his perception for AI's difficulty is given by Assumption 2 in the main text.

Micro-foundation for Assumption 2. Assume the principal does not perfectly observe

$\delta^A(t)$ but receives a noisy signal of AI difficulty,

$$z_t = \delta^A(t) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0, (\sigma^s)^2\big).$$

Human difficulty is a natural anchor when the principal assesses AI difficulty, so we assume a normal prior centered around δ^H, i.e., $\delta_0^A(t) \sim \mathcal{N}\big(\delta^H(t), (\sigma^H)^2\big)$. By conjugacy, the posterior mean of AI difficulty is

$$\tilde{\delta}^A(t) = \lambda\, \delta^H(t) + (1-\lambda)\, z_t, \qquad \text{where} \quad \lambda = \frac{(\sigma^s)^2}{(\sigma^s)^2 + (\sigma^H)^2}.$$

Taking expectations with respect to the signal noise yields

$$\mathbb{E}[\tilde{\delta}^A(t) \mid \delta^H(t), \delta^A(t)] = \lambda\, \delta^H(t) + (1-\lambda)\, \delta^A(t),$$

which is the functional form in Assumption 2 once we plug the posterior mean difficulty into the mapping p^A. Note that λ is increasing in the precision of the prior, $1/(\sigma^H)^2$, and decreasing in the precision of the signal, $1/(\sigma^s)^2$.

Human Projection. Together, Assumptions 1 and 2 constitute what we call *Human Projection* (HP): a tendency to rely on human features when evaluating AI performance. We distinguish two components. First, *Ability–Model projection* refers to applying a human-relevant domain structure and the associated unidimensional Ability Model when forming beliefs about AI performance within $\mathcal{T}$. Second, *Difficulty projection* refers to anchoring perceived AI difficulty on human difficulty, as captured by Assumption 2 through $\lambda > 0$. Misspecification can arise through either component. Ability–Model projection is misspecified if the inference structure that is valid for humans within $\mathcal{T}$ does not align with AI's true performance structure (e.g., because the relevant AI domains differ or because performance is not well summarized by a single within-domain ability). Difficulty projection is misspecified if AI task difficulties within $\mathcal{T}$ are not well-approximated by human difficulties, so that δ^H provides a misleading anchor for δ^A.

We let priors over ability differ across agents ($G^H \neq G^A$), which would happen if one

has higher uncertainty about AIs, or thinks machines are generally better than humans in the relevant domain. We also allow the mappings themselves (p^A and p^H) to be agent-specific. The Ability Model imposes only a light structure on updating: it assumes a consistent ranking of tasks by difficulty and of agents by ability within $\mathcal{T}$, which is natural when the principal observes few performance signals relative to the number of tasks and cannot feasibly track task-specific success rates. It reflects the logic behind the use of "tests" for students or job candidates: strong (and often reliable) inferences about the quality of the candidate are made on the basis of few signals, because the tasks the candidate is tested on are representative of the relevant domain.

The statements of predictions as in the main text are thus:

Proposition 1. *Under projection ($\lambda > 0$), the expected success rate is decreasing in δ^H for both humans and AI.*
Formally, $\forall i \in \{H, A\}$:

$$\frac{\partial \mathbb{E}_{G^i}[p^i(\theta^i, \tilde{\delta}^i)]}{\partial \delta^H} \leq 0.$$

Proof of Proposition 1. Fix $i \in \{H, A\}$ and a task with human difficulty δ^H. For humans, we have $\tilde{\delta}^H = \delta^H$ by definition. For AI, Assumption 2 implies that $\tilde{\delta}^A(\delta^H)$ is weakly increasing in δ^H, with $\frac{\partial \tilde{\delta}^A}{\partial \delta^H} = \lambda \in [0, 1]$. By Assumption 1, for each agent i the success function $p^i(\theta^i, \delta)$ is continuously differentiable and strictly decreasing in δ for all θ^i in the support of G^i. We then have

$$\frac{\partial}{\partial \delta^H} \mathbb{E}_{G^i}\left[p^i\left(\theta^i, \tilde{\delta}^i(\delta^H)\right)\right] = \mathbb{E}_{G^i}\left[\frac{\partial p^i}{\partial \delta}\left(\theta^i, \tilde{\delta}^i(\delta^H)\right) \frac{\partial \tilde{\delta}^i}{\partial \delta^H}\right] \leq 0.$$

since for $\partial \tilde{\delta}^i / \partial \delta^H \geq 0$ for both humans and AI, and $\partial p^i / \partial \delta < 0$ by Assumption 1. □

Proposition 2. *Consider any two tasks t^{δ^-} (easier) and t^{δ^+} (harder), with $\delta^- < \delta^+$, within the domain $\mathcal{T}$. Given observed performance x, let $\Pr(t = 1 \mid x) \equiv \mathbb{E}_{G_{|x}}\big(p(\theta, \delta(t))\big)$ denote posterior success rates. Then, under Human Projection and for any prior G and any task $t \in \mathcal{T}$:*

1. $\Pr(t = 1 \mid t^{\delta^-} = 1) \leq \Pr(t = 1 \mid t^{\delta^+} = 1)$,
2. $\Pr(t = 1 \mid t^{\delta^-} = 0) \leq \Pr(t = 1 \mid t^{\delta^+} = 0)$.

Proposition 2 holds for posterior rates on any task $t \in \mathcal{T}$. Through the MLRP, the posterior distribution on ability following an observed success on the harder task first-order stochastically dominates the one following a success on the easier task (and conversely for failures). This mechanism is driven by Ability–Model projection: once performance within $\mathcal{T}$ is summarized by a single latent ability parameter, observations on one task discipline beliefs about performance on all other tasks through the common mapping $p(\theta, \delta)$. The types of extrapolations highlighted by Proposition 2 are only an issue for belief misspecification if the *perceived* co-variation of performance and task features departs from the *actual* co-variation—for instance, if AI performance does not vary with task difficulty the way that human performance does, or if the unidimensional Ability Model itself is a poor description of AI performance within $\mathcal{T}$. Human Projection is not limited to difficulty or to AI as agent: in our conclusion we discuss examples of other projections which could be motivation for future work.

Proof of Proposition 2. We prove part (1); the argument for part (2) is analogous, using the failure likelihood $1 - p(\theta, \delta)$ instead of $p(\theta, \delta)$.

Let $\Theta = [\underline{\theta}, \overline{\theta}]$ and prior density g on Θ. After observing a success on a task of difficulty δ, Bayes' rule implies that the posterior CDF of ability is

$$G(\theta \mid t^{\delta} = 1) = \frac{\int_{\underline{\theta}}^{\theta} g(s)p(s,\delta)\,ds}{\int_{\underline{\theta}}^{\overline{\theta}} g(u)p(u,\delta)\,du} = \frac{1}{1 + R(\theta,\delta)},$$

where

$$R(\theta,\delta) := \frac{\int_{\theta}^{\overline{\theta}} g(s)p(s,\delta)\,ds}{\int_{\underline{\theta}}^{\theta} g(u)p(u,\delta)\,du} = \int_{\theta}^{\overline{\theta}} \left(\int_{\underline{\theta}}^{\theta} \frac{g(u)p(u,\delta)}{g(s)p(s,\delta)}\,du \right)^{-1} ds. \tag{1}$$

By Assumption 1 (MLRP in θ with respect to δ), for any $u \leq s$ the likelihood ratio $\frac{p(u,\delta)}{p(s,\delta)}$ is strictly decreasing in δ. Hence for each fixed $s \geq \theta$, the inner integral in (1) is strictly decreasing in δ, so its inverse is strictly increasing in δ and thus $R(\theta, \delta)$ is strictly increasing in δ. It follows that $G(\theta \mid t^{\delta} = 1)$ is strictly decreasing in δ for all θ. Thus for $\delta^{+} > \delta^{-}$:

$$G(\theta \mid t^{\delta^{+}} = 1) \leq G(\theta \mid t^{\delta^{-}} = 1) \quad \forall \theta,$$

with strict inequality on a set of θ of positive measure. This is precisely $G_{|t^{\delta^+}=1} \succ_{\text{FOSD}} G_{|t^{\delta^-}=1}$, which, combined with the fact that $p(\theta, \delta(t))$ is increasing in θ for all tasks $t \in \mathcal{T}$, implies $\Pr(t = 1 \mid t^{\delta^-} = 1) \leq \Pr(t = 1 \mid t^{\delta^+} = 1)$. This proves part (1). □

A.2 Framework in Field Context

Setup. Assume a similar setup to that of the basic framework, where a principal is trying to predict the performance of an agent over a domain of tasks.[17] Let $\mathcal{Q}$ and $\mathcal{A}$ denote the (finite) sets of questions and answers: for each question $q_i \in \mathcal{Q}$, assume there exists a subset of useful answers $A_i \subset \mathcal{A}$. Performance is observed at the question–answer level: given a question–answer pair (q_i, a_j), the principal's utility is given by $\overline{u} > 0$ if $a_j \in A_i$, and 0 otherwise. Assume each pair of answers (a, a') has a degree of *similarity* given by $S : \mathcal{A} \times \mathcal{A} \to [0, 1]$, with $S(a, a) = 1$ for all $a \in \mathcal{A}$. The function S captures the human similarity of answers, depending on semantic overlap, contextual proximity, shared meaning, etc. Define the *reasonableness* of answer a_j to question q_i by

$$r_{ij} \;:=\; \max_{a \in A_i} S(a_j, a) \in [0, 1].$$

In words, an answer's reasonableness is given by its similarity to the most similar useful answer. Among useful answers ($a_j \in A_i$), we have $r_{ij} = 1$ by construction. This adapted framework thus focuses on inference from *failures*: cases where the agent misunderstands the question and provides an answer that is not useful ($a_j \notin A_i$). Reasonableness is simply a particular way to measure similarity to useful answers: it is the most "forgiving" one, because it considers the maximum possible similarity of the (useless) answer to any useful answer.

For each question q_i and ability θ, let $\tilde{p}(\theta, r)$ denote the probability that the agent's answer to q_i has reasonableness $r \in [0, 1]$. We assume that all useful answers are symmetric: for any $a_j \in A_i$, the probability of producing that specific useful answer is $\tilde{p}(\theta, 1)$. Hence the probability of giving any useful answer to q_i is $P_\theta(\text{useful} \mid q_i) = \sum_{a_j \in A_i} \tilde{p}(\theta, 1) =$

[17] The agent belongs to a population which can be ordered by ability θ as before. We omit agent subscripts for clarity.

$|A_i|\,\tilde{p}(\theta,1)$, and the probability of a useless answer is $P_\theta(\text{useless} \mid q_i) = \sum_{a_j \notin A_i} \tilde{p}(\theta, r_{ij})$. Conditional on θ, the distribution of an answer depends on its content only through its reasonableness r_{ij}; thus r_{ij} is a sufficient statistic for the answer when updating beliefs about θ. We can now state the analogue of the MLRP part of Assumption 1 in this context.

Assumption 4. *For any $\theta' > \theta$ and any $r' > r$, we have $\frac{\tilde{p}(\theta',r')}{\tilde{p}(\theta',r)} > \frac{\tilde{p}(\theta,r')}{\tilde{p}(\theta,r)}$.*

This assumption says that the relative probability of giving a *more* reasonable answer (higher r) is higher for higher-ability agents. As before, denote the principal's posterior over the agent's ability by $G_{|x}$, given data $x = (r_{ij}^1, \ldots, r_{ij}^T)$ consisting of observed reasonableness scores. The subjective probability of providing a useful answer to q_i given data x is then

$$p(q_i \mid x) \;\equiv\; \mathbb{E}_{G_{|x}}\big[P_\theta(\text{useful} \mid q_i)\big] = \mathbb{E}_{G_{|x}}\big[|A_i|\,\tilde{p}(\theta,1)\big].$$

We thus obtain the following result about belief updating, which is the formal statement of Prediction 3 in the main text.

Proposition 3. *Let $r' > r$. Then for any $q_i \in \mathcal{Q}$, $p(q_i \mid r') > p(q_i \mid r)$.*

MLRP implies that observing a less reasonable failure (lower r) induces stronger negative inference over θ. Since $p(q_i \mid x)$ is increasing in θ, this translates into a lower perceived chance of providing a useful answer after a less reasonable answer. Further assuming that trust in the agent and willingness to ask questions are increasing functions of the perceived likelihood of a useful answer, we obtain the remaining parts of Prediction 3 in the main text.

Proof of Proposition 3. Fix a question q_i and ability level θ. As noted above, $P_\theta(\text{useful} \mid q_i) = |A_i|\,\tilde{p}(\theta,1)$, and $P_\theta(\text{useless} \mid q_i) = \sum_{a_j \notin A_i} \tilde{p}(\theta, r_{ij})$. Consider the odds ratio

$$\frac{P_\theta(\text{useless} \mid q_i)}{P_\theta(\text{useful} \mid q_i)} = \frac{\sum_{a_j \notin A_i} \tilde{p}(\theta, r_{ij})}{|A_i|\,\tilde{p}(\theta,1)} = |A_i|^{-1} \sum_{a_j \notin A_i} \frac{\tilde{p}(\theta, r_{ij})}{\tilde{p}(\theta,1)}.$$

For any useless answer $a_j \notin A_i$ we have $r_{ij} < 1$. Applying Assumption 4 with $(r', r) = (1, r_{ij})$, for any $\theta' > \theta$: $\frac{\tilde{p}(\theta',1)}{\tilde{p}(\theta',r_{ij})} > \frac{\tilde{p}(\theta,1)}{\tilde{p}(\theta,r_{ij})} \iff \frac{\tilde{p}(\theta',r_{ij})}{\tilde{p}(\theta',1)} < \frac{\tilde{p}(\theta,r_{ij})}{\tilde{p}(\theta,1)}$. Thus, for each $a_j \notin A_i$, the ratio $\tilde{p}(\theta, r_{ij})/\tilde{p}(\theta, 1)$ is strictly decreasing in θ, and therefore the odds ratio is strictly decreasing in θ. It follows that $P_\theta(\text{useful} \mid q_i)$ is strictly increasing in θ. Let $G_{|r}$ and $G_{|r'}$ denote the posterior distributions of θ after observing a single answer of reasonableness r and r' respectively. By Assumption 4 and the same MLRP argument as in Proposition 2, we have $r' > r \implies G_{|r'} \succ_{\text{FOSD}} G_{|r}$. Since the expectation of any increasing function is higher under an FOSD-dominating distribution, we obtain

$$p(q_i \mid r') = \mathbb{E}_{G_{|r'}}[P_\theta(q_i)] > \mathbb{E}_{G_{|r}}[P_\theta(q_i)] = p(q_i \mid r),$$

where $P_\theta(q_i) := P_\theta(\text{useful} \mid q_i)$, which completes the proof. □

A.3 Medium-Run Adoption Framework

A.3.1 Proof of Main Result

Proof of Theorem 1. We first derive KL minimization and study how beliefs induce best-responses in terms of adoption decisions. Then, we use those best responses to prove statements (i)-(iii) of the Theorem. Fix $\mathbf{p}^A = (p_1^A, \ldots, p_n^A)$ and the human benchmark θ^H with $p_i^H = p(\theta^H, \delta_i)$.

Step 1: KL minimization. For any nonempty adoption set $x \subseteq \{1, \ldots, n\}$,

$$K(x, \theta) = \sum_{i \in x} \Big[p_i^A \ln \frac{p_i^A}{p(\theta, \delta_i)} + (1 - p_i^A) \ln \frac{1 - p_i^A}{1 - p(\theta, \delta_i)} \Big].$$

Differentiating with respect to ability gives

$$\frac{\partial K(x, \theta)}{\partial \theta} = \sum_{i \in x} w_i(\theta) \left(p(\theta, \delta_i) - p_i^A \right), \quad w_i(\theta) \equiv \frac{\partial p(\theta, \delta_i)/\partial \theta}{p(\theta, \delta_i)\big(1 - p(\theta, \delta_i)\big)}.$$

Assumption 1 implies $p(\theta, \delta_i) \in (0, 1)$ and is strictly increasing in θ, hence $w_i(\theta) > 0$. Each term is strictly convex in $p(\theta, \delta_i)$ and $p(\theta, \delta_i)$ is strictly monotone in θ, so $K(x, \cdot)$ is

strictly convex and has a unique minimizer $\theta^*(x)$. Evaluating the derivative at θ^H,

$$\frac{\partial K(x,\theta^H)}{\partial \theta} = \sum_{i\in x} w_i(\theta^H)\left(p_i^H - p_i^A\right) = \sum_{i\in x} w_i(\theta^H)p_i^H - \sum_{i\in x} w_i(\theta^H)p_i^A.$$

By strict convexity, the sign at θ^H pins down the KL minimizer:

$$\sum_{i\in x} w_i(\theta^H)p_i^A \begin{cases} > \sum_{i\in x} w_i(\theta^H)p_i^H & \Rightarrow\ \theta^*(x) > \theta^H, \\ < \sum_{i\in x} w_i(\theta^H)p_i^H & \Rightarrow\ \theta^*(x) < \theta^H, \\ = \sum_{i\in x} w_i(\theta^H)p_i^H & \Rightarrow\ \theta^*(x) = \theta^H. \end{cases}$$

The weights of Theorem 1 are exactly $w_i(\theta^H)$, so the hyperplane $\mathcal{H}$ is the set of $\mathbf{p}$ for which $\theta^*(\{1,\ldots,n\}) = \theta^H$.

Step 2: Best responses given beliefs. Given belief θ, the subjective success of AI on task i is $p(\theta,\delta_i)$ and that of the human is $p_i^H = p(\theta^H,\delta_i)$. By Assumption 1, $p(\theta,\delta_i)$ is strictly increasing in θ, hence for all i,

$$\theta > \theta^H \Rightarrow p(\theta,\delta_i) > p_i^H, \quad \theta < \theta^H \Rightarrow p(\theta,\delta_i) < p_i^H.$$

Because the objective is separable across tasks, the best response is $\theta > \theta^H \Rightarrow x = \{1,\ldots,n\}$ and $\theta < \theta^H \Rightarrow x = \emptyset$. On the knife-edge $\theta = \theta^H$, the principal is indifferent task-by-task, and we adopt a consistent tie-breaking rule (e.g., assign knife-edge tasks to AI when $\theta^*(x) \geq \theta^H$ and to human when $\theta^*(x) \leq \theta^H$).

Step 3: Partial adoption is never a BkNE (part (i)). Take any nonempty proper x with $\emptyset \subsetneq x \subsetneq \{1,\ldots,n\}$. In a Berk–Nash equilibrium with action x, the belief must be $\theta(x) = \theta^*(x)$. By Step 1, $\theta^*(x)$ is either $> \theta^H$, $< \theta^H$, or $= \theta^H$ (knife-edge). By Step 2:

$$\theta^*(x) > \theta^H \Rightarrow \text{best response is } \{1,\ldots,n\}, \quad \theta^*(x) < \theta^H \Rightarrow \text{best response is } \emptyset,$$

and if $\theta^*(x) = \theta^H$ the tie-breaking rule again yields either full or no adoption, never x. Hence no partial adoption x can be both KL-consistent and optimal, so partial adoption

is never a Berk–Nash equilibrium.

Step 4: Full adoption above $\mathcal{H}$ (part (ii)). Consider $x = \{1, \ldots, n\}$. From Step 1 applied to this x,

$$\sum_{i=1}^{n} w_i(\theta^H) p_i^A > \sum_{i=1}^{n} w_i(\theta^H) p_i^H \quad \iff \quad \theta^*(\{1, \ldots, n\}) > \theta^H.$$

When $\mathbf{p}^A$ is above $\mathcal{H}$ (left-hand inequality holds), the KL-minimizing belief under full adoption satisfies $\theta^*(\{1, \ldots, n\}) > \theta^H$, and by Step 2 full adoption is then the unique best response. Hence $(x, \theta(x)) = (\{1, \ldots, n\}, \theta^*(\{1, \ldots, n\}))$ is a BkNE. Furthermore, by Step 3 partial adoption cannot be a BkNE, and no adoption requires beliefs under which humans are (weakly) preferred everywhere, which is not possible when $\mathbf{p}^A$ is above $\mathcal{H}$. Thus, among equilibria with experimentation, full adoption is the unique BkNE when $\mathbf{p}^A$ lies strictly above $\mathcal{H}$.

Step 5: No adoption below $\mathcal{H}$ and under pessimistic priors (part (iii)). Now suppose $x = \{1, \ldots, n\}$ and $\mathbf{p}^A$ lies strictly below $\mathcal{H}$, i.e. $\sum_{i=1}^{n} w_i(\theta^H) p_i^A < \sum_{i=1}^{n} w_i(\theta^H) p_i^H$. By Step 1 this implies $\theta^*(\{1, \ldots, n\}) < \theta^H$, and by Step 2 the best response is $x = \emptyset$. Together with Step 3, this means that any equilibrium with experimentation converges to no adoption, and $x = \emptyset$ is a BkNE when $\mathbf{p}^A$ lies strictly below $\mathcal{H}$.

Finally, $K(\emptyset, \theta) \equiv 0$ for all θ: any belief is KL-minimizing without experimentation. For a prior G^A sufficiently pessimistic such that $\mathbb{E}_{G^A}[p(\theta^A, \delta_i)] \leq p_i^H \ \ \forall i$, no adoption is both ex ante optimal and a best response regardless of the true $\mathbf{p}^A$. A sufficient condition for it is $\mathbb{E}_{G^A}[\theta^A] \leq \theta^H$ when $p(\theta, \delta)$ is concave in θ, since then by Jensen's inequality and monotonicity, $\mathbb{E}_{G^A}[p(\theta^A, \delta_i)] \leq p(\mathbb{E}_{G^A}[\theta^A], \delta_i) \leq p(\theta^H, \delta_i) = p_i^H$.

In sum, we have shown that: (i) partial adoption is never a BkNE; (ii) full adoption is the unique BkNE with experimentation when $\mathbf{p}^A$ lies strictly above $\mathcal{H}$; (iii) no adoption is a BkNE when $\mathbf{p}^A$ lies strictly below $\mathcal{H}$ and can also be sustained above $\mathcal{H}$ in the no-data case, under sufficiently pessimistic priors. □

A.3.2 Extension: Dynamics of Adoption

We extend the static model of Section 6 to characterize the adoption path under Human Projection as AI technology improves over time, comparing it to the optimal benchmark.

Setup. Consider a continuous-time model over $\tau \in [0, \overline{\tau}]$. In each period τ, there exists a convex set $\mathcal{P}(\tau) \subseteq [0,1]^n$ of feasible AI success-rate vectors $\mathbf{p}^A$. We assume $\mathcal{P}(0) = \{\mathbf{0}\}$ and $\mathcal{P}(\overline{\tau}) = [0,1]^n$, representing technological progress from no capability to universal capability. The correspondence $\mathcal{P} : [0, \overline{\tau}] \rightrightarrows [0,1]^n$ satisfies:

1. Monotonicity: (i) For any $\tau' > \tau$, $\mathcal{P}(\tau) \subseteq \mathcal{P}(\tau')$; (ii) For any $\mathbf{p} \in \mathcal{P}(\tau)$, if $\mathbf{p}'$ is dominated by $\mathbf{p}$ (i.e., $p'_i \leq p_i$ for all i), then $\mathbf{p}' \in \mathcal{P}(\tau)$.

2. Continuity: The correspondence $\mathcal{P}(\tau)$ is continuous (both upper and lower hemi-continuous) in τ.

Denote the *technological frontier* (set of Pareto-undominated success-rate vectors) in period τ by $\mathcal{P}^*(\tau) \equiv \{\mathbf{p} \in \mathcal{P}(\tau) : \nexists \mathbf{p}' \in \mathcal{P}(\tau) \text{ s.t. } \mathbf{p}' \geq \mathbf{p}, \mathbf{p}' \neq \mathbf{p}\}$, and assume $|\mathcal{P}^*(\tau)| > 1$ for all $\tau > 0$.[18] We characterize the optimal adoption path absent projection, then compare it to the equilibrium path under Human Projection.

Theorem 2 (Optimal Adoption Path)**.** *Let $\mathbf{p}^H \in (0,1)^n$. There exist $\tau_1, \tau_2 \in (0, \overline{\tau})$ with $\tau_1 < \tau_2$ such that:*

1. *For $\tau \in [0, \tau_1]$, no adoption is optimal.*
2. *For $\tau \in (\tau_1, \tau_2]$, partial adoption is optimal for a non-empty subset of $(\tau_1, \tau_2]$.*
3. *For $\tau \in [\tau_2, \overline{\tau}]$, full adoption is optimal.*

The optimal path follows an intuitive progression: as AI improves, it first surpasses human performance on some tasks (triggering partial adoption), then eventually on all tasks (triggering full adoption). Under Human Projection, this pattern is disrupted:

[18] This ensures the frontier is not a singleton, which is plausible given that different technological approaches may achieve different task-specific success rates.

Theorem 3 (Adoption Path Under Human Projection)**.** *Under Human Projection with complete projection ($\lambda = 1$), there exists $\tilde{\tau} \in (\tau_1, \tau_2]$ such that:*

1. *If $\tau < \tilde{\tau}$, then for any $\mathbf{p}^A \in \mathcal{P}^*(\tau)$, no adoption is the unique Berk-Nash equilibrium.*
2. *If $\tau \geq \tilde{\tau}$, then there exists $\mathbf{p}^A \in \mathcal{P}^*(\tau)$ such that full adoption is a Berk-Nash equilibrium.*

This theorem establishes two comparative statics relative to the optimal path: (i) *delayed adoption*for $\tau \in [\tau_1, \tilde{\tau})$, some adoption is optimal yet no adoption is the unique BkNE; and (ii) *over-adoption*for $\tau \in [\tilde{\tau}, \tau_2]$, full adoption is a BkNE even though partial adoption dominates it (except in the knife-edge case $\tilde{\tau} = \tau_2$). Thus, conditional on adoption occurring, Human Projection leads principals to assign too many tasks to AI prematurely.

Proof of Theorem 2 (Optimal path with n tasks). Define $\overline{p}_i(\tau) \equiv \max\{p_i : \mathbf{p} \in \mathcal{P}(\tau)\}$ as the highest achievable success rate for task i in period τ.

(1): No adoption is initially optimal. Since $\mathcal{P}(0) = \{\mathbf{0}\}$, continuity of $\mathcal{P}(\tau)$ implies there exists a range where no adoption is optimal. Assume by contradiction that for all $\tau > 0$, $\mathcal{P}(\tau)$ includes some $\mathbf{p}$ with $p_i \geq p_i^H$ for some $i \in \{1, ..., n\}$. Without loss of generality, take $i = 1$. Consider the sequence $(a_k)_{k=1}^{\infty}$ where $a_k = 1/k \to 0$, and define $b_k = (p_1^H, 0, ..., 0) \in \mathcal{P}(a_k)$ for all k. Then $b_k \to b \equiv (p_1^H, 0, ..., 0)$. Upper hemicontinuity implies $b \in \mathcal{P}(0)$, but $(p_1^H, 0, ..., 0) \notin \mathcal{P}(0) = \{\mathbf{0}\}$, a contradiction.

(3): Full adoption is eventually optimal. Since $(1, ..., 1) \in \mathcal{P}(\overline{\tau})$ and $\mathbf{p}^H \in (0, 1)^n$, there exists a range where full adoption is optimal. Assume by contradiction that for all $\tau < \overline{\tau}$, either $\overline{p}_i(\tau) < p_i^H$ for some i. Let $a_k = \overline{\tau} - 1/k \to \overline{\tau}$ and $b = (1, ..., 1) \in \mathcal{P}(\overline{\tau})$. By lower hemicontinuity, there exists a subsequence (a_{k_j}) and sequence (b_j) with $b_j \in \mathcal{P}(a_{k_j})$ and $b_j \to (1, ..., 1)$. But if full adoption is never optimal for $\tau < \overline{\tau}$, there exists a lower bound on $\|b_j - (1, ..., 1)\|$, so $b_j \not\to (1, ..., 1)$, a contradiction.

(2): Partial adoption is optimal in-between. Define $\tau_i \equiv \inf\{\tau : p_i^H \in \mathcal{P}_i(\tau)\}$, where $\mathcal{P}_i(\tau) \equiv \{p_i : \mathbf{p} \in \mathcal{P}(\tau)\}$ is the projection onto the i-th coordinate. If $\tau_i < \tau_j$ for some $i \neq j$, then for $\tau \in [\tau_i, \tau_j]$, partial adoption (AI for task i only, if no other tasks have crossed) is optimal. It remains to show that not all τ_i coincide, i.e., $\tau_1 = \cdots = \tau_n \equiv \tau_H$ leads to a contradiction.

Suppose $\tau_1 = \cdots = \tau_n = \tau_H$. By the lemma below, $\mathbf{p}^H \in \mathcal{P}^*(\tau_H)$. Since $|\mathcal{P}^*(\tau_H)| > 1$,

there exists $\mathbf{p}' \in \mathcal{P}^*(\tau_H)$ with $\mathbf{p}' \neq \mathbf{p}^H$. As both are on the frontier, $p_i' > p_i^H$ for some i. Without loss of generality, take $i = 1$. Choose an open ball $B(\mathbf{p}')$ such that $p_1 > p_1^H$ for all $\mathbf{p} \in B(\mathbf{p}')$. By lower hemicontinuity, for $\tau < \tau_H$ sufficiently close to τ_H, we have $\mathcal{P}(\tau) \cap B(\mathbf{p}') \neq \emptyset$. But since $\tau_H = \inf\{\tau : p_1^H \in \mathcal{P}_1(\tau)\}$ and $\mathcal{P}(\tau) \subseteq \mathcal{P}(\tau')$ for $\tau < \tau'$ (monotonicity), we cannot have $p_1 > p_1^H$ for any $\mathbf{p} \in \mathcal{P}(\tau)$ when $\tau < \tau_H$, a contradiction.

Lemma 1. *At τ_H, $\mathbf{p}^H \in \mathcal{P}^*(\tau_H)$.*

Proof. Assume $\mathbf{p}^H \notin \mathcal{P}^*(\tau_H)$. Let $\tilde{\mathbf{p}} \in \mathcal{P}^*(\tau_H)$ strictly dominate $\mathbf{p}^H$. Then there exists an open neighborhood $N(\tilde{\mathbf{p}})$ such that all $\mathbf{p} \in N(\tilde{\mathbf{p}})$ strictly dominate $\mathbf{p}^H$. By monotonicity, if $\mathbf{p}^H \notin \mathcal{P}(\tau)$, then $N(\tilde{\mathbf{p}}) \cap \mathcal{P}(\tau) = \emptyset$. Since $\mathbf{p}^H \notin \mathcal{P}(\tau')$ for all $\tau' < \tau_H$, we have $N(\tilde{\mathbf{p}}) \cap \mathcal{P}(\tau') = \emptyset$ for all $\tau' < \tau_H$. Thus all open neighborhoods of τ_H contain points whose image does not intersect $N(\tilde{\mathbf{p}})$, contradicting lower hemicontinuity. □

□

Proof of Theorem 3 (Path under projection with n tasks). Existence of initial and final ranges. By Theorem 1, for $\mathbf{p} = \mathbf{0}$, no adoption is the unique BkNE. Using continuity arguments analogous to Theorem 2, there exists a range $[0, \tau')$ where for all $\mathbf{p} \in \mathcal{P}^*(\tau)$, no adoption is the unique BkNE. Similarly, for $\mathbf{p} = (1, ..., 1)$, full adoption is a BkNE, and there exists a range $[\tau'', \overline{\tau}]$ where full adoption is a BkNE for all $\mathbf{p} \in \mathcal{P}^*(\tau)$.

Existence and uniqueness of $\tilde{\tau}$. Define $\tilde{\tau} \equiv \inf\{\tau : \mathcal{P}^*(\tau) \cap \{\mathbf{p} : \mathbf{p} \text{ strictly above } \mathcal{H}\} \neq \emptyset\}$ as the first time the frontier crosses above the hyperplane $\mathcal{H}$ from Theorem 1. If for some τ, there exists $\mathbf{p} \in \mathcal{P}(\tau)$ lying above $\mathcal{H}$, then by monotonicity $\mathbf{p} \in \mathcal{P}(\tau')$ for all $\tau' > \tau$. By Theorem 1, for $\tau < \tilde{\tau}$, all $\mathbf{p} \in \mathcal{P}^*(\tau)$ lie below $\mathcal{H}$ and no adoption is the unique BkNE. For $\tau \geq \tilde{\tau}$, some $\mathbf{p} \in \mathcal{P}^*(\tau)$ lies above $\mathcal{H}$ and full adoption is a BkNE.

Verification that $\tilde{\tau} \in (\tau_1, \tau_2]$. Using notation from Theorem 2, not all τ_i coincide. Let $\tau_1 = \min_i \tau_i$ and $\tau_2 = \max_i \tau_i$. Without loss of generality, assume $\tau_1 = \tau_j$ and $\tau_2 = \tau_k$ for some $j \neq k$. At time τ_1, the frontier $\mathcal{P}^*(\tau_1)$ first touches $\mathbf{p}^H$ in coordinate j, so $\overline{p}_i(\tau_1) < p_i^H$ for all $i \neq j$. Thus all points in $\mathcal{P}^*(\tau_1)$ lie strictly below $\mathcal{H}$, implying $\tilde{\tau} > \tau_1$.

For the upper bound, suppose $\tilde{\tau} > \tau_2$. At time τ_2, we have $\overline{p}_i(\tau_2) \geq p_i^H$ for all i, so $\mathcal{P}(\tau_2)$ contains points strictly above $\mathcal{H}$ (e.g., near $(1, ..., 1)$ by continuity and monotonicity). By definition of $\tilde{\tau}$, this means $\tilde{\tau} \leq \tau_2$, a contradiction. Therefore $\tilde{\tau} \in (\tau_1, \tau_2]$. □

B Performance in Mathematics

B.1 Task Dataset

We construct a benchmark of standardized mathematics problems from released items in the *Trends in International Mathematics and Science Study* (TIMSS). All released items were accessed through the TIMSS portal, and items for the 2015 and 2019 test waves were obtained through direct request to the IEA. To reduce benchmark contamination, a substantial subset of items was obtained through direct request to the IEA and is not easily accessible online.

We manually re-transcribe all selected items into a uniform format, while trying as much as possible to preserve wording, tables, and symbols used. We restrict attention to multiple-choice questions with a unique objectively correct answer (4 or 5 options, labeled A–E), which yields a binary performance measure. We excluded items with components which could not be processed in textual form by ChatGPT, e.g., with visual charts or geometric shapes.

Our final benchmark contains 414 items spanning 4^{th} grade (29%), 8^{th} grade (58%), and High School (13%), and covers the main topic categories in the TIMSS frameworks.

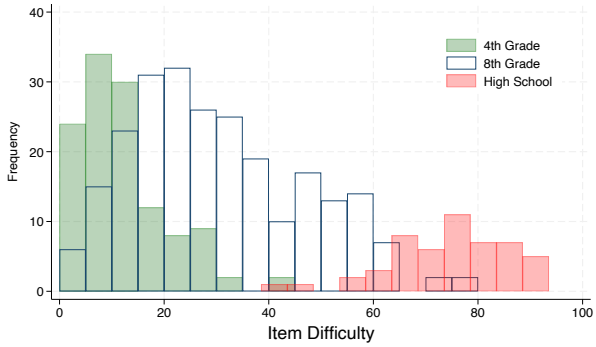
To allow for better comparison between human and AI performance, we experimented beforehand to make sure ChatGPT was able to correctly process inputs and was not tricked by minor formatting issues. We prompted ChatGPT with one task at a time, following the order used in the released items documents. Full logs of conversations are available upon request.

B.2 Human Performance

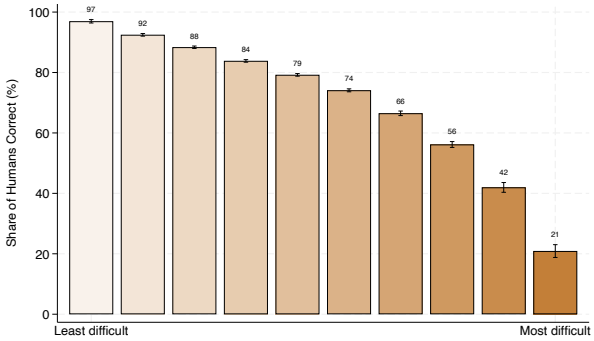
Descriptive Results. For human performance data, the number of answers per item ranges from 24 to 191, with an average of 41.5 and a median of 34. High-School problems tend to have more answers per question, while 8^{th}-grade problems have fewer, due to (i) the number of released problems not being evenly split by grade level (see Figure 7); (ii)

Figure 7: Distribution of Item Difficulty and Average Human Performance by Difficulty

(a) Histogram of Item DIfficulty



(b) Average Human Perf. by Difficulty Decile



Notes: **Left panel:** Overlaid histograms of item difficulty by grade level (4th grade, 8th grade, high school). The y-axis reports the number of questions and the x-axis reports human difficulty. **Right panel:** Average share of subjects answering items correctly within each decile of human difficulty. Deciles are constructed from the item difficulty measure; each decile contains around 41 items.

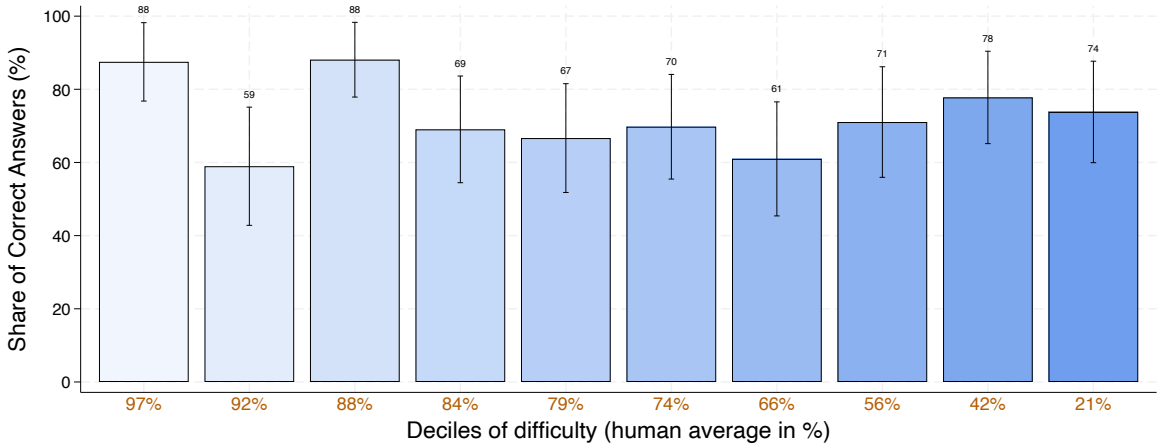
the beliefs experiment sampling H.S. items more heavily to ensure each subject sees the whole gradient of difficulty.

The average task difficulty in the sample is 30.3. There are 7 items with a difficulty level of 0 (meaning every subject who attempted it answered correctly) and the highest difficulty is 92. Panel A of Figure 7 shows the origin of questions for each difficulty level, while panel B plots the average share of correct answers for each difficulty decile: the 10% easiest items have a success rate of around 95%, while the 10% most difficult ones have around 21%.

B.3 Human-likeness of AI performance

Approach. We combine item-level human response distributions from the TIMSS benchmark with item-level correctness and answer choices for several LLMs of various AI labs (GPT-3.5 Turbo, GPT-4o, Claude 3.7, Mistral Large, Llama 3.1, o3-mini; all accessed through their respective APIs). For each of 414 items, we observe (i) the share of human test-takers answering correctly, which we treat as the item's empirical "easiness" for humans, and (ii) the full distribution of human answer choices (modal option and its share), alongside the model's chosen option and whether it matches the correct key.

Figure 8: ChatGPT 3.5 Performance by Human Difficulty Deciles (API)



Notes: The figure plots ChatGPT (3.5 Turbo, collected in May 2025 via the API) share of correct answers within each decile of human difficulty. Each decile contains between 40 and 42 items. Numbers under each bar report the average human success rate in the corresponding decile. Human difficulty is computed from an average of 41.5 answers per item.

We compare model outputs across multiple prompt variants: text prompts (standard prompt vs. one suppressing chain-of-thought by asking models to only output the final answer) and image prompts (a screenshot of the problem with a standardized layout).

Measures. We then compute several complementary notions of "human-likeness," all at the item level. First, we compare how model success co-varies with human success across items using two notions of correlation: the Spearman rank correlation between model correctness and average human correctness, capturing whether models tend to succeed on the same items humans find easier, and a tetrachoric correlation between model correctness and a binary "easy-for-humans" indicator (human accuracy at least 50%). Then, we summarize the "shape" of model performance as a function of human performance by performing OLS regressions of model correctness on human correctness at the item level and reporting the slope β, standard error, and R^2. Finally, we compute two measures of "agreement" between human and AI answers. The first "agreement rate" measures how much of the mass of the human distribution of answers lies in the model's chosen option, meaning it is high when the model tends to pick options that many humans pick; we compute it both overall (among all items) and conditional on the model picking the wrong option. The second is the probability the model chooses the

Table 5: Human-likeness of GenAI performance

		Correlations		OLS		Agreement Rate		AI Picks Human Mode	
Model	Mean Perf.	Spearman	Tetrachoric	β (s.e.)	R-squared	Overall	AI wrong	Overall	AI wrong
Text prompt (standard)									
GPT-3.5 Turbo	0.729	0.045	-0.051	0.048 (0.094)	0.001	0.568	0.107	0.718	0.390
GPT-4o	0.978	0.060	0.181	0.040 (0.031)	0.004	0.675	0.149	0.886	0.286
Claude 3.7	0.993	0.032	-1.000	0.007 (0.018)	0.000	0.679	0.146	0.894	0.333
Mistral Large	0.928	0.018	0.125	0.014 (0.055)	0.000	0.673	0.137	0.885	0.429
Llama 3.1	0.935	0.133	0.224	0.154 (0.052)	0.021	0.677	0.192	0.887	0.545
o3-mini	0.993	-0.075	-1.000	-0.025 (0.018)	0.005	0.678	0.041	0.896	0.500
Text prompt (no chain-of-thought)									
GPT-3.5 Turbo	0.432	0.089	0.181	0.228 (0.104)	0.011	0.376	0.112	0.428	0.335
GPT-4o	0.734	0.128	0.213	0.239 (0.093)	0.016	0.549	0.130	0.688	0.376
Claude 3.7	0.920	0.128	0.235	0.153 (0.057)	0.017	0.647	0.132	0.845	0.273
Mistral Large	0.763	0.243	0.351	0.452 (0.087)	0.061	0.584	0.157	0.728	0.448
Llama 3.1	0.696	0.258	0.353	0.515 (0.094)	0.068	0.589	0.126	0.736	0.390
o3-mini	0.995	-0.043	-1.000	-0.013 (0.015)	0.002	0.679	0.031	0.896	0.000
Image prompt									
GPT-4o	0.630	0.208	0.352	0.452 (0.100)	0.047	0.618	0.117	0.777	0.308
Claude 3.7	0.696	0.180	0.289	0.357 (0.096)	0.032	0.621	0.128	0.782	0.353
Mistral Large	0.621	0.225	0.355	0.477 (0.100)	0.052	0.604	0.138	0.759	0.338
Llama 3.1	0.589	0.198	0.214	0.390 (0.102)	0.034	0.590	0.147	0.720	0.471
o3-mini	0.693	0.245	0.348	0.486 (0.095)	0.060	0.623	0.117	0.787	0.278

Notes: Each row reports item-level "human-likeness" statistics comparing a model to humans on the same multiple-choice items ($n = 414$). Average human performance is 0.679 overall, to be compared with each model's mean performance. *Correlations* columns report Spearman's rank correlation between human and model correctness, and the tetrachoric correlation between model correctness and a human-defined "easy" indicator equal to one if human accuracy is at least 50%. *OLS* columns report the slope (s.e.) and R^2 from regressing model correctness on human accuracy across items. *Agreement Rate* columns report the average human mass on the model's chosen option (Overall), and conditional on model error (AI wrong). *AI Picks Human Mode* columns reports the probability the model selects the human modal option (Overall), and conditional on model error, the probability it selects the most common human *incorrect* option (AI wrong). Higher values generally indicate closer alignment with human difficulty and answer-choice patterns.

most common human answer, i.e., the human mode. We compute it both overall, and conditional on model error, the probability it selects the most common human *incorrect* option, to see whether humans and AIs tend to fail in the same way.

Results. Results are reported in Table 5. Under the standard text prompt, the most advanced models achieve extremely high mean accuracy (roughly 0.93 to 0.99), yet item-level alignment with human difficulty is close to zero (Llama 3.1 being the highest): Spearman correlations are tiny (often near 0 and sometimes negative), OLS slopes are near 0, and R^2 is essentially null. Thus, *which* items humans find easy versus hard barely predicts *which* items the models answer correctly—strong evidence that "difficulty" is not a shared ordering even when the model nearly saturates the benchmark. The tetrachoric boundary value ($\rho = -1$) for Claude 3.7 and o3-mini under the standard prompt reflects a near-separation, small-error regime and is substantively consistent with the raw counts: all observed errors (4/4 and 3/3) occur on human-easy items, with none on human-hard items.

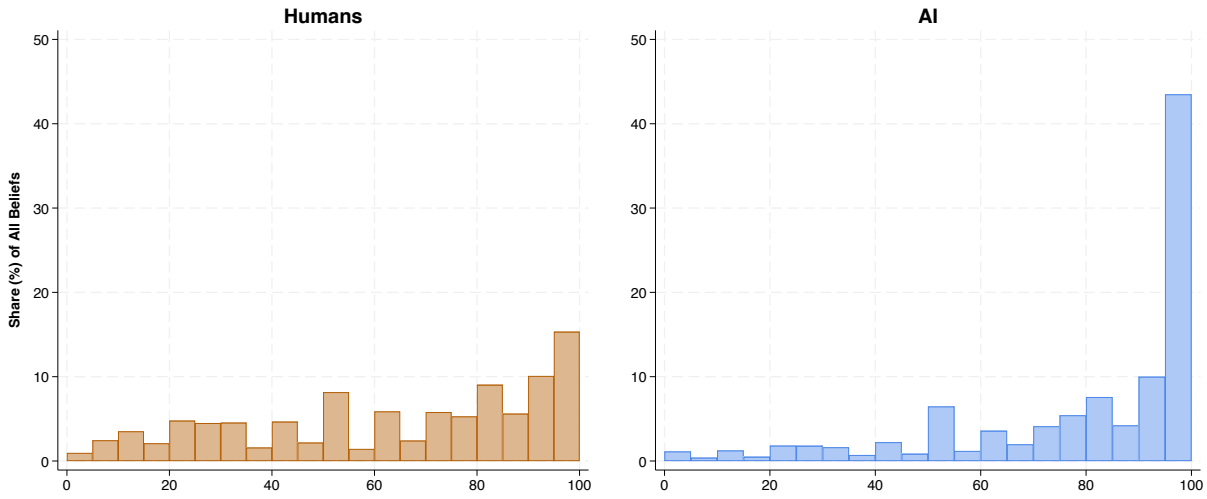
Overall answer-choice overlap is fairly high, but weakens markedly on mistakes: the agreement rate is large unconditionally yet much lower conditional on model error. A similar divergence is seen for agreement with the human mode. Although models often match the human modal answer overall, their mistakes are only partially human-like: conditional on error, agreement with the most common human wrong option is far from one and varies substantially across models (≈ 0.29 to 0.55 under the standard prompt), implying heterogeneous divergence in error modes. Finally, we note that prompting choices have large consequences for performance. Suppressing chain-of-thought substantially reduces accuracy for most models (except for o3-mini, consistent with this manipulation constraining visible output rather than internal reasoning). It tends to increase alignment with humans among difficulty measures (higher Spearman and steeper β, with larger R^2) but decrease it among answer-choice alignment measures (reduction in overall agreement rates). Image prompting further reduces average accuracy and, relative to the standard text prompt, increases correlation-based difficulty alignment while modestly reducing overall choice-based overlap.

These results, leveraging an unusually rich human benchmark, echo recent evidence that model "difficulty" and error patterns can be systematically misaligned with human intuitions even in strong models (Yang et al., 2024; Williams and Huckle, 2024; Berglund et al., 2023; Wang and Sun, 2025; McKenzie et al., 2023; Li et al., 2025).

C Beliefs Experiment

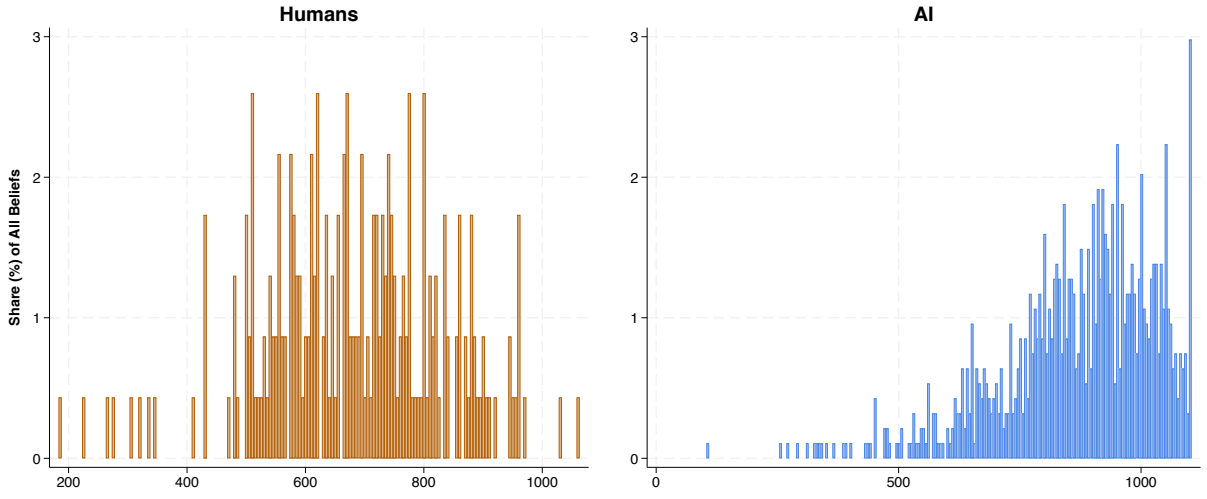
C.1 Additional Results

Figure 9: Histograms of Prior Beliefs in Performance



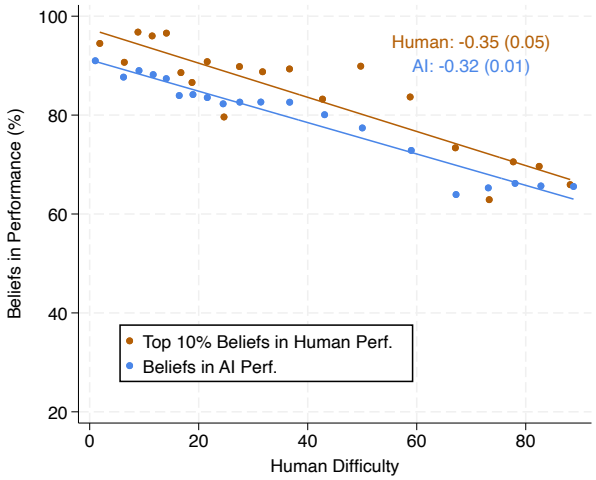
Notes: The figure plots histograms of prior beliefs in agent performance. On the x-axis are beliefs in %. On the y-axis is the share of all beliefs contained in each bin. $n = 2442$ for humans, and $n = 10021$ for AI.

Figure 10: Histograms of the Sum of Prior Beliefs in Performance



Notes: The figure plots histograms of the individual sum of their prior beliefs in agent performance: the maximum value is $11 \times 100 = 1100$. On the y-axis is shown the share of all beliefs contained in each bin.

Figure 11: Average Beliefs about AI vs. Optimistic Beliefs about Humans



Notes: This figure compares all beliefs about ChatGPT with a subset of the most optimistic subjects in the *Human* treatment. We order subjects by their average reported belief and keep the top 10% of subjects, then plot all of their beliefs. Sample sizes are $n = 245$ for humans, and $n = 10021$ for AI. The average difficulty of questions seen by a subject is added as control. We report coefficient estimates and their standard errors on the top right.

Separating ability priors from projection. In Figure 11 we plot the top 10% most optimistic subjects in *Human*, on the basis of mean reported belief, against the full sample in *AI*. As we control for the mean difficulty of tasks seen, this is a way to proxy for prior belief in agent ability. We find beliefs about performance to be highly similar, both in slope and levels, between optimistic subjects predicting humans and all subjects predicting AI. While not a causal exercise, this finding is consistent with subjects holding a more optimistic prior over AI ability which could partly explain differences noted in Figure 2.

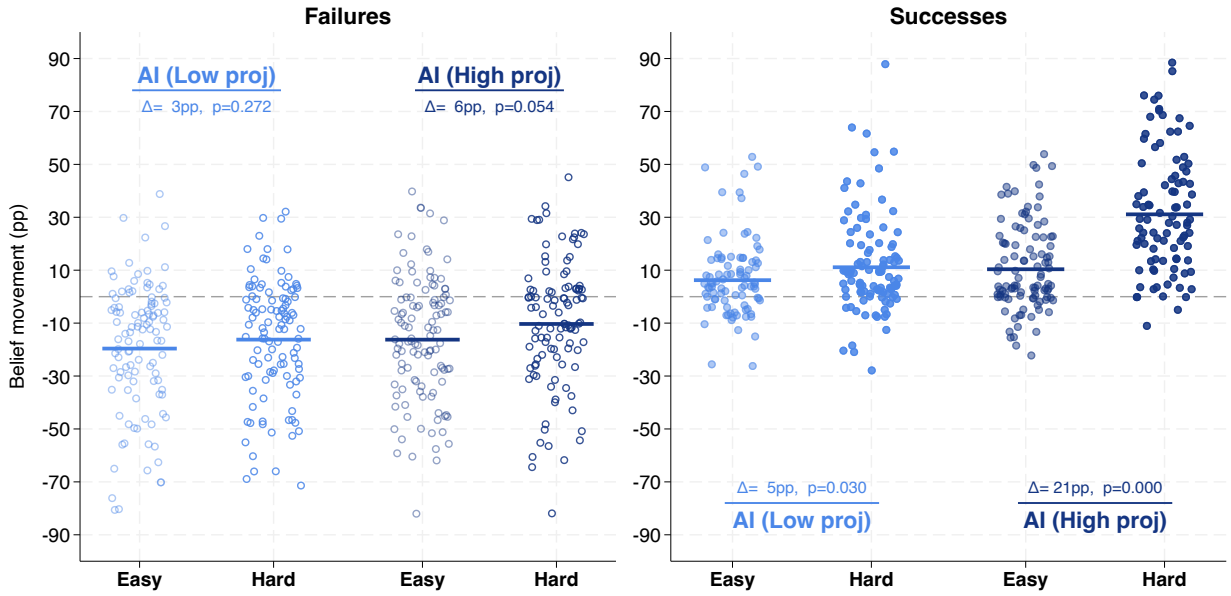
Belief Accuracy. We quantitatively assess the accuracy of subjects' beliefs by regressing actual performance on elicited beliefs. We estimate the following specification:

$$Y_t = c_0 + c_1 \text{Beliefs}_{it} + d_i + \epsilon_{it},$$

where Y_i is agent performance on task t and d_i are subject fixed effects. Estimated coefficients for both humans and AI are positive and significant: respectively 0.58 (0.01) and 0.06 (0.01), but there is stark contrast in predictive power: the R^2 is 0.391 for hu-

mans, while only 0.002 for AI. As subjects rely on human difficulty to predict (mostly uncorrelated) AI performance, their answers carry very little predictive power overall.

Figure 12: Belief Updating after Exogenous Signals of Performance



Notes: The figure presents jittered dot plots of belief movement, defined as the difference between posterior and prior, winsorized at percentiles 2.5 and 97.5. Sample sizes from left to right are $n = 107$, $n = 109$, $n = 114$, $n = 107$, $n = 89$, $n = 103$, $n = 103$, $n = 92$. Colored horizontal lines denote condition means; Δ are mean differences (hard vs easy) within failures and within successes, with p-values from two-sided tests of equality.

Heterogeneous updating by projection strength. We estimate within-subject the projection strength using an empirical Bayes approach, and split the sample into above and below median projection. We test the hypothesis that subjects projection more exhibit a stronger effect for Prediction 2, compared to those with lower projection. Results presented in Figure 12 are consistent with this hypothesis, and most of the differential effect appears concentrated among reactions to successes.

Table 6: Beliefs Experiment - Balance Table

		Humans		AI	
Group	**Variable**	*Mean*	*S.E.*	*Mean*	*S.E.*
Gender	Female	46.8	3.6	50.1	1.8
	Male	50.5	3.7	48.5	1.8
Race	White	73.4	3.2	78.6	1.5
	Black	13.8	2.5	12.8	1.2
Highest Education	HS or more	98.9	0.8	99.6	0.2
	College or more	62.8	3.5	58.9	1.8
Highest Math Class	High school or more	96.8	1.3	97.8	0.5
	College or more	46.3	3.6	47.3	1.8
	Graduate level or more	1.1	0.8	2.7	0.6
AI Familiarity	Only heard of AI	–	–	34.3	1.7
	Some understanding, never used	–	–	40.1	1.8
	Interacted with AI	–	–	24.0	1.5
	Mean Age	40.9	1.0	42.9	0.5
	Observations	188		784	

Notes: This table provides demographics of participants in the beliefs experiment, run in October 2023. The AI familiarity question was only asked in the *AI* treatment. The sample is for belief updating results: it excludes subjects failing comprehension checks and the updating comprehension question. Averages are qualitatively similar over the broader sample, and are consistent with overall demographics of Prolific participants.

D ParentData

D.1 Institutional Context and Interaction Data

This appendix provides institutional background and descriptive details for the ParentData setting that are not required for the main text exposition.

ParentData.org. ParentData.org is a self-described "data-driven guide through pregnancy, parenthood, and beyond." Created in 2020 by American economist Emily Oster—and launched as its own website in 2023—its goal is to translate scientific evidence into rigorous answers to questions asked by current or expecting parents. The platform provides multiple formats (articles, newsletters, podcasts) with several tiers of paid subscription. It also offers an AI chatbot service that is free to use and openly accessible.

User base. The pool of users is almost exclusively composed of people who are trying to conceive, currently expecting, or parents of young children. The website does not collect user demographics. A survey from July 2024 and discussions with the ParentData team

describe the prototypical user as (i) a woman in her 30s living in the U.S.; (ii) either expecting or a mother of young children; (iii) educated; and (iv) with higher income. We recruit all experimental participants on the basis of these demographic features.

The AI chatbot. The website hosts Dewey, an "AI librarian," which is a LLM-based AI chatbot.[19] Dewey was built by first summarizing parenting materials (books and articles) into a set of questions and answers, which were then verified and vetted by humans from the ParentData team. Dewey is then available to answer users' questions by matching each incoming query to the archive of pre-made questions using a confidence score, and displaying the highest-score answer if it exceeds a confidence threshold. The matching relies on a measure of textual similarity which is related but not identical to what humans understand by "similarity."[20]

Interaction data snapshot. Our interaction data consist of a snapshot of approximately 30,000 conversations, where a "conversation" is defined as one user query and one AI answer, occurring between December 31st 2023 and April 30th 2024. The Experimental Materials Appendix presents descriptive statistics for this dataset, including the most frequent questions asked to the chatbot.

D.2 Selecting matched conversation pairs

This appendix documents the pre-registered procedure used to construct matched *pairs* of real ParentData conversations that differ in the *human reasonableness* of the chatbot's misunderstanding while holding constant the user's intent and the (low) usefulness of the answer. Throughout, a "conversation" refers to one user query and one chatbot answer.

Step 1: Initial coding and candidate pairs.. We began from a corpus of roughly 22,000

[19]This service is provided by the Dewey Labs company, of which ParentData.org is a client. Dewey has since the time of the experiment been renamed "Petey."

[20]The matching process is a custom approach to vector embeddings, whose details were not disclosed. We confirm that the AI score correlates with various textual similarity measures. Below the confidence threshold, the chatbot either suggests related questions or displays the message: "I'm sorry. It looks like we don't have an answer in the ParentData archives. However, your question has caught our attention and will be shared with Emily and the team. It could be a great topic for a future newsletter!"

real ParentData conversations. We manually coded each conversation for: (i) *intent* (a short label capturing what the user is asking), and (ii) a binary indicator for whether the chatbot *misunderstood* the question (a proxy for the answer being unhelpful/useless for that specific query). We then created an initial list of candidate *same-intent* pairs in which *both* answers were coded as misunderstandings. This produced approximately 40 candidate pairs.

Step 2: Reasonableness ratings. We recruited participants from the population of interest (parents of young children or currently expecting) to rate the *reasonableness* of each misunderstood answer. The survey presented a single conversation at a time in a display mimicking the ParentData website. To avoid within-subject contrast effects, participants did not see both sides of a pair: we split each pair into two pools and randomly drew conversations from one pool per participant. Instructions emphasized that the object is not whether the answer is helpful, but whether the *specific misunderstanding* is a plausible human error. *Reasonableness prompt (verbatim).* "The AI gave answers that were deemed unhelpful. What do you think is the percent (%) chance that a **reasonable human** would misunderstand the questions in the way the AI did? Choose a % between 0 and 100."

Step 3: Usefulness ratings. In a separate study with the same target population and a nearly identical display, we elicited perceived *usefulness* of each answer for the specific question on a 1–5 scale. We included a small number of genuinely useful conversations in the rating pool both (i) to obtain usefulness ratings for the "success" conversations used in the main experiment and (ii) to avoid an unnatural rating environment in which all answers are useless, which could invite demand effects. Instructions stressed that usefulness is about answering the specific query, not about sounding generally sensible or containing good generic parenting advice. *Usefulness prompt (verbatim).* "Assess the answer's **usefulness**: does its content answer *that specific question?* **Read carefully:** some answers may *appear* useful at a glance, even though they are not! Indicate your answer on the [1–5] scale."

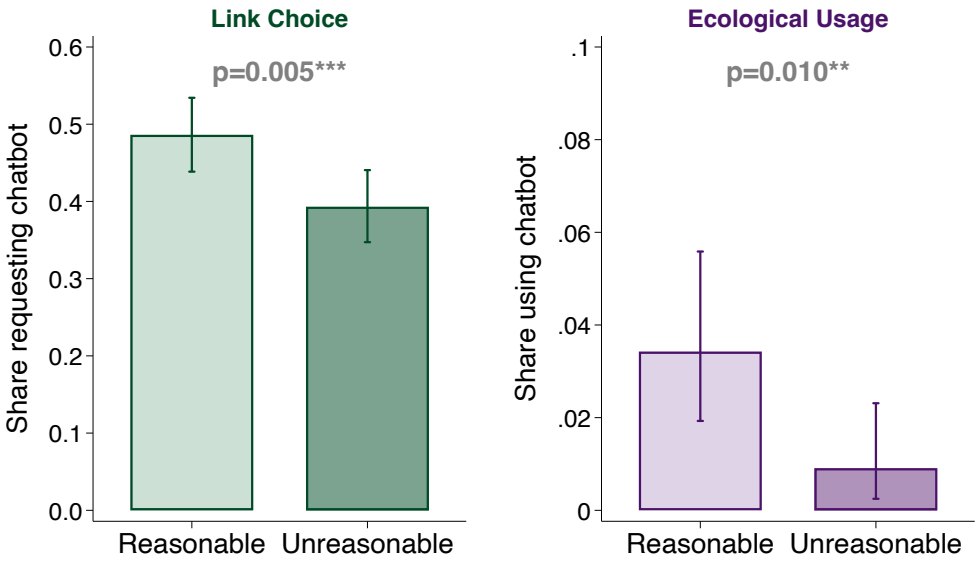
Final selection rule and implemented pairs. We retained five final pairs satisfying three criteria: (i) the two conversations within a pair share the same intent (from Step 1), (ii)

both sides are comparably unhelpful (same *median* usefulness, equal to 1 or 2), and (iii) the pair exhibits a large difference in *median* reasonableness. As pre-registered, medians are the primary matching criterion to reduce sensitivity to outliers and scale use. As robustness, we also checked that (a) average usefulness differences within a pair are not statistically significant at the 90% level, while (b) average reasonableness differences are statistically significant. We additionally set aside three "success" conversations (median usefulness 4 or 5) used as common baseline content in the engagement experiment.

Materials. The full list of conversations used in the experiment (five pairs plus the baseline successes), along with their usefulness and reasonableness ratings, is available is the separate Experimental Materials Appendix.

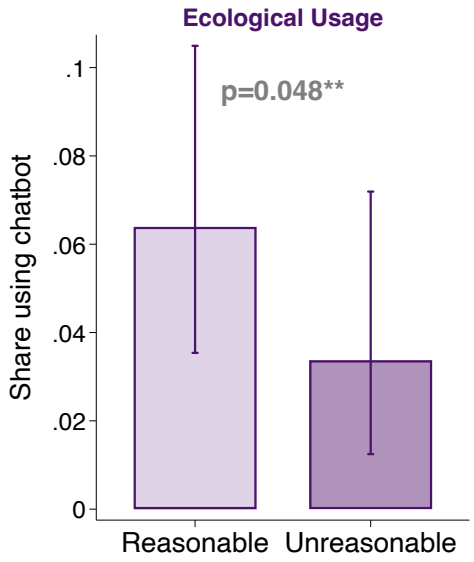
D.3 Additional Experimental Results

Figure 13: Treatment Effect on Engagement Excluding Subjects Familiar with ParentData



Notes: The figure plots the share of subjects in each treatment choosing to receive the link to the chatbot and actually using the chatbot, excluding subjects who reported prior familiarity with ParentData. Sample sizes are $n = 438$ for *Reasonable* and $n = 440$ for *Unreasonable*. p-values of two-sided tests of proportions are reported.

Figure 14: Treatment Effect on Engagement - Conditional on Link Request



Notes: The figure plots the share of subjects in each treatment who uses the chatbot post-experiment, conditional on requesting the chatbot link, and excluding subjects with prior familiarity with ParentData. Sample sizes are $n = 213$ for *Reasonable* and $n = 173$ for *Unreasonable*. p-value of two-sided test of proportions is reported.

Table 7: Demographics of Participants in Engagement Experiment

	Unreasonable		*Reasonable*	
	Value	std. err.	Value	std. err.
Gender				
Female	85.0	(1.7)	83.8	(1.7)
Male	14.1	(1.6)	16.0	(1.7)
Race				
White	72.7	(2.1)	74.7	(2.0)
Black	18.1	(1.8)	18.8	(1.8)
Education				
HS or more	99.8	(0.2)	99.3	(0.4)
Bachelor or more	59.3	(2.3)	59.6	(2.3)
Parenting status				
Trying to conceive	24.2	(2.0)	22.2	(2.0)
Currently expecting	6.8	(1.2)	7.3	(1.2)
Have children below 18	79.1	(1.9)	77.6	(2.0)
AI familiarity				
Only heard of AI/LLMs	98.5	(0.6)	97.8	(0.7)
General understanding	77.1	(2.0)	76.3	(2.0)
Interacted with AI	38.1	(2.3)	36.1	(2.3)
ParentData familiarity				
Knew of the website	3.1	(0.8)	2.9	(0.8)
Age (years)	34.3	(0.3)	34.0	(0.3)
Observations	454		451	

Notes: This table provides demographics of participants (current or expecting parents) in the engagement experiment. Samples recruited to rate the usefulness and reasonableness of conversations are distinct but highly similar in terms of demographics. "Value" reports cell mean and "std. err." reports the standard error of the mean.

Table 8: Misperceptions about human difficulty: proxy validity

Proxy: $\Delta^H \equiv p^H(hard) - p^H(easy)$	Estimate	s.e.	p-value	N
Panel A: Misperceptions proxy ($\Delta^H \geq 0$)				
Baseline wave mean (%)	56.3	(2.8)		426
Replication wave mean (%)	17.1	(3.5)		426
Replication – Baseline (pp)	-39.2	(4.5)	0.000	426
Panel B: Treatment does not affect the proxy				
Black Box effect in baseline (pp)	-0.7	(5.7)	0.903	426
Black Box effect in replication (pp)	3.1	(7.0)	0.655	426
Diff. across waves: $BlackBox \times Rep$ (pp)	3.8	(9.0)	0.671	426

Notes: The misperceptions proxy equals one if the subject reports prior human performance in the hard pool at least as high as in the easy pool (ties counted as misperceptions; they are rare and mostly in baseline: $N = 47$ overall: 43 baseline, 4 replication). Panel A reports wave means and the replication–baseline difference from a pooled regression of the proxy on indicators for Black Box and Replication. Panel B reports within-wave *Black Box* effects and the across-wave difference from a pooled regression of the proxy on indicators for Black Box, Replication and their interaction. Robust standard errors are in parentheses.

E Adoption Experiment

This appendix provides evidence on the role of perceptions of human task difficulty on final adoption choices, as well as additional results on adoption shares and delegation behavior, and the balance table of covariates.

E.1 Misperceptions of human difficulty and adoption

Proxy for misperceptions. Table 8 documents the large share of misperceptions in the baseline sample (56.1%), proxied by $\Delta^H \geq 0$: the difference between reported beliefs about *human* performance on easy vs hard tasks, *before* the delegation phase thus before receiving endogenous signals of performance. The replication wave, which made human task difficulty more salient within the same design, strongly reduces these misperceptions (to 17.1%) but they remain non-trivial. Importantly, the treatment does not affect the share of misperceptions neither in baseline nor replication, lending credibility to the proxy, and results are robust to classifying ties as correct perceptions.

Table 9: Belief-Conditional Treatment Effects on Adoption

	***Black Box* effect for participants with:**						
	Incorrect perception		Correct perception		**Difference**		
Outcome	Coefficient	s.e.	Coefficient	s.e.	Coeff.	s.e.	N
All-or-nothing adoption:	-15.9**	(6.8)	-15.3***	(5.7)	0.6	(8.8)	426
— Full Adoption	-18.8***	(6.6)	-16.9***	(5.3)	1.9	(8.5)	426
— No Adoption	2.8	(2.2)	1.5	(2.7)	-1.3	(3.5)	426
Partial adoption:	15.9**	(6.8)	15.3***	(5.7)	-0.6	(8.8)	426
— Only Hard	-6.5	(7.2)	12.4**	(6.2)	18.9**	(9.5)	426
— Only Easy	22.4***	(5.5)	2.9	(4.5)	-19.5***	(7.1)	426

Notes: Each row reports coefficients (in pp) from pooled linear regressions of the form $y = \alpha + \beta\, BB + \gamma\, Correct + \delta(BB \times Correct) + \rho\, Rep + X'\kappa + \varepsilon$, estimated separately for each adoption outcome y. BB, Rep, $Correct$, are indicators for Black Box framing, replication wave, and correct perception of human difficulty (= 1 if the subject reports belief in human performance to be lower on human-hard tasks than on human-easy tasks) respectively. Demographic controls X include age, gender, bachelor's degree indicator, white indicator, and self-reported AI familiarity. The "incorrect perception" column reports β, "correct" reports $\beta + \delta$, and "difference" reports δ, i.e., the change in the BB effect when moving from incorrect to correct perceptions. Robust standard errors are in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Belief-conditional partial adoption. Using this proxy, Table 9 presents the main evidence towards the role of misperceptions. Each row reports estimates of the *Black Box* treatment effect from pooled linear specifications of the form:

$$y = \alpha + \beta\, BB + \gamma\, Correct + \delta(BB \times Correct) + \rho\, Rep + X'\kappa + \varepsilon,$$

estimated separately for each adoption outcome y. BB, Rep, and $Correct$ are respectively indicators for Black Box, replication wave, and the proxy for correct perceptions of human difficulty. First, the overall reduction in all-or-nothing adoption is stable: *Black Box* reduces Full Adoption and increases partial adoption by similar magnitudes for correct and incorrect perceivers (interactions near zero). Second, misperceptions strongly moderate the *direction* of partial adoption. Among subjects with incorrect initial beliefs about human difficulty, *Black Box* reallocates mass primarily into the normatively wrong margin ("Only Easy" increases substantially), while "Only Hard" does not rise. Among subjects with correct priors, the reallocation flips: "Only Hard" increases, while

"Only Easy" is essentially unchanged. The interaction coefficients quantify this reversal directly, and the pattern is stable with and without demographic controls. Our conservative interpretation is that *Black Box* reliably reduces all-or-nothing adoption, consistent with our hypothesis, but that whether resulting partial adoption is normatively correct hinges on by correct perceptions of the human task difficulty.

E.2 Additional results

Adoption shares. Table 10 presents the main adoption results of Figure 6 using raw shares, confirming the effects discussed in the main text.

Mechanism. Table 4 in the main text compares projection proxies across arms at three belief-elicitation phases: Initial (pre-delegation), Interim (after 30 problems), and Final (after 60 problems). All outcomes use 0–100 belief reports and are scaled so that higher values indicate more projection. Let Δ_t^i represent the belief gap (Easy - Hard) for agent $i \in \{H, A\}$ at phase t. *Ability–Model projection*. "Single-index AI ability:" Spearman correlation between each subject's AI-beliefs for Easy and Hard pools. "Belief ordering:" share of subjects with $\Delta_t^A \geq 0$ (consistent with human difficulty). "Difficulty-gradient precision:" $1 - \text{IQR}(\Delta_t^A)/200$ (decreasing in subjects' dispersion in their implied difficulty gradient). *Difficulty projection*. "Mapping (rank):" cross-subject Spearman correlation between Δ_t^H and Δ_t^A. "Mapping (level):" $1 - \text{Med}(|\Delta_t^A - \Delta_t^H|)/200$ (decreasing in AI gaps' divergence from human gaps). *p-values:* Fisher z tests for correlation differences; robust two-sided test from $\mathbf{1}\{\Delta_t^A \geq 0\}$ on Black Box; Brown–Forsythe test for equal dispersion of Δ_t^A; and two-sided Wilcoxon rank-sum test on $|\Delta_t^A - \Delta_t^H|$.

Delegation behavior and realized AI signals. Table 11 shows that *Black Box* reduces delegation to the AI by 9.6pp overall ($p < 0.001$), with a larger decline on the easy pool (14.7pp) than on the hard pool (4.6pp). Conditional on delegating to the AI, realized AI failure rates are statistically similar across framings, both overall and by pool (Panel B). Thus, the framing primarily shifts the propensity to delegate rather than generating large differences in realized AI performance feedback conditional on delegation.

Table 10: Adoption Results – Full Sample

	Baseline Sample						**Replication Sample**					
	Anthrop.		*Black Box*		**Difference**		*Anthrop.*		*Black Box*		**Difference**	
Outcome	Share	s.e.	Share	s.e.	Share	s.e.	Share	s.e.	Share	s.e.	Share	s.e.
All-or-nothing	37.3	(3.9)	24.5	(3.4)	-12.8**	(5.2)	37.9	(6.4)	18.6	(5.1)	-19.3**	(8.1)
Full Adoption	35.3	(3.9)	19.5	(3.1)	-15.8***	(5.0)	34.5	(6.2)	15.3	(4.7)	-19.2**	(7.8)
No Adoption	2.0	(1.1)	5.0	(1.7)	3.0	(2.1)	3.4	(2.4)	3.4	(2.4)	-0.1	(3.4)
Only Hard	48.7	(4.1)	47.2	(4.0)	-1.5	(5.7)	55.2	(6.5)	72.9	(5.8)	17.7**	(8.7)
Only Easy	14.0	(2.8)	28.3	(3.6)	14.3***	(4.6)	6.9	(3.3)	8.5	(3.6)	1.6	(4.9)
Observations	150		159		—		58		59		—	

Notes: Each cell reports the share of subjects (in p.p.) choosing the corresponding final adoption pattern, with standard errors computed within each treatment arm and sample. Significance stars are based on a two-samples tests of proportions: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 11: Adoption Experiment – Delegation Behavior

	Treatment			
	Anthropomorphic	*Black Box*	**Diff.**	*p*
Panel A: Delegation to AI (%)				
AI delegation (overall)	61.2	51.6	**-9.6**	<0.001
AI delegation (easy pool)	47.2	32.5	**-14.7**	<0.001
AI delegation (hard pool)	75.2	70.7	**-4.6**	0.008
Easy–hard delegation gap	-28.0	-38.2	**-10.1**	<0.001
Panel B: AI failures among delegations (%)				
AI failure rate (overall)	34.1	34.6	**0.5**	0.322
AI failure rate (easy pool)	35.8	38.2	**2.4**	0.171
AI failure rate (hard pool)	33.7	34.4	**0.7**	0.236

Notes: The table reports mean outcomes by treatment arm and their difference. Panel A reports the share of tasks delegated to the AI (in p.p.), overall and separately for the easy vs. hard task pools; the "easy–hard delegation gap" is the easy-pool delegation rate minus the hard-pool delegation rate. Panel B reports realized AI failure rates (in %) conditional on delegating to the AI. *p*-values are from two-sided *t*-tests by treatment arm.

Table 12: Adoption Experiment – Balance Table

	Baseline Sample				**Replication Sample**			
	Anthropomorphic		*Black Box*		*Anthropomorphic*		*Black Box*	
	Value	s.e.	Value	s.e.	Value	s.e.	Value	s.e.
Demographics								
Age (mean)	41.4	(1.0)	37.8	(1.0)	39.1	(1.8)	37.7	(1.7)
Female (%)	52.7	(4.1)	49.1	(4.0)	48.3	(6.6)	54.2	(6.5)
White (%)	58.0	(4.0)	60.4	(3.9)	74.1	(5.7)	64.4	(6.2)
Black (%)	34.7	(3.9)	31.4	(3.7)	22.4	(5.5)	32.2	(6.1)
Socioeconomic and education								
Income (mean category)	2.6	(0.1)	2.5	(0.1)	2.2	(0.2)	2.5	(0.2)
HS or more (%)	100.0	(0.0)	99.4	(0.6)	100.0	(0.0)	98.3	(1.7)
Bachelor or more (%)	74.0	(3.6)	76.1	(3.4)	60.3	(6.4)	74.6	(5.7)
Graduate or more (%)	25.5	(3.6)	22.2	(3.3)	15.5	(4.8)	23.7	(5.5)
Math background								
HS math or more (%)	99.3	(0.7)	98.7	(0.9)	100.0	(0.0)	98.3	(1.7)
College math or more (%)	52.7	(4.1)	62.9	(3.8)	46.6	(6.5)	64.4	(6.2)
Graduate math or more (%)	4.0	(1.6)	5.0	(1.7)	3.4	(2.4)	3.4	(2.4)
AI familiarity								
Heard of AI/LLMs (%)	99.3	(0.7)	98.7	(0.9)	100.0	(0.0)	100.0	(0.0)
General understanding (%)	76.7	(3.5)	72.3	(3.5)	79.3	(5.3)	78.0	(5.4)
Interacted with AI (%)	34.7	(3.9)	30.8	(3.7)	39.7	(6.4)	45.8	(6.5)
Observations	150		159		58		59	

Notes: The table reports baseline covariate balance by treatment arm, separately for the Baseline sample and the Replication sample. "Value" is the cell mean; "s.e." is the standard error of the mean (for binary indicators, computed from the Bernoulli standard error). Shares are shown in percent; age is in years; income is the categorical index used in the survey.