Human Learning about AI

Bnaya Dreyfuss
Harvard University

Raphaël Raux
Harvard University

February 12, 2025

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.

Harvard University: bdreyfuss@g.harvard.edu & raphael_raux@g.harvard.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, Kai Barron, 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 help with the field experiment. 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: copyright © 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

One of the greatest strengths of AI models—particularly of Large Language Models (LLMs)—is the diversity of their potential uses. This versatility makes predicting their performance difficult for experts (Chang et al., 2024) and even more so for users, who have little understanding of the underlying technology. Inaccurate evaluation of AI performance has important economic consequences: Dell'Acqua et al. (2023) find that giving highly trained consultants access to GPT-4 results in a net productivity *loss* when they misjudge whether a task is outside AI's competence; Agarwal et al. (2023) show that providing radiologists with a high-quality AI predictions does not increase diagnostic accuracy, as physicians misweight AI signals.

Decisions to adopt AI for certain tasks depend on how its performance on these tasks compares to humans. Since task features that predict human performance do not necessarily predict AI performance, learning and predicting AI performance is far from trivial. For example, an AI may succeed in a task that is difficult for most humans, such as passing an MBA exam, while failing basic spelling tests almost no human would fail.[1] Similarly, two texts deemed similar by humans may not be seen as similar by AI, which can cause an AI to misunderstand a user's question in humanly unreasonable ways. Inferences from human-relevant features can therefore be misguided: an AI failing in humanly-unreasonable ways may still be worth adopting, as it can outperform humans at human-hard tasks. In other words, learning about AI may require people to *ignore* task features that are relevant in learning about humans.

This paper studies the hypothesis that people over rely on human-relevant task features when forming expectations about AI performance. Under this mental model of *Human Projection* (HP), people believe that task features that predict human performance also predict AI performance, potentially distorting expectations and subsequent adoption decisions. In lab experiments where ChatGPT's performance on math tasks is uncorrelated with human difficulty, we find that subjects still heavily rely on human difficulty when predicting AI performance resulting in misspecified beliefs. This misspecification persists and can lead to over- or under-adoption even after observing many signals of performance, and we find over-adoption decreases when experimentally manipulating AI to appear less "anthropomorphic." In a field experiment with an AI chatbot giving parenting advice, we find that potential users react to the human likeness of AI errors. Among equally uninformative answers, those that are less

[1] ChatGPT has passed a Wharton MBA Operations Management course (Terwiesch, 2023), while surprisingly failing to count the number of "r" letters in the word "strawberry" (see, e.g.: https://www.inc.com/kit-eaton/how-many-rs-in-strawberry-this-ai-cant-tell-you.html).

reasonable—less humanly-similar to the correct, expected answer—significantly reduce trust and continued engagement with AI. Consistent with the projection of human similarity onto AI, users reject a high-performing AI because its mistakes are unreasonable—i.e., highly dissimilar to a helpful answer—from a human perspective.

HP can result from a combination of two main factors. First, predictions of AI performance are highly complex *ex ante*, especially for users with limited experience; this can lead rational but constrained agents to fall back on cognitive defaults (Woodford, 2020), in this context, predictions made about other humans. Second, AI systems are often *designed* to mirror humans, making projection more likely. Our results imply that while AI "anthropomorphism" tends to induce trust (Chugunova and Sele, 2022), it can backfire by de-aligning human expectations away from actual AI capabilities.

We present the basic theoretical framework for HP in Section 2. We model a principal predicting an agent's performance in a domain of tasks, where performance is binary (success or failure) and perfectly observed. The principal is subject to Human Projection, which has two components. First, the principal perceives the agent's success rate as a function increasing in the agent's (unknown) ability and decreasing in the task's difficulty feature (known). This function also satisfies a Monotone Likelihood Ratio Property (MLRP): successes on harder tasks are more diagnostic of high ability, while failures on easier tasks are more diagnostic of low ability. This "Ability Model" is consistent with models used in Item Response Theory (Lord and Novick, 2008) to evaluate human performance on standardized tests.[2] The second component is the *projection* of a human-relevant task feature onto AI. In our lab context the principal projects human task difficulty: they believe what is difficult for humans is also difficult for AI.

Projection of human features is the source of belief misspecification. In the Ability Model, features affect inferences over agent ability, and ability enters into the success rate perceived by the principal. In the case of difficulty, an error—failing a task—is more diagnostic of low ability when the task failed was human-easier. While we focus on specific features projected onto AI, projection can extend to any informative feature, and to any type of agent. For example, basic grammar mistakes made by a non-native speaker might be wrongly seen as evidence of poor education when projecting the notion of "basic" seen from a native speaker's perspective.

Our framework delivers two testable predictions about beliefs when projecting human difficulty onto AI. The first states that prior beliefs about the agent's success are decreasing in the task's human difficulty. The second states that: (1) following a failure on a task, posterior beliefs *decrease less* if the human difficulty of the task is higher; (2) following a success, posterior beliefs *increase more* if task difficulty is higher. This result is driven by the MLRP: "easy mistakes" are more diagnostic of low ability, and "hard successes" of high ability. Resulting expectations of AI performance are inaccurate whenever human difficulty significantly differs from AI difficulty.

[2]The Ability Model implies a consistent performance ranking among agents across all tasks and among tasks for all agents, which may not hold if, e.g., different agents excel at different types of tasks. Whether there exists a consistent ranking remains a largely open question in the AI evaluation literature; see, e.g., Martinez-Plumed and Hernandez-Orallo (2018) in the context of video game benchmarks.

We empirically test these predictions using a domain of tasks sourced from standardized math tests. In Section 3, we describe our dataset of 414 multiple-choice problems sourced from the *Trends in International Mathematics and Science Study* (TIMSS). Following standard practice (Bachman, 1990), we define human task difficulty as the share of humans providing an incorrect answer on that task, using data from an incentivized test we administered to Prolific subjects. We then (zero-shot) prompt ChatGPT 3.5 with our set of tasks. ChatGPT performs better than the average human (82% correct vs. 67%) but under-performs on human-easy tasks. Overall its performance is uncorrelated with human difficulty (OLS coeff: -0.001, SE = 0.001).[3]

We test for projection of human difficulty onto AI in Section 4, using TIMSS problems as tasks. In a between-subject experiment, we incentivize participants to predict the performance of an agent: a randomly chosen human (*Human* treatment; $N = 222$) or ChatGPT (*AI* treatment; $N = 911$). We measure belief updating in three steps: we first elicit prior beliefs on a random initial task; we then reveal agent performance on a different task, varying both performance and (human) difficulty of the task revealed; we finally elicit posteriors on the initial task.

Results are consistent with our two predictions. Prior beliefs in performance strongly decrease with human task difficulty, both in *Human* (OLS coeff: -0.65; SE = 0.01) and *AI* (-0.32; 0.01). Then, subjects update more negatively following a failure on an human-easier task, and more positively following a success on a human-harder task. Expectations thus depart from belief benchmarks that ignore human task difficulty, and are inaccurate in this context because AI performance is not correlated with human difficulty.

We then turn to study the consequences of HP for equilibrium adoption in Section 5. We build a "medium-run" adoption model, where the principal (e.g., a firm) engages in fixed difficulty projection but receives many signals of performance. These signals are endogenous to their decisions to delegate each of two production tasks—one human-easy, one human-difficult—to a human or to an AI. We characterize the Berk-Nash equilibrium (Esponda and Pouzo, 2016) under HP. We show that partial adoption—using AI only for one of the two tasks—is never a Berk-Nash equilibrium, even when it maximizes profits. Upon observing AI perform better than humans in one task, the principal infers that AI has higher ability, and must therefore perform better in the other task as well. Human Projection thus leads to an "all-or-nothing" adoption decision, where absolute advantage in one task must imply—through higher inferred ability—absolute advantage in the other. We then assume that a less "anthropomorphic" appearance of AI leads to a lower degree of Human Projection (consistent with Chugunova and Sele, 2022) and obtain a testable prediction: a non-anthropomorphic AI decreases the share of all-or-nothing (Full or No Adoption) decisions in equilibrium.

We test this prediction with a repeated-learning design using the same pool of math tasks. We construct two sets of tasks: one human-easy, where humans dominate (presented as "blue"

[3]This evidence is not meant as an accurate description of the frontier of LLM performance—which evolves quickly—but rather as a symptom of a deeper misalignment between AI performance and human difficulty, which remains relevant at the SOTA at the time of writing this paper (Mialon et al., 2023; Xie et al., 2024). Our argument is that people rely on human difficulty when forming beliefs about AI performance, *even when the two are not correlated*.

tasks; success rate of 78%, vs. 66% for AI), and one human-hard where AI dominates ("green" tasks; 23% vs. 66%). We choose relative success rates to maximize the baseline likelihood of over-adoption (i.e., Full Adoption). Consistently with ChatGPT's overall math performance, it is optimal to adopt AI only on human-hard tasks. Participants make a series of 60 delegation decisions (30 from each set) all following the same process. They see a task along with its type (green or blue) and delegate it to either a human or to AI. They then learn the performance (success or failure) of their pick, and receive a small bonus for each success. At the end, participants make a final decision to adopt either humans or AI for 10 random tasks of each type. Our treatment manipulates the degree of Human Projection. In the baseline condition, *Anthropomorphic* ($n = 59$), we use a human-like AI framing similar to those used by ChatGPT or Claude. We endow AI with a name, a "typewriter" effect when providing answers, and use the active voice to describe its behavior. In the *Black box* condition ($n = 58$) we remove all human references: we present AI as a neutral "black box" and use the passive voice to describe its behavior. This manipulation decreases the degree of projection of human difficulty, thereby making participants more agnostic regarding expected patterns of AI performance.

Adoption decisions are consistent with our prediction: the share of Full Adoption (over-adoption) is significantly lower in *Black box*, compared to *Anthropomorphic* (15% vs. 34%, p-value = 0.016). Subjects in *Black box* are significantly more likely to take the optimal adoption decision, i.e., to choose AI for human-hard tasks, and humans for human-easy tasks (55% vs 73%, p-value = 0.046). Results suggest Human Projection can lead to suboptimal adoption decisions even after observing a relatively large number of signals. Removing AI's anthropomorphic features lowers projection and makes participants better able to realize that AI's success rate can be uncorrelated with human difficulty, reducing potential AI misuse.

We move to the field in Section 6 and complement our lab findings with evidence on real user engagement with AI. We use "Dewey," an AI chatbot specialized in parenting questions hosted on the website ParentData.org. Dewey answers questions and provides human-vetted advice on topics related to conception, pregnancy, or child-rearing. Dewey sometimes misunderstands the user's prompt and provides an answer to the wrong question, useless given what the user asked. We study inferences from such errors when users project human *textual similarity* onto AI, i.e., when they believe two pieces of text which appear similar to humans also do to AI.

To obtain a prediction on user inferences under HP, we first define an answer's (human) *reasonableness* as its human similarity to a correct (i.e., useful) answer. As a task feature, reasonableness (i.e., similarity) plays the same role that task difficulty played in the math context: it enters the principal's subjective success rate in the same way and we measure it consistently with how we measure difficulty.[4] A similar belief updating prediction therefore follows when human similarity is projected onto AI. Among AI errors (useless answers), those that are less reasonable from a human perspective—i.e., less humanly-similar to useful answers—further reduce beliefs in performance and willingness to engage with AI. Relying on human similarity to assess AI's

[4]We measure task difficulty based on performance, as the "share of humans who would fail the task." We measure reasonableness directly, as the "share of reasonable humans who would misunderstand in the way the AI did."

quality can lead to inaccurate expectations when notions of similarity differ between humans and AI, which is the case here. When projecting human similarity, humanly-unreasonable AI errors are seen as highly diagnostic of low AI quality, although they might not be.

We test the above prediction in a field experiment where we show participants AI errors from real human-AI conversations (user question and AI answer). Using current or expecting parents as subjects, we manipulate the human reasonableness of AI errors they observe and measure subsequent engagement. To build intuition for our design, consider two wrong answers to the question: "Which car seat brand should I buy?" The first answer gives advice on where to install a baby car seat (front or back seat). The second answer discusses which is the best baby food brand. While neither contains the desired information, the former is more humanly-reasonable than the latter, because it at least shares the same context and thus is more humanly-similar to a useful answer (e.g., one discussing which car seat is the best). We predict under Human Projection that engagement is lower after observing the latter answer.

As in the above example, we construct pairs of conversations from real user-Dewey interactions for which: (i) both queries in the pair ask the same question, only worded differently; (ii) both AI answers are failures, rated (by parents) as equally useless; (iii) one answer is rated as significantly more reasonable than the other. Our design then manipulates between subjects the side of the pair—i.e., the type of error—displayed: *Reasonable* ($n = 451$) or *Unreasonable* ($n = 454$). After each conversation, we elicit beliefs in performance and trust in the chatbot. After the last conversation, we measure willingness to keep engaging with AI: participants choose to receive either a link to the chatbot or to parenting articles. We also observe subsequent engagement with Dewey using participants' IP addresses.

We find significant effects of human reasonableness on beliefs, trust, willingness to engage, and actual engagement. Errors lead to decreases in beliefs and trust, which are significantly more pronounced when they are less humanly reasonable (80% larger drop for beliefs, and 75% for trust). As a result, participants in *Unreasonable* are significantly less likely to choose the chatbot link (39% vs. 49%, $p = 0.005$), and to use Dewey at all post-experiment (1.3% vs. 3.3%, $p = 0.045$). In other words, conditional on receiving a useless answer, subjects' inferences strongly depend on the human similarity of the answer to a correct (useful) response. These inferences overlook the fact that unreasonable answers are not as diagnostic of a low AI ability as they are for a human: Dewey is overall highly accurate (a large majority of its answers are rated as highly useful), despite some of its mistakes being un-human-like.

Taken together, our findings indicate that people engage in Human Projection when forming expectations about AI performance, which affects AI usage and adoption. Our work has three practical implications. First, HP can distort adoption timing and utilization: in Appendix A we consider a dynamic setting where technology improves over time and show HP delays adoption compared to a rational benchmark. Once adoption occurs however, AI gets over-adopted, even for tasks where humans outperform AI. Second, AI anthropomorphism may backfire by de-aligning human expectations and AI capabilities: making AIs appear human-like increases human trust (Chugunova and Sele, 2022), but also leads users to over-react to its (sometimes

un-human-like) performance. Third, human-AI interactions can be designed to improve the accuracy of human expectations. Current LLM training may be improved to include human features that are relevant for people's assessment of AI performance. On the other side, users can be trained to reduce their degree of projection. We further discuss these issues in conclusion.

Related Literature. Our work relates to several strands of literature. First, we add to the literature on the use of "mental models" in economic decision-making. (Johnson-Laird, 1983; Hanna et al., 2014; Bordalo et al., 2016; Enke, 2020; Imas et al., 2022; Andre et al., 2023a,b; Graeber, 2023; Barron and Fries, 2023; Esponda et al., 2024; Kendall and Oprea, 2024) We characterize a general mental model used in performance evaluation, and apply it to AI technology. We also add to the—mostly theoretical—economic literature on learning under (potentially misspecified) mental models (Esponda and Pouzo, 2016; Gagnon-Bartsch et al., 2018; Fudenberg and Lanzani, 2023; Lanzani, 2022; Esponda et al., 2024), where we add one of the first pieces of empirical evidence on Berk-Nash equilibrium decisions.

We then document a new type of projection in the behavioral economics literature on "projection bias" (Loewenstein et al., 2003; Grable et al., 2004; Gruber, 2009; Acland and Levy, 2015; Gagnon-Bartsch and Bushong, 2022; Gagnon-Bartsch and Rosato, 2022; Bushong and Gagnon-Bartsch, 2024). While earlier work has focused on the projection of subjective preferences onto one's future self or other people, we show projection can occur with variables deemed relatively "objective," such as task difficulty or answer similarity. Rather than a simple bias, we argue projection can arise from users' limited information when predicting AI performance.

We also find that the degree of projection is influenced by AI "anthropomorphism," which connects our findings to the literature on Human-Machine Interaction (De Visser et al., 2016; Kulms and Kopp, 2019; Natarajan and Gombolay, 2020; Chugunova and Sele, 2022). We highlight a potential drawback of anthropomorphic traits in machines: while they have been shown to increase user trust and pro-social behavior, they can trigger reliance on mental models such as HP, leading to distorted beliefs and adoption decisions. Such distortions can lead to both overly optimistic and overly pessimistic beliefs in performance, which could help explain conflicting evidence on algorithm appreciation (Logg et al., 2019; You et al., 2022) and aversion (Dietvorst et al., 2015; Dietvorst and Bharti, 2020).

Our findings speak to a growing literature studying AI adoption and task delegation (Green and Chen, 2019; Lai et al., 2022; Noy and Zhang, 2023; Brynjolfsson et al., 2023; Agarwal et al., 2023; Dell'Acqua et al., 2023; Bick et al., 2024). Namely, we provide a behavioral mechanism explaining the difficulty of delegating tasks to AI. While people tend to delegate tasks which they think AI will perform well (Wang et al., 2021; Lai et al., 2022), misspecification induced by Human Projection can lead to delegation errors and AI "misuse." Our work is complementary with Vafa et al. (2024), which also documents patterns of human inferences from LLM performance. They find human inferences only arise across some domains of tasks (e.g., from math to physics, but not from math to literature). We provide an intuitive structure—the Ability Model—to help explain their findings: the sparsity of inferences is consistent with math and physics abilities be-

ing seen as more correlated than math and literature abilities. They find that human predictions of LLM performance can be inaccurate, especially for larger AI models. We report a similar finding in the math domain, and overall their evidence complements and broadens the scope for HP misspecification we document in this paper.

Finally, our work speaks to the literature looking to align AI systems with human intentions and values (Gabriel, 2020; Terry et al., 2023; Wang et al., 2024). We highlight the role of performance alignment as a type of robustness (Ji et al., 2023) of AI models, and our results support model human-centered evaluation precedures (Shankar et al., 2024; Wallach et al., 2025).

2 Theoretical Framework

In this section we develop the framework for Human Projection (HP), which is composed of two components: (i) a model of performance, the "Ability Model," involving task features; (ii) projection of the relevant (human) feature onto AI. The Ability Model is a general model of performance evaluation, which can be used to assess humans or technology. The feature projection, source of misspecification, may vary across contexts. In this section we consider the projection of (human) task difficulty, relevant for performance on mathematical problems.

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 $\theta \in \Theta \subseteq \mathbb{R}$, and is 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. We denote the difficulty of problem t_k by $\delta^i(t_k)$, and a problem with a level of difficulty δ by t^δ. We let this difficulty be agent-specific (i.e., a human-difficult task may not be AI-difficult) and "difficulty" will refer to *human* difficulty unless specified otherwise. The probability that an agent with ability θ succeeds in solving a given task t^δ is given by $p : \Theta \times \Delta \to [0,1]$. Assumption 1 imposes structure on the mapping from ability and difficulty to success rates:

Assumption 1 (Ability Model)**.** *The function $p(\theta, \delta)$ satisfies the following:*

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

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 from principles of 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.[5]

The principal is trying to learn about agent ability θ^i, where agent $i \in \{H, A\}$ can be a human or an AI. He receives binary signals of agent performance on tasks—success or failures—within domain $\mathcal{T}$, and updates his prior G^i according to Bayes rule.

We now introduce the "projection," source of belief misspecification about AI. We assume the principal perfectly observes human difficulty δ^H but not AI difficulty δ^A:

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): a tendency to project human features when forming expectations about AI performance.

Assumption 1—the Ability Model— 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 principal faces large cognitive constraints—typically, when the number of tasks within the domain is large—this assumption is relatively weak. 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. Assumption 2—feature projection—is

[5] 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.

stronger in comparison. It posits that task difficulty for AI is assessed (partially or totally) from a human perspective.

2.2 Predictions

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

Prediction 1. *The predicted success rate is decreasing in δ^H for both humans and AI. Formally, $\forall i \in \{H, A\}$:*

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

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

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

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

Proposition 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—driven by the MLRP—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. Formulated in terms of belief movement (difference between posterior and prior beliefs), Proposition 2 states that: (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. We test this version of the prediction in Section 4.1.

3 Constructing a Domain of Tasks: Mathematics

In this section, we describe how we construct the domain of tasks assumed in our framework. We focus on standardized math problems, for which we can obtain clean measures of (binary) performance and task difficulty. We use this domain in both the beliefs experiment (Section 4)

and adoption experiment (Section 5). We first present the task dataset, then our measures of human difficulty and of AI performance. Appendix B contains further details regarding task dataset and performance data. We will hereafter use the terms "task," "item," and "question" interchangeably.

3.1 Task Dataset

We briefly describe the construction of our dataset, and relegate details and problem examples to Appendix B. 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. To obtain a consistent measure of human difficulty, we focus on multiple-choice items, which have either 4 or 5 possible answers, denoted by A, B, C, D, or E.[6] 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 range of usual mathematical topics.

Our dataset presents a number of advantages as a domain of tasks. First, performance is binary, since each item has a unique and objectively correct answer. Second, the format of questions is kept constant as the item writing guidelines are similar across tests. Third, tasks span a broad range of topics and difficulty, and are purposefully designed to be independent of national or cultural contexts. Fourth, a significant part of the items are not accessible online and were obtained through direct request to the IEA: this means they could not be part of ChatGPT's training set, which could invalidate our measure of AI performance. Fifth, problems are designed to assess math ability, which is estimated using models which satisfy restrictions of the Ability Model (Assumption 1).

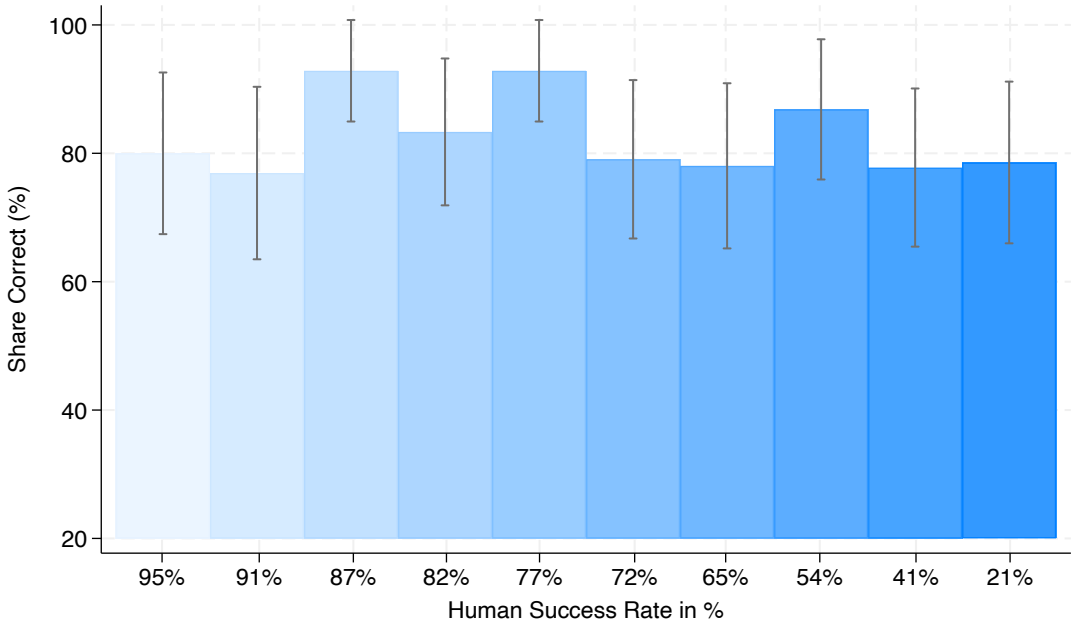
3.2 Measuring Human Difficulty and AI Performance

We construct our own human performance measure using adult participants recruited online on Prolific. The data comes from two samples: (i) an initial mathematics test we administered during the month of October 2023; and (ii) performance data we collected during the beliefs experiment. Test conditions, described below, are highly similar across samples so we pool samples together. Appendix C.2 presents test instructions as well as screenshots of decision screens.

Test Conditions. Initial test conditions aim to recreate an actual examination. We recruit adult participants on Prolific to take a test composed of 30 randomly sampled questions, 10 for each grade level. Subjects are instructed to "approach this as [they] would a real-life math test," and to put a reasonable amount of effort into each question while remaining time-conscious. We limit use of external assistance through several features: (i) all problems are screenshots, preventing copy-pasting into e.g., Google or ChatGPT; (ii) monetary incentives are kept modest—5 cents per

[6] We set aside open-ended problems, whose difficulty is hard to compare to that of M.C.Q.. We also exclude items with significant visual components, such as drawing or reading pictures or charts, as data collection was conducted in August 2023, before updates allowing ChatGPT to process images.

Figure 1: ChatGPT Performance by Human Difficulty Deciles



Notes: The figure plots ChatGPT's share of correct answers among each decile of human difficulty. Each decile contains between 40 and 42 items, and each item was zero-shot prompted once in conditions described in this section. Numbers under each bar represent the average human success rate for each decile. The average number of human answers per item is 41.5.

correct answer—and complemented with feedback on performance relative to other participants; (iii) we elicit self-reported reliance on external assistance after the test, and exclude those (2% of total subjects) from the sample. Participants in the beliefs experiment face identical conditions, but see only 11 random problems (3 of High School, 8 of other grade levels).

Difficulty Measure. Following the standardized test literature (Bachman, 1990), we define *human* task difficulty based on average subject performance: for task t, $\text{Difficulty}_t = \text{Share Incorrect}_t$. We exclude test-takers who: (i) reported using external help; (ii) fail the comprehension question on test instructions; (iii) fail attention checks. The average number of human answers per item is 41.5.[7] All results presented in Section 4.2 are robust to alternative measures of difficulty which do not exclude low-quality test-takers. Hereafter the mention of "difficulty" will always refer to this measure of human difficulty. Descriptive results for task difficulty are: 30.3 average, 24 median, 0 lowest and 92 highest; see Appendix B for more details.

AI Performance. We collect AI performance data by zero-shot prompting ChatGPT (3.5, August 2023 version) with all items from our dataset. Each item was given as a separate prompt, including only the stem (the question itself) and possible answers. We performed minor formatting changes to help process special symbols and tables, after some experimentation to ensure

[7] High-School problems have more answers per question, while 8$^{\text{th}}$-grade problems have fewer. The lowest number of answers is 24 and highest is 164, median of 34. The variation comes from two sources: (i) the number of released problems is not split evenly by grade level, as shown in Figure 13 in Appendix B; (ii) the beliefs experiment samples more heavily from HS level, to ensure each subject sees the whole gradient of difficulty.

ChatGPT was able to correctly process inputs. We classified an answer as correct if the AI designated the appropriate answer key as its answer, and as incorrect otherwise. The soundness of "reasoning" leading up to the final answer was ignored in order to most closely mirror test conditions faced by human subjects.

Figure 1 plots ChatGPT's average performance across the deciles of difficulty. For reference, human success rates in the relevant decile are reported below each bar. Overall, ChatGPT was correct on over 82% of the tasks; human difficulty does not correlate with AI performance, and displays very small predictive power. (OLS coefficient: -0.001, SE = 0.001, $R^2 = 0.002$). Appendix B displays a similar graph for human performance. Unlike any of our human test-takers, ChatGPT consistently fails around 20% of tasks across the difficulty gradient.

This evidence (which comes from an older model) is not meant as a general claim about the correlation between AI performance and human difficulty. It is rather symptomatic of a deeper misalignment between human and AI difficulty, which has been known for a long time as Moravec's paradox, and maintains its relevance today with the latest models (Mialon et al., 2023). Xie et al. (2024) namely remark: "We identify notable disparities in the perceived difficulty of tasks between humans and AI agents. Tasks that are intuitively simple for humans often present substantial challenges to agents, and conversely, tasks that humans find demanding can be more straightforward for agents to execute." Our argument (supported by the evidence in the next section) is that people rely on human difficulty in order to form beliefs about AI performance, *even when it is uncorrelated with human difficulty*. This leads to misspecification is our case, and in all other cases where AI's patterns of performance are un-human-like.

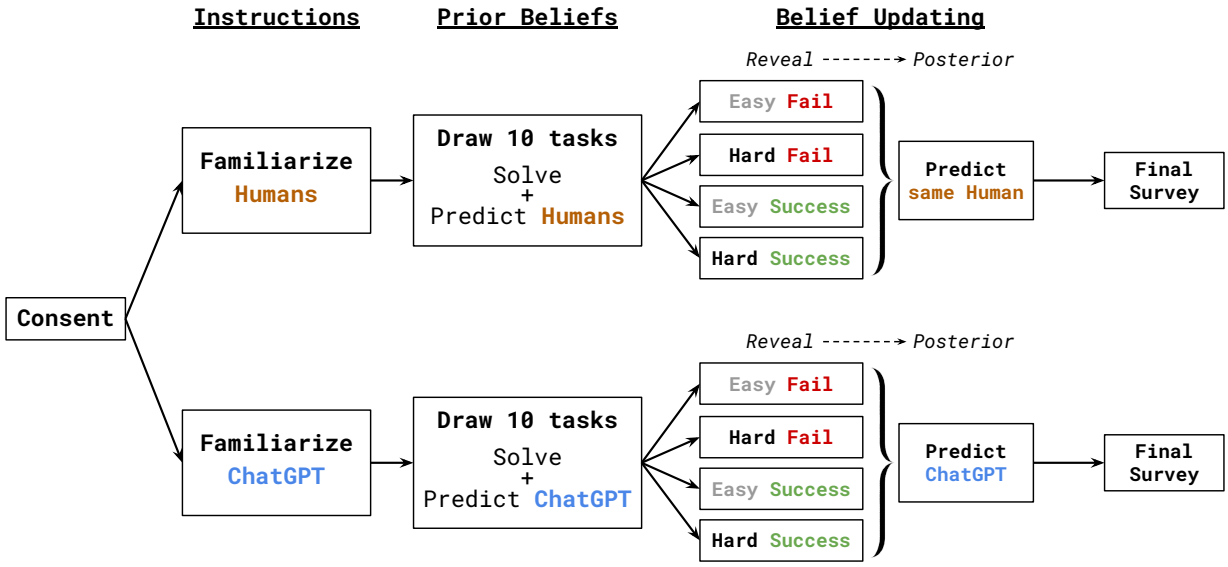
4 Documenting Human Projection: Beliefs in Performance

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

We ask participants to predict an agent's performance over tasks of varying human difficulty. We randomly vary between subjects the identity of the agent whose performance they are to predict: a randomly-selected participant who attempted the question in the *Human* treatment, and ChatGPT in the *AI* treatment. We test that *both* beliefs about human and AI performance follow patterns of Predictions 1 and 2 for human task difficulty. For humans this test is a sanity check, while for AI it is a test of Human Projection: whether beliefs about performance are influenced by the human difficulty of problems. To reduce concerns about experimenter demand effects, treatments are completely blind to each other: AI is never mentioned to subjects in *Human*, while human test-takers are never mentioned to subjects in *AI*. We also never refer to

Figure 2: Flowchart of Experimental Design



Notes: Subjects are randomly assigned to the *Human* or *AI* treatment right after indicating consent. Tasks for prior beliefs are drawn at the start of the survey block. Subjects are then assigned to one of the four belief updating conditions after completing the prior beliefs part. The final survey collects socio-demographic variables and open-ended comments.

tasks' (human) difficulty explicitly.

We show the structure of the experiment on Figure 2: we familiarize participants with the agent during instructions, then elicit prior beliefs and belief updating. We describe each of these steps in more detail below, and include screenshots of decision screens in Appendix C.

We introduce the agent to participants to increase realism and reduce noise in elicited beliefs. In *AI*, we provide a general definition of AI and LLMs, then provide animated GIFs showing ChatGPT answering common prompts.[8] To avoid prior contamination, these examples are not related to the math problems and we make no comment regarding the quality of answers. In *Human*, we describe the human test-taker sample, the test conditions they faced, and provide basic demographic information.

Prior Beliefs. Each participant randomly draws 10 problems, presented on the same screen. To ensure that each subject sees questions spanning the full range of difficulty, 2 are drawn from the High-School pool, and 8 from the 4^{th}- and 8^{th}-grade pool. For each problem, subjects are asked to report their answer to the problem, and their probabilistic belief about the agent's performance (both incentivized). We elicit the latter using the following language: "What do you think is the % chance [a random participant/ChatGPT] answered correctly?" This is the only difference between *Human* and *AI* treatments. Asking subjects to solve problems before

[8]We show a total of 3 prompts. The first two are taken directly from a recent news article and ask ChatGPT to write a cover letter and to explain the physics concept of wormholes. We add a third prompt about a famous logical puzzle—the "twelve-coin problem"—to show that ChatGPT can also attempt to solve problems, beyond providing descriptive content. For all of them, we give subjects the option to freeze the GIF and examine the answer.

stating their beliefs gives them a signal of task difficulty and generates additional performance data which we add to our difficulty measure.

Belief updating. Within each treatment, subjects are randomly assigned to one of four possible conditions which vary the signal of agent performance: Easy Fail, Hard Fail, Easy Success, and Hard Success. We proceed in three steps to measure belief movement, defined as the difference between posterior and prior beliefs. First, subjects draw a "prediction" question from a pool of 10 of the hardest tasks (on the basis of initial test data). They are incentivized to solve it and provide their prior beliefs, as in the previous part. Second, we reveal agent performance on a different task—the "signal" task. This task can be either easy (lowest difficulty decile) or hard (highest decile), and performance on the task can be a success or a failure. Signal tasks are randomly sampled from four pools of 10 questions each.[9] They are presented along with the agent's answer and an indication of whether this answer is correct or incorrect.[10] Third, the prediction question is presented again and we elicit posterior beliefs in performance: "Given what you saw, what do you think is the % chance that [the same participant/ChatGPT] answered this question correctly?"

Incentives. Performance on problems is rewarded with 5 cents for each correct answer: we keep stakes low on purpose to be consistent with our initial test and to deter cheating as discussed in Section 3. Following the belief elicitation literature (Hossain and Okui, 2013; Erkal et al., 2020), we incentivize predictions using a binarized scoring rule.[11] Subjects can earn 10 cents per prior belief elicitation, and 30 cents for the belief updating prediction. While we keep these stakes small for power purposes, recent experimental work using very large incentives finds that stake size has little effect on belief updating errors (Enke et al., 2023).

Logistics. We programmed both the initial test and experiment using Qualtrics, and recruited participants on the Prolific platform during the months of October (initial test) and December 2023 (experiment). The test's pre-registration specifies test conditions, while the experiment's includes sample sizes, main hypotheses, and quality checks. We initially recruited a total of 244 subjects in *Human*, and 971 in *AI*. After dropping subjects failing attention checks and the relevant comprehension questions, the final sample sizes for *Human* and *AI* are respectively of 222 and 911 for the prior beliefs part, and of 231 and 809 for belief updating.[12] The initial test was designed to be completed in around 30 minutes with a base pay of $4 with a maximum

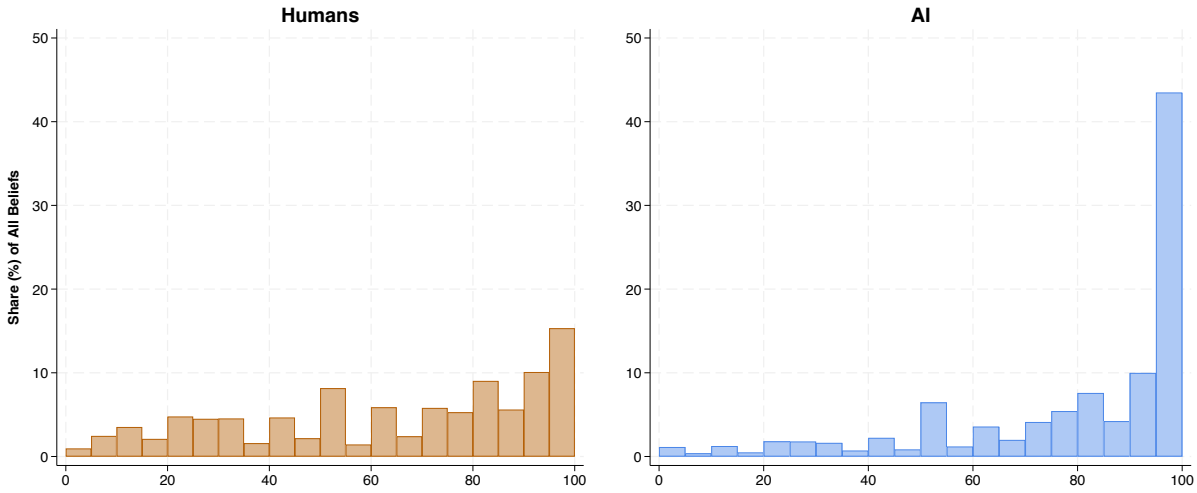
[9] Questions in this part are thus drawn from five disjoint sets of 10 questions: one pool of (hard) prediction questions, and four pools of signal questions. We chose hard prediction questions to avoid mechanical ceiling effects, as prior beliefs in AI performance on easier tasks are already close to 100%.

[10] In *AI*, we also include a screenshot of the prompt and ChatGPT's output, meant to increase the signal's credibility, as some of its mistakes may be difficult to believe ex ante. This does not lead to an increase in time spent on the page (median time of 35s in *AI* vs. 37s in *Human*), suggesting the signal does not significantly differ from a binary one.

[11] This scheme is complex, and can lead subjects to (wrongly) engage in strategic behavior (Danz et al., 2020). We thus describe it in intuitive terms, and provide details behind a clickable button.

[12] We include one comprehension question for each of the main tasks, and we drop (and replace) data from subjects who fail the corresponding question, leading to different sample sizes between priors and belief updating. For *AI*, as subjects failed the relevant comprehension question in higher proportions in failure conditions, additional participants were recruited to compensate.

Figure 3: 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.

potential bonus of $30 \times 0.05 = \$1.5$. The experiment was designed to last around 15 minutes, with a base pay of \$2 and a maximum potential bonus of $10 \times (0.05 + 0.1) + 0.3 = \$1.8$. We also gave participants feedback on their test performance relative to other subjects. As stated in the IRB submission and instructions, the experiment did not rely on deception, as we use real human or AI answers as signals of performance and to incentivize decisions.

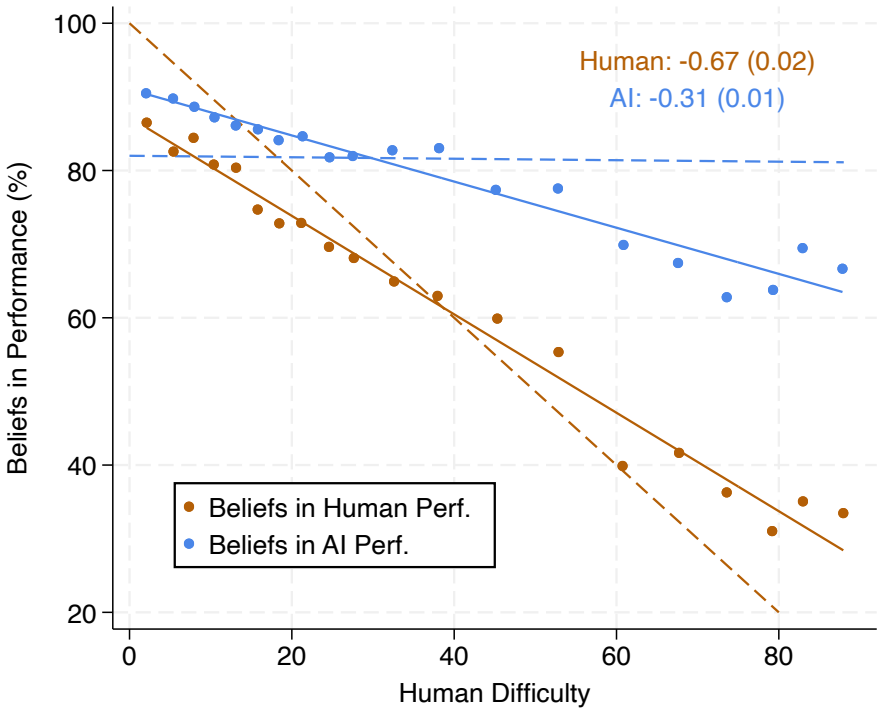
4.2 Results

Benchmark. Predictions 1 and 2 state that patterns of beliefs depend on *human* task difficulty. We expect this to be the case for beliefs about humans, since difficulty is defined from performance. Regarding beliefs about AI—here ChatGPT—we consider a benchmark where people are aware that it does not "reason" like humans do, and is more likely to answer correctly tasks that are more similar to those included within ChatGPT's training set, mostly composed of information available online.[13] Under this benchmark, perceived likelihood of AI success on a task should not depend on its human difficulty. This implies a "flat" prior on AI performance, whose intercept depends on prior over AI ability, and updating patterns which only depend on performance (success or failure).

Prior beliefs. We first plot the distribution of prior beliefs, presented in Figure 3. A significant share of beliefs about AI are exactly 100% (43% in *AI* vs. 14% in *Human*). This does not seem to be driven by subjects who would systematically report 100% on all tasks, as their proportion is small (3% in *AI* and 0.5% in *Human*). Overall, subjects place higher expectations on ChatGPT

[13] OpenAI's FAQ reads: "ChatGPT and our other services are developed using (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or human trainers provide."

Figure 4: Binned Scatter Plots of Prior Beliefs in Performance



Notes: The figure 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. Slopes for actual human and AI performance are represented by the dashed lines. We report coefficients from the basic OLS regression with standard errors clustered at the subject level.

than on a random human, with a mean belief of 64% in *Human* compared to 80% in *AI*.

We then turn to Prediction 1: Figure 4 presents binned scatter plots showing strong negative relationships between both priors in performance and human difficulty, with the slope being around twice as large in *Human* than in *AI*. Overall, ChatGPT is perceived to be performing strictly better than the average human across the whole difficulty gradient: predicted performance on the easiest questions is around 90% (85% for humans) compared to 65% (30% for humans) on the most difficult ones. Prior beliefs about AI thus depart from our flat benchmark which ignores human task difficulty. Given the actual patterns of performance, represented by the dashed lines, beliefs about AI are misspecified.

Beliefs about humans are consistent with the "hard-easy" effect (Lichtenstein and Fischhoff, 1977), underestimating performance on easiest tasks and overestimating it on hardest tasks. This pattern is reversed for AI: subjects overestimate AI performance on the easier tasks and underestimate it on the harder ones.

Table 1 presents regression evidence on priors, where we estimate the following specification:

$$Y_{it} = a_0 + a_1 \text{Difficulty}_t + d_i + \mathcal{X}_i + \epsilon_{it}, \tag{1}$$

Table 1: Correlations of Beliefs in Performance with Item Difficulty

	Dep. Var: Beliefs in Performance			
	Humans		**AI**	
	(1)	(2)	(3)	(4)
Difficulty	-0.665***	-0.664***	-0.316***	-0.319***
	(0.021)	(0.022)	(0.011)	(0.011)
Controls	Yes	Yes	Yes	Yes
Subject FE	No	Yes	No	Yes
R^2	0.389	0.602	0.125	0.440
Observations	2442	2442	10021	10021

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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

where Y_{it} are beliefs in agent performance (ranging from 0 to 100) of subject i on task t. a_1 is the main coefficient, associated with our measure of difficulty, and we include subject fixed effects (d_i) and socio-demographic controls ($\mathcal{X}_i$). Beliefs about performance are strongly negatively correlated with question difficulty for both treatments, and the coefficient for humans is about twice as large in magnitude. Task difficulty alone explains around 12% of the variation in beliefs for AI, and 39% for humans. Overall, these results are consistent with Prediction 1. The feature of human task difficulty is perceived by subjects as relevant for both human and AI success rates, which we interpret as evidence for Human Projection.

Belief Updating. We turn to Prediction 2, which compares the effect of different signals of performance on posterior beliefs. To estimate the signal effect we primarily look at the subject's *belief movement*, defined as $\text{Movement}_p \equiv \text{Posterior}_p - \text{Prior}_p$ for the same question p. We expect this movement to be negative in case of a failure, and positive in case of a success.

Figure 5 plots all belief movements across the four conditions, along with differences in means and p-values from the relevant t-tests. We observe, consistent with our prediction: (i) a larger increase following a Hard Success compared to an Easy Success; (ii) a smaller decrease following a Hard Fail compared to an Easy Fail. For both *Human* and *AI*, differences are less marked between failures than between successes, an effect which is consistent with discrepancies in average difficulty between task pools. Indeed, while the average difficulty is comparable for the Easy Fail and Easy Success pools (9.1 vs. 7.1), task difficulty in Hard Fail is significantly lower than in Hard Success (77.0 vs. 87.3). The difference in signal difficulty is thus smaller between failures ($77 - 9.1 = 67.9$) than between successes ($87.3 - 7.1 = 80.2$) which, under our

framework, leads to smaller effects for failures.[14]

We further test the prediction by regressing belief movement on our continuous measure of difficulty; we run the following specification separately for successes and failures:

$$Y_{ipst} = b_0 + b_1 \text{Difficulty}_t + \gamma_p + d_i + \mathcal{X}_i + \epsilon_{ips}, \tag{2}$$

where Y_{ipst} is belief movement of subject i on prediction question p, binary signal $s \in \{0, 1\}$ (failure or success), and t the signal question. The main coefficient is b_1, which is predicted to be positive in all cases. After failures belief movement is negative, but less so if the signal question t is more difficult. After successes movement is positive, and stronger if the signal question is more difficult. We include task and subject fixed effects, and individual controls.

Table 2 reports results: coefficients are smaller for failures than for successes, again for both humans and AI. Beyond discrepancies in task pools, results thus point towards a more general asymmetry in the effect of task difficulty between successes and failures.[15] Including prior beliefs as additional control delivers qualitatively similar results, presented in Appendix C.

Additional results. We investigate whether the smaller slope on prior beliefs observed in *AI* (Table 1) is due to partial projection of difficulty ($0 < \lambda < 1$), or to different priors over human and AI ability ($G^H \neq G^A$).[16] The latter would arise if subjects are more uncertain about AI capabilities or if they think, as our results suggest, that AI is better at math than humans in general. We thus proxy for (mean) ability priors using subjects' *average* reported belief, holding fixed average question difficulty. Comparing the most "optimistic" subjects in *Human* (reporting high average beliefs in performance) to all subjects in *AI* should make (mean) priors over ability more comparable across treatments. Figure 18 in Appendix C replicates Figure 4, but restricting the *Human* sample to the top 10% most optimistic subjects on the basis of average reported beliefs. The belief curves are very similar, both in slopes and levels (-0.35 (0.03) with OLS constant of 90 for *Human* vs. -0.32 (0.01) and constant of 86 for *AI*). This suggests that the belief gap is partially driven by differences in priors over ability, consistent with the view that people believe machines to be "better at math" than humans in general.

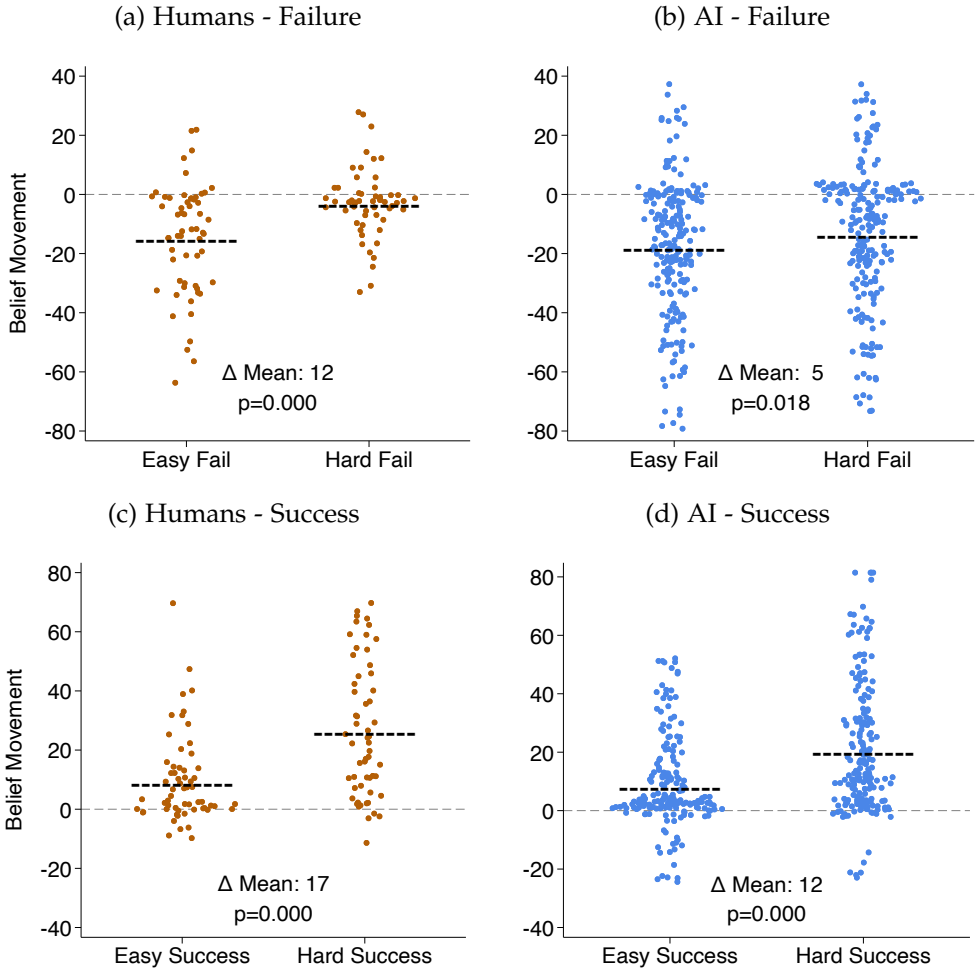
Finally, in Appendix C we quantitatively assess the accuracy of subjects' beliefs by regressing actual performance on elicited beliefs. Beliefs about both humans and AI are correlated with

[14] This discrepancy is due to using pools of signal tasks on the basis of performance data coming from the initial test (see Section 4.1). The beliefs experiment generated more data to refine the difficulty measure, leading to some changes in task difficulty.

[15] Two lines of interpretation for the asymmetry appear plausible in our context. The first points to a smaller "diagnosticity gap" (for ability) between hard and easy failures, compared to the gap between hard and easy successes. The second would assume that people also update "task to task" directly, in addition to updating over ability, and that similar tasks (difficulty-wise) lead to stronger direct updating. Under this lens, a Hard Success would have a stronger influence than an Easy Success both because the Hard Success is more predictive of high ability *and* because the signal task is more similar to the prediction task (both being hard tasks). An *Easy Failure* would have a stronger effect than a *Hard Failure* because it is more predictive of a low ability, but counteracted by the fact that the signal task (easy) is *less similar* to the prediction task (hard).

[16] We do not elicit the subjective mapping between difficulty and performance but simply a probabilistic estimate, which given our incentive scheme should be the expected value under the distribution of the ability prior G^i. Then, $G^H \neq G^A \implies \mathbb{E}_{G^A}[p(\theta^A, \delta)] \neq \mathbb{E}_{G^H}[p(\theta^H, \delta)]$ even in cases where mappings are the same ($p^A = p^H$).

Figure 5: Dot plots of Belief Movement after a Signal 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. From (a) to (d), $n = 117$, $n = 114$, $n = 386$, $n = 425$. Dashed lines represent average movements. Differences in means and p-values for one-sided t-tests are reported at the bottom.

actual agent performance, but the former have much larger predictive power. By strongly relying on human difficulty to predict AI performance, subjects largely overestimate AI on easier tasks and underestimate it on harder ones.

5 AI Adoption in the Medium-Run under Human Projection

We study adoption decisions under Human Projection in the "medium-run", where we allow for many signals of performance but maintain projection fixed. We still consider the projection of human difficulty onto AI, using TIMSS problems as tasks. Instead of exogenous, one-shot signals of performance, we consider a setting where the principal receives (many) signals that are endogenous to their adoption decisions. We first theoretically characterize the equilibrium adoption decisions under Human Projection and compare them to an optimal benchmark. We

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

	Dep. Var: Belief Movement			
	Humans		**AI**	
	Success	**Failure**	**Success**	**Failure**
	(1)	(2)	(3)	(4)
Task Difficulty	0.205***	0.159***	0.164***	0.076**
	(0.052)	(0.048)	(0.028)	(0.031)
Controls	Yes	Yes	Yes	Yes
Prediction Task FE	Yes	Yes	Yes	Yes
R^2	0.235	0.179	0.100	0.074
Observations	117	114	386	425

Notes: Belief movement is the difference between posterior and prior (positive for successes and negative for failures). The main independent variable is the human difficulty of the signal task, for which performance was revealed. Controls include socio-demographic variables, and familiarity with AI (only in *AI*). Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

then run a lab experiment manipulating the degree of projection: we vary the "anthropomorphism" of AI and compare equilibrium adoption patterns.

5.1 Theoretical Framework

We minimally characterize the setup and focus on delivering intuition for the results we test in our experiment. We relegate the formal presentation of the model and proofs of statements to Appendix A.

Setup. A principal (e.g., a firm) is faced with two types of tasks: a human-easy task (t_E) and a human-difficult task (t_D),[17] which can be performed by either humans or AI. As in the basic framework of Section 2, the principal entertains the Ability Model: they perfectly observe signals of performance—success and failures—on tasks delegated to either agent, and update over ability θ (with prior distribution G^i). Signals of performance are endogenous: if the principal only delegates one type of task to AI, they only observe AI performance on this type of task.

We make some simplifying assumptions to be consistent with our experimental design. We assume the principal uses the same subjective mapping for both agents ($p^A = p^H = p$) and knows human ability θ^H but not AI ability θ^A. We abstract away from the costs of deploying human or AI technology by assuming that the principal's choices are made over (cost-weighted)

[17] To avoid confusion with the H denoting "Human," we use subscript D but still use "hard" and "difficult" interchangeably.

success rates (p_E^H, p_D^H) for humans and (p_E^A, p_D^A) for AI. The principal's objective function is increasing in expected success rates on each task, and there is no extra cost to an agent performing both tasks at once. The problem therefore boils down to using the best-performing agent for each task, which requires learning about whether $p_j^A \lesseqgtr p_j^H,\ \forall j \in \{E, D\}$.

Equilibrium. We characterize the pure strategy Berk-Nash Equilibrium (BkNE) (Esponda and Pouzo, 2016) with the principal as the single player of the game. *Actions* correspond to the four possible AI adoption decisions: No Adoption, Only Easy, Only Hard, and Full Adoption. *Consequences* refer to binary performance outcomes, which are perfectly observed. *Objective and subjective distributions* are bivariate and uncorrelated Bernoullis with true and perceived success rates as respective parameters. These distributions coincide for humans (since human ability is known), but can differ for AI when the principal engages in Human Projection: perceived rates $(\hat{p}_D(\theta; x), \hat{p}_E(\theta; x))$ may then depart from their true level.

The heuristic statement of a pure BkNE is an adoption decision such that there exists a belief in AI ability θ^A for which: (i) the adoption action is optimal given the belief; (ii) the belief best explains the consequences generated by the adoption decision.[18]

Adoption benchmark. When the principal's model is correctly specified, i.e. when not projecting human difficulty ($\lambda = 0$), they adopt the optimal strategy in equilibrium. For example, if success rates are such that AI is better at the human-hard task and humans are better at the human-easy task ($p_D^A > p_D^H$ and $p_E^A < p_E^H$), the principal chooses Only Hard in equilibrium.

Adoption under Human Projection. The following theorem characterizes the Berk-Nash equilibrium in the case where the principal fully projects ($\lambda = 1$) human difficulty:

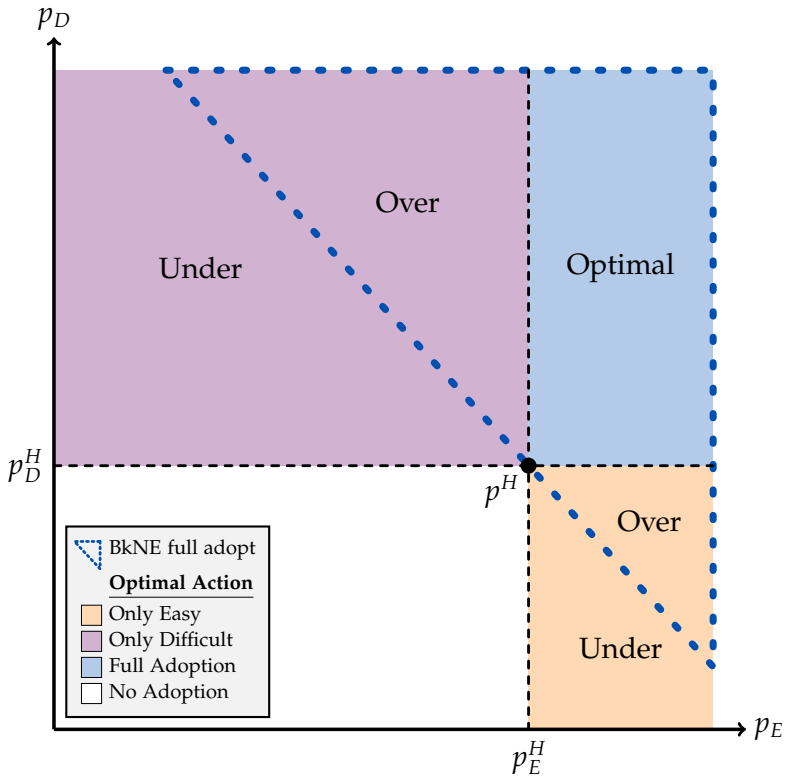
Theorem 1. *Let $\boldsymbol{p}^H$ and $\boldsymbol{p}^A$ denote human and AI success rate pairs. The Berk-Nash equilibrium is characterized as follows:*

1. *No adoption is always a Berk-Nash equilibrium.*
2. *Partial adoption is never a Berk-Nash equilibrium.*
3. *There exists a downward sloping hyperplane L that passes through $\boldsymbol{p}^H$ such that full adoption is a Berk-Nash equilibrium iff $\boldsymbol{p}^A$ lies above L.*

To build intuition, consider first the no adoption case. Since AI is not used for any task, no signals of AI performance are observed, such that any belief in AI ability—however pessimistic—can be sustained as BkNE. Consider then partial adoption, which is never an equilibrium. Signals of AI performance are observed on one of the tasks, and the principal forms a belief over θ^A which best explains this data. But there are only two possible outcomes (excluding the knife-edge case): either $\theta^A|_x > \theta^H$, in which case both $\hat{p}_D(\theta; x) \geq p_D^H$ and $\hat{p}_E(\theta; x) \geq p_E^H$ and the

[18] The sense in which the belief best explains consequences is that it minimizes the Kullback-Leibler divergence between $Q(\cdot|x)$ and $Q_\theta(\cdot|x)$, the objective and subjective distributions given observed consequences x. See Appendix A for details and a formal statement of the definition.

Figure 6: Adoption Equilibrium Region Under Human Projection



Notes: We present a stylized example of equilibrium adoption decisions under full human projection. The dashed blue triangle delimits all AI success rates (p_E^A, p_D^A) for which full adoption is sustained as BkNE. Points outside of the triangle are those for which no adoption is sustained as BkNE. The slope of the line passing through p^H is determined by the exact specifications of the Ability Model as well as distribution G. Colored zones represent optimal adoption decisions given the AI's success rates. Over-adoption arises when full adoption is sustained as BkNE while humans have a higher success rate than the AI either on the easy or difficult task. Under-adoption arises when no adoption is sustained as BkNE while AI has higher success rate either on the easy or difficult task.

principal fully adopts; or $\theta^A|_x < \theta^H$, and the principal does not adopt at all. Partial adoption cannot be sustained because the principal, as a result of Human Projection, draws an "all or nothing" conclusion and adopts accordingly. Finally for full adoption, the principal observes AI performance on both types of tasks, and forms a belief in ability which best explains the data leading again to an all-or-nothing conclusion. Figure 6 presents regions of the AI technology space for which full adoption is sustained as BkNE, and compares it with benchmark adoption zones. Human Projection creates patterns of both over- and under-adoption, as the principal fails to realize that absolute advantage in one task does not imply it in the other task.

These results assume a fixed AI technology (pair of success rates). In Appendix A we consider an extension where we allow for the AI "possibility frontier" to expand over time. We show first that, relative to the benchmark of no projection, AI adoption gets delayed. Then, upon crossing the downward sloping hyperplane (Figure 6), AI gets over-adopted: used even for the task where it performs worse than humans.

Experimental prediction. Our results have so far been about point predictions, comparing extreme cases where the principal does not project (adoption benchmark) or fully projects (Theorem 1). We make the following assumption about projection:

Assumption 3. *Anthropomorphic appearance of AI increases the likelihood of Human Projection (i.e., increases λ)*

This assumption is consistent with findings of experimental psychology: anthropomorphic traits (e.g., a name, a voice, the ability to explain or apologize, etc) of machines lead humans to behave more similarly to how they would behave with other humans (for a review, see Chugunova and Sele, 2022). In our context, this means that people are more likely to believe that what is difficult for humans is also difficult for AI, when AI appears more human-like. We can now state the comparative static prediction we test in our experiment:

Prediction 3. *A non-anthropomorphic presentation of AI decreases the share of "all or nothing" adoption (e.g., full adoption), by reducing the share of subjects fully projecting human difficulty.*

5.2 Adoption Experiment

We test Prediction 3 using TIMSS problems as mathematical tasks. We set up task difficulty levels to specifically test for a reduction in the share of Full Adoption, and we manipulate projection through the appearance of AI.

5.2.1 Design

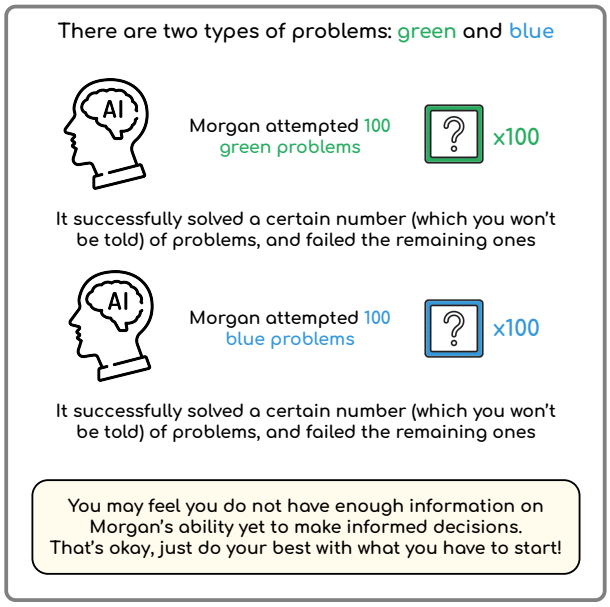
Our design provides participants with a costly opportunity to acquire signals about agent performance and inform a one-shot adoption decision they take at the end of the experiment. We vary the prevalence of Human Projection at the treatment level. Figure 27 in Appendix shows the structure of the experiment: the experimental variation is performed during instructions, followed by a "training phase" where participants can obtain signals of performance, which ends with a final adoption decision testing the equilibrium prediction.

Treatment variation. We vary the AI's presentation, which under Assumption 3 reduces the share of subjects fully projecting human difficulty (and increases average λ in the sample). In the *Anthropomorphic* condition, we use a standard AI framing: we present it as "Morgan," a neutral AI based on "currently available LLM-technology." We use the active voice to describe its behavior, and depict it using a human-like logo. As in the beliefs experiment of Section 4, we show non-math-related examples of AI answers using a dynamic "typewriter" effect.[19] This anthropomorphic framing is similar to those used by leading LLM companies, such as Anthropic's *Claude* or OpenAI's *ChatGPT*. It is consistent with findings of experimental psychology, linking

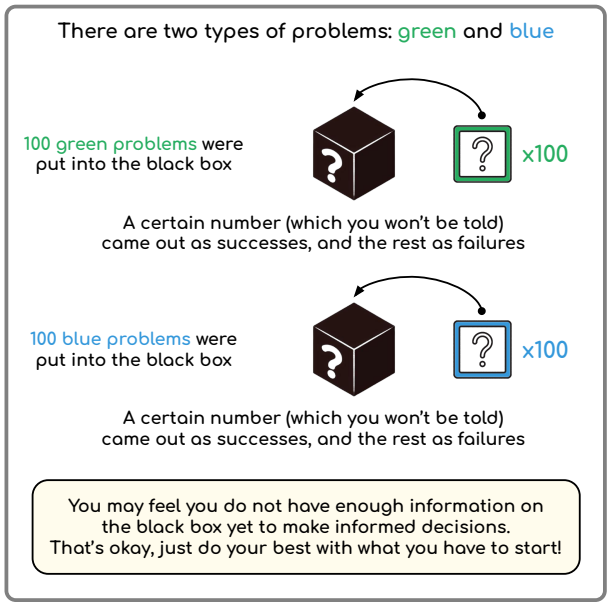
[19]We provide subjects with an example of a common prompt—e.g., writing a cover letter—and a clickable button to generate the answer. The pre-loaded answer then dynamically appears using Javascript. The answer text is the same as the one used in the beliefs experiment.

Figure 7: Treatment Variation in AI Framing

(a) Anthropomorphic Framing



(b) Black Box Framing



Notes: The figure presents the visual presented to subjects at the end of instructions. Alongside these pictures are presented 5 examples of both blue (easy) and green (hard) problems. This visual constitutes the main treatment difference: see Appendix D for more details on visuals used and the full text of experimental instructions.

anthropomorphic features to increased user trust in machines (e.g., Waytz et al., 2014; Kulms and Kopp, 2019; Troshani et al., 2021). In the other condition, *Black Box*, we present AI as a "black box:" we use the passive voice when describing its behavior, and omit any reference to a human being. We aim to make participants more agnostic about the relationship between human and AI difficulty, which implies a reduction in Human Projection. Figure 7 displays the main visual, presented before the start of the training phase, in which we use language inspired by recent work from Esponda et al. (2024).

Delegation and Adoption Decisions. Problems can be of two types: blue problems are (relatively) human-easy (mostly 4^{th}- and 8^{th}-grade, 78% human success rate), and green problems are human-hard (mostly High-School, 23% human success rate). AI success rate is kept constant at 66% across both types, so that humans are better in blue (easy) and AI is better in green (hard). We design the relative difficulty of tasks to focus on how the share of "full adoption" varies across treatments. In each treatment, we give the same information about the kinds of problems participants will encounter. We show a total of 6 examples of each type, and for one of them we also give human success rates (28% for green and 75% for blue) and AI performance (success on each). We then elicit prior beliefs in both human and AI performance, for each pool of problems, on a 0-100% scale: "What do you think is the success rate of [Morgan/the black box] in the [blue/green] problems?"

The training phase is composed of a fixed set of 60 problems, 30 of each type, presented in

random order. Each problem is shown with a colored border indicating its type: the participant chooses to delegate it to either a random human or the AI, and sees the performance of their pick (success or failure) on the next screen[20] We incentivize delegation decisions by providing a small bonus for each success obtained. As for priors, we elicit interim beliefs in the middle, and posterior beliefs at the end of the training phase. After the training phase, participants take a one-shot decision (with larger incentives) to adopt either a human or the AI to solve 10 tasks from each pool. The four possible AI adoption decisions are: No Adoption, Only Easy, Only Hard, and Full Adoption.

Logistics. We programmed the experiment using Qualtrics and recruited participants on Prolific during the month of August 2024. The tests pre-registration specifies the design, sample sizes, and main hypotheses, and the survey was designed to last around 15 minutes. We collected two samples for this design. The first has sizes of 150 in *Anthropomorphic* and 159 in *Black Box*. While we find our main pre-registered result with this sample, we also report non-optimal adoption behavior which we did not pre-register (we discuss this in Section 5.2.2). This behavior appears consistent with subjects misperceiving the relative *human* difficulty of the problem pools. To confirm this we collected a second sample, with sizes 59 and 58 respectively, with an identical survey design and procedures, except that human difficulty is made more salient: "blue" problems are (truthfully) said to be of "grade school" level, while "green" problems are said to be of "high school" level. We include results from the first sample and evidence for misperceptions in Appendix D. On average, subjects earned a total of \$4.2. The base fee was of \$1.7 with a potential bonus of \$4.1 (\$0.03 per success observed in the training phase, \$0.05 per accurate belief in performance, and up to \$2 for the adoption decision).

5.2.2 Results

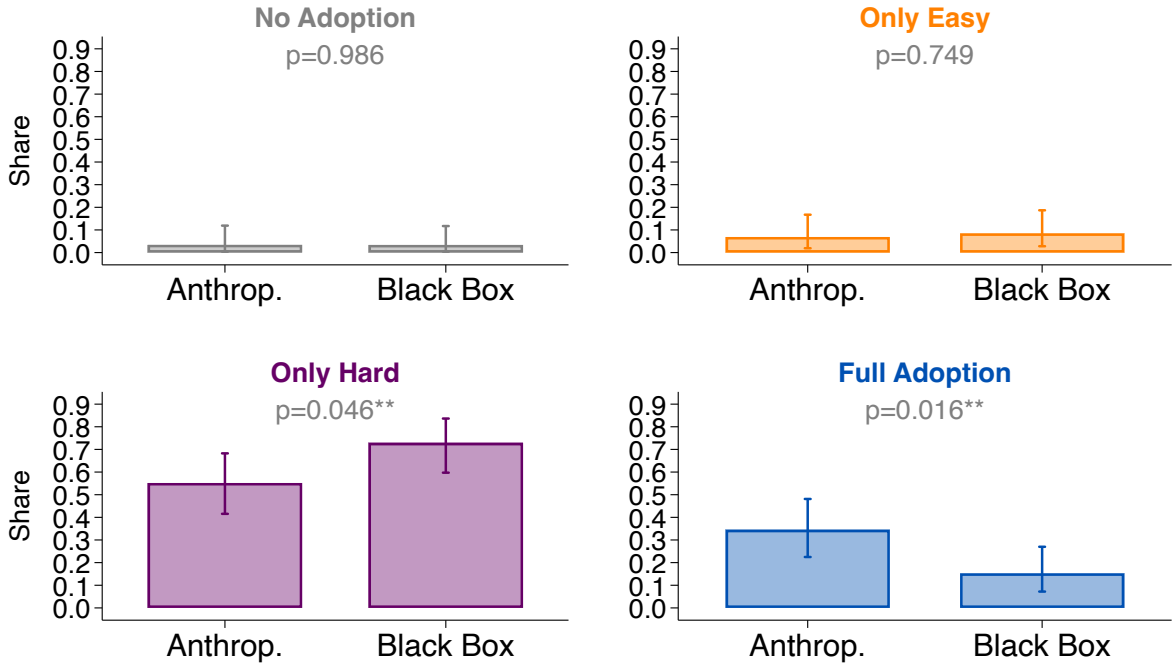
Proposition 3 predicts a lower share of Full Adoption in *Black Box* as it reduces λ, the degree of projection. Figure 8 reports results consistent with the prediction: 34% of participants fully adopt in *Anthropomorphic*, significantly more than the 15% in *Black box*. While prior beliefs do not matter for full adoption in the model, they may influence outcomes as the number of signals of performance is limited.[21] A mechanical effect of the *Black box* treatment is to reduce priors in performance: as we aim to make participants more agnostic regarding success rates, we also mechanically push their priors towards 50%. We control for prior influence by estimating the following specification:

$$\text{Full Adoption}_i = a_0 + a_1 \text{Anthropomorphic}_i + a_2 \text{Priors}_i + \mathcal{X}_i + \epsilon_i. \tag{3}$$

[20] Performance corresponds to actual performance gathered in Section 3. For humans, we draw from a Bernoulli with actual success rate.

[21] The Berk-Nash equilibrium is a fixed point argument, which means prior beliefs over ability do not affect convergence within the full adoption case. Our experiment includes a large but necessarily limited number of signals of performance, making prior beliefs a potential concern.

Figure 8: Shares of Adoption by Treatment



Notes: The figure presents the shares between Black Box and Anthropomorphic treatments for each type of adoption. Sample sizes are $n = 59$ for Anthropomorphic and $n = 58$ for Black Box. Confidence intervals at the 95% level are included.

where we regress the binary outcome of full adoption on the treatment dummy, prior beliefs P_i, and individual controls. Table 3 reports results: the prior shift partially explains the larger share of full adoption in *Anthropomorphic*, but the treatment coefficient remains significant when controlling for priors. We conclude that removing anthropomorphic features reduces the likelihood of "all-or-nothing" adoption and increases the optimality of adoption decisions when patterns of AI performance are un-human-like.

In Appendix D we report results from a first sample which was collected using an identical design, but which did not make salient the *human* difficulty of the blue and green pools of tasks. There, we find our main pre-registered effect—the lower share of Full Adoption in *Black Box*—but do not find an increase in the optimal choice of Only Hard, which appears consistent with subjects misperceiving the relative *human* difficulty of the task pools.[22] When accounting for such misperceptions, we find that *Black Box* indeed increases the share of (mis)perception-optimal adoption (see Figure 26). Results above come from the second sample obtained using a near-identical experimental procedure: when presenting blue and green tasks, we mention they are of "grade school level" and "high school level" respectively, thereby improving the salience of the human difficulty of task pools.

[22]To keep survey length within reasonable limits, this design (as opposed to the beliefs design of Section 4.1) does not ask nor incentivize subjects to solve the tasks. As a result, we report a significant share of subjects misperceiving the relative human difficulty of task pools (on the basis of reported priors in *human* performance).

Table 3: Treatment Effect on Full Adoption

	Dep. Var: Full Adoption			
	(1)	(2)	(3)	(4)
Anthrop. Framing	0.192**	0.208**	0.199**	0.197**
	(0.079)	(0.082)	(0.082)	(0.086)
Controls	No	Yes	No	Yes
Priors	No	No	Yes	Yes
R^2	0.050	0.094	0.059	0.096
Observations	117	112	117	112

Notes: This table reports OLS estimation results. Dependent variable is a dummy for full AI adoption, and the independent variable is the *Anthropomorphic* treatment dummy. Robust standard errors in parentheses. Prior beliefs in human and AI performance are included as control. Other controls include age, gender, income, and familiarity with AI. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Additional results. We also include some descriptive results on training phase behavior: beliefs in AI and humans are equal—by design—across treatments, but beliefs in AI significantly lower in *Black box*. These lower beliefs are consistent with subjects choosing to delegate fewer problems to AI during the training phase. Finally, as additional evidence for the treatment's effect on Human Projection, we look at the correlation between beliefs in AI performance. We simply regress reported beliefs about AI performance on hard tasks, on beliefs about AI performance in easy tasks. We find (in Appendix D) that beliefs are more strongly correlated across types of task in *Anthropomorphic*, and posterior beliefs even end up uncorrelated in *Black Box*. We interpret it as participants being better able to realize that better performance in one type of task does not imply it in the other (as it would under Human Projection, since these tasks differ by their human difficulty).

6 Projection of Similarity: Field Consequences for AI Engagement

In previous sections we study inferences from human-AI interactions (prompt and AI answer) that are fully exogenous from participants' perspective. We complement these results with evidence from a realistic field setting: parents asking an online AI chatbot for parenting advice. We study reactions to errors (unhelpful answers) when users project the feature of human textual similarity onto AI. When provided by a human, not all unhelpful answers are equal: inferences about the human's competence are more negative when the answer is highly *dissimilar* to the expected answer. The same logic applies to AI's unhelpful advice—leading to potentially in-

accurate expectations—when projecting what "similar" means to humans. In what follows we first provide some background on the context, then derive a belief updating prediction under projection of human textual similarity, and present both our field experiment design and results.

6.1 Background

ParentData.org is a self-defined "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 the latest scientific findings into rigorous answers to questions asked by current or expecting parents. It provides various services (articles, newsletters, podcasts) with several tiers of paid subscription. It also offers an AI chatbot service which is currently free to use and openly accessible.

User base. The pool of users is almost exclusively composed of people who are either trying to conceive, currently expecting, or parents of young children. The website does not collect user demographics, but a (non-representative) survey from July 2024 and discussions with the ParentData team describe the prototypical user as a (i) woman in her 30s, living in the U.S.; (ii) either expecting or mother of young children; (iii) educated; (iv) with higher income. We recruit all our experimental participants on the basis of these demographic features.

The AI. The website hosts Dewey, an "AI librarian," which is a LLM-based AI chatbot.[23] It initially summarized a variety of parenting material—books and articles—into a series of questions and answers, which were then verified and vetted by humans from the ParentData team. It is then available to answer user's questions. Upon receiving one, Dewey matches it to all premade questions using a confidence score, and displays 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."[24]

Interaction data consists of a snapshot of around 30000 conversations (defined as one user query and one AI answer) taking place between December 31st 2023 and April 30th 2024. Appendix E presents descriptive statistics, including the most frequent questions asked to Dewey.

6.2 Theoretical Framework in Context

We slightly adapt our framework to account for the specifics of this context. Here, task difficulty and answer correctness are highly subjective and complex to determine, since they relate to advice on e.g., best pregnancy diets, or how to discipline children. AI performance is also peculiar: Dewey can only provide human-vetted, high-quality answers, but it can fail to understand the

[23] This service is provided by the Dewey Labs company, of which ParentData.org is a client.

[24] The matching process is a custom approach to vector embeddings, whose details were not disclosed. We confirm the AI score correlates with various textual similarity measures. Below the threshold, the chatbot either gives question suggestions, 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!"

question and provide a useless answer (a good answer, but to a different question). We briefly describe the setup and prediction below and relegate formal details to Appendix A.

Setup. The principal (e.g., a user) is trying to predict the probability that an agent will provide a *useful* answer to a question. Performance is defined as usefulness and assumed to be binary. This simplification mirrors our experimental design which studies inferences from errors. 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 $A_i \subset \mathcal{A}$. 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.

Similarity projection. Assume each pair of answers (a, a') has a degree of *similarity* S, with $S(a, a) = 1$. S captures the *human* similarity of answers, depending on various factors such as semantic overlap, contextual proximity, shared meaning, etc. Define the *reasonableness* of answer a_j to question q_i by $r_{ij} = \max_{a \in A_i} S(a_j, a)$.[25] In other words, an answer's reasonableness is simply a certain way—the most "forgiving" one—to measure human similarity of a wrong answer to a correct one, by taking the maximal (human) similarity to an answer within the useful set A_i. If there is only one correct (useful) answer to question q_i, reasonableness and similarity coincide.

Dewey matches queries to answers using a score which is highly correlated with textual similarity measures, and displays the highest-score answer.[26] The AI answer to a user's question is thus the most reasonable—among *all* premade answers—as seen from the AI's perspective: i.e., maximally *AI-similar* to a useful answer. Therefore, in the absence of projection, *none* of the premade AI answers with below-maximal scores should be deemed more humanly-reasonable than the top-score answer, and user inferences should not depend on human reasonableness.

If instead people project human similarity onto AI then, holding fixed both the question and the answer's usefulness (performance), answers deemed less humanly-reasonable should lead to stronger negative inferences, which translate to lower trust and engagement with AI.

To gain intuition for this idea, which we test in experimental design, consider the following example question: "Which is the best car seat brand?" The first possible answer gives advice on where to install a baby car seat (front or back seat). The second one discusses which is the best baby food brand. The former is deemed more reasonable than the latter, because it at least shares the same context and thus appears more humanly-similar to a useful answer (which would discuss several seat brands and provide a recommendation).[27]

We now outline the rest of the theoretical setup and formulate our prediction. As before, we assume the principal perfectly observes agent performance (answer's usefulness). We define the Ability Model as follows: the probability that an agent with ability θ provides answer a_j to question q_i is given by $\tilde{p}(\theta, r_{ij})$, where $\sum_{j:a_j \in \mathcal{A}} \tilde{p}(\theta, r_{ij}) = 1$. We assume the subjective mapping

[25] This framework is designed to study inferences from AI failures, so among useful answers $r_{ij} = 1$ by construction.

[26] More precisely, Dewey matches the user's question to all its premade questions, and displays the answer corresponding to the question with the highest score. Since there is a one to one mapping between premade questions and answers, we directly focus on answer similarity.

[27] This example is part of the list of real conversations between users and Dewey that we use in the experiment; see next sections for details.

$\tilde{p}$ is increasing in θ and satisfies a MLRP assumption: the relative probability of giving a *more reasonable* answer is higher if the agent has higher ability. Under projection of human similarity, the human reasonableness of AI answers influences the principal's inferences over AI performance. We derive the following experimental prediction, which is the analog to Prediction 2 for AI failures:

Prediction 4. *When the principal observes a 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.*

6.3 Empirical Approach

We test Prediction 4 in a field experiment with users interacting with Dewey.[28] We randomly show participants (real) conversations, only varying the reasonableness of AI's answers—how humanly-similar they are to a useful answer—and measure subsequent engagement.

6.3.1 Selecting conversation pairs

The main challenge is to identify actual *pairs* of conversations[29] for which: (1) users are asking the same question, which got misunderstood by AI; (2) AI answers are (not) useful to the same extent; (3) one answer is significantly more reasonable—from a human perspective—than the other. We describe below the three-step process which we pre-registered. All participants we recruit (to rate conversations and in the experiment) exhibit demographics matching that of the actual ParentData.org user base, namely are parents of young children or currently expecting. The final list of conversations, experimental instructions and screenshots of decision screens are in Appendix E.

Initial labeling. We started by manually labeling around 2200 conversations for their "intent" (what the question is asking) and a binary measure of whether the question was misunderstood by the AI (a proxy for usefulness). We establish a list of around 40 same-intent pairs with both answers labeled as misunderstandings.

Measuring reasonableness. We then recruit parents to rate the reasonableness of conversations. Instructions put heavy emphasis on the fact that we were not asking for how useful the answers are, but rather in how reasonable the misunderstanding was perceived to be. Elicitation screens are identical for each conversation: one random conversation is shown,[30] using a layout that mimics the ParentData website, and we ask:

[28] Conversation data between users and Dewey unfortunately does not contain *any* information about users—beyond the queries they sent—which does not allow us to pursue a purely observational approach. This is because Dewey is a relatively recent service (introduced in May 2023) and is provided by a separate company. The web-behavior and chatbot sides of the data have not yet been merged, which prevent us from knowing important determinants of user engagement such as paid subscription status, history of prior engagement, demographics, etc.

[29] In what follows, a "conversation" will always refer to exactly one user query and one AI answer.

[30] To avoid contrast effects (which are absent from the engagement experiment), we split each pair down the middle to create two pools, from which conversations are randomly drawn for each subject.

> 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.

Measuring usefulness. We use an almost identical design to measure usefulness.[31] Instructions put heavy emphasis on eliciting pure usefulness and paying close attention to the AI answer. Elicitation screens are identical, and we ask:

> 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 list. We establish a final list of 5 pairs of conversations for which the sides of the pair: (i) have the same intent; (ii) are equally useless; (iii) differ in reasonableness. As pre-registered, we rely on medians as the main criterion: all pairs have the same median usefulness (either 1 or 2) but large differences in median reasonableness. As robustness, we confirm that all pairs have non-significant differences in average usefulness at the 90% level, but significant differences in reasonableness. In addition, we set aside 3 useful conversations which we also use in the experiment to increase baseline willingness to engage with AI. Appendix E presents the full list of conversations used, along with measures of usefulness and reasonableness.

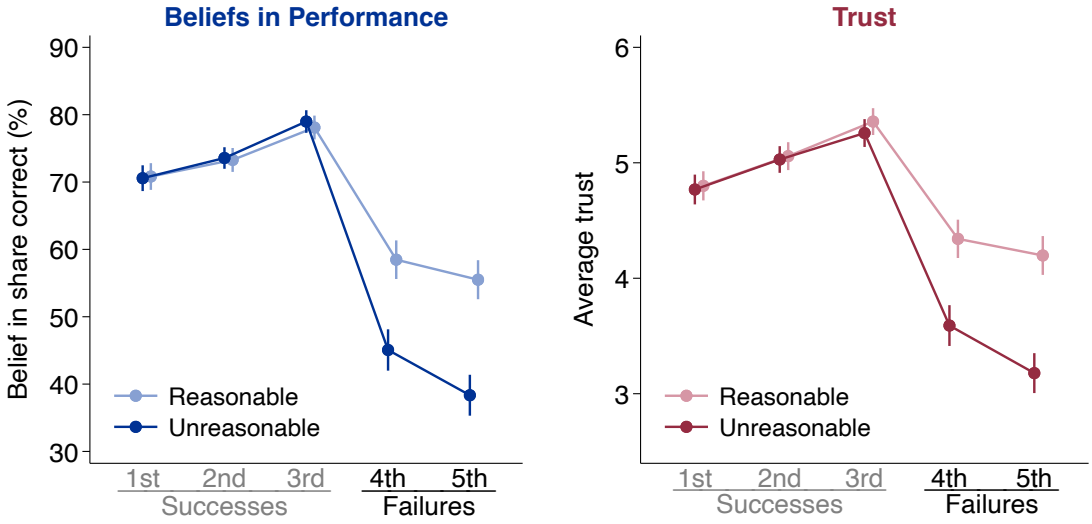
6.3.2 Engagement Experiment

Design. The experiment implements a between-subject design which only varies the type of AI failures. After instructions, subjects see a total of 5 real conversations between user and AI. The first 3 are "successes" (median usefulness of 4 or 5), held fixed across treatments. The last 2 are "failures" (median usefulness of 1 or 2). Failures are drawn from the final list of 5 pairs established above, and the treatment varies which side of the pairs gets displayed: subjects in *Reasonable* see the side deemed relatively more reasonable, while those in *Unreasonable* see the less reasonable side. Each conversation page initially displays the user query with a "Generate" button.[32] Upon clicking, the AI's answer dynamically appears using a "typewriter effect" similar to the one used on ParentData.org. We urge participants to read the answer carefully and then elicit beliefs in performance ("What do you think is the % chance the chatbot answers a random parenting question correctly?"; 0-100% scale) and trust ("How much do you trust the chatbot?"; 1-7 scale). We incentivize beliefs by drawing 100 conversations at random and eliciting usefulness from the same population of interest.

[31] The pool of conversations also included 4 useful conversations, for two reasons. First, to obtain a measure of usefulness for the 3 useful conversations used in the engagement experiment. Second, it would have been unnatural for subjects to only be presented with relatively useless conversations. Mentioning that conversations had been selected for being useless would have introduced concerns for demand effects.

[32] To increase attention to the mistaken conversations (the last 2), we use a visual nudge that we hold fixed across treatments: the user's question is shown in a box of a different color from the first 3.

Figure 9: Post-Conversation Beliefs and Trust



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.

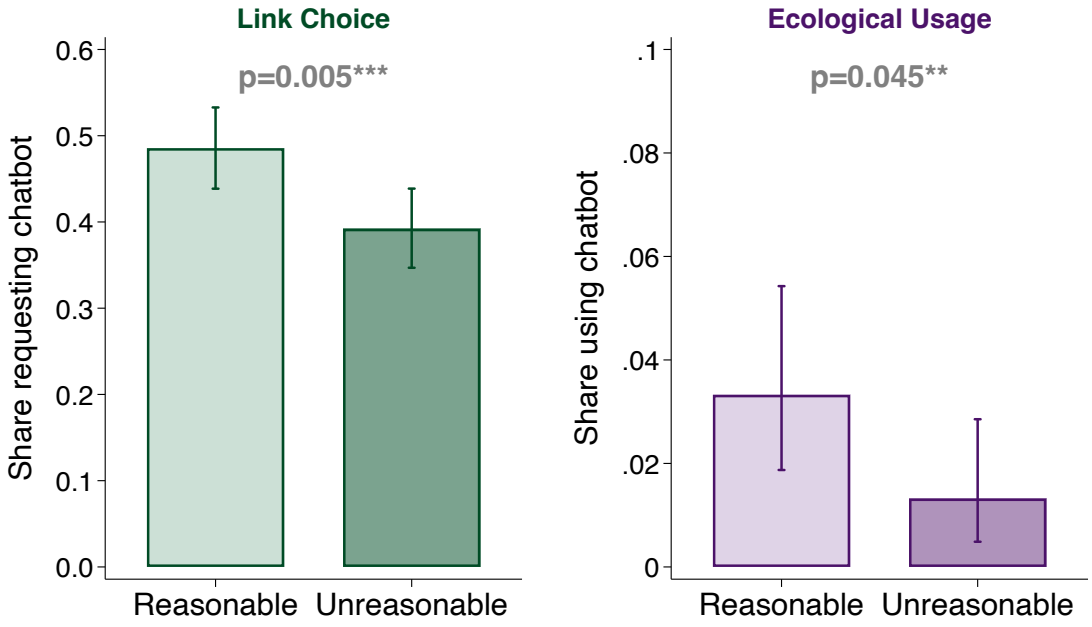
We elicit the main engagement decision right after the last conversation: subjects choose to receive either a link to the AI chatbot or to "a large list (500+) of good-quality parenting articles."[33] The chosen link opens in a new tab as subjects exit the survey. We make clear that both options are free and that there is no obligation to use them. As a secondary measure of engagement, we track actual AI usage by matching the IP address used during the experiment to those appearing in ParentData chatbot conversations.[34] We conclude the experiment by eliciting usual demographics and measures of prior familiarity with AI and with ParentData.org.

Logistics. We recruited all subjects (for usefulness and reasonableness ratings, and for the experiment) on Prolific during the months of July and August 2024. We target the population of interest in two steps: we pre-screen for a sample 85% female, living in the U.S., between 18 and 45 years old, and self-reporting as being in a relationship. We then screen at the beginning of the survey for current or expecting parents, excluding those reporting either "I am a parent of a child aged more than 18 years" or "I am not a parent and I am not trying to have children." Pre-registration includes: (1) the entire process for selecting pairs of conversations along with elicitation procedures; (2) the design of the engagement experiment; (3) sample sizes; (4) main hypotheses regarding engagement, beliefs and trust, along with secondary heterogeneity analysis. We recruited 905 subjects, 451 for *Reasonable* and 454 for *Unreasonable* treatments. The survey experiment took a little more than 5 minutes on average, with average earnings of \$0.87. Participants were paid a base fee of \$0.7, and could get a bonus of up to \$0.5 ($5 \times \$0.1$) for each

[33] Links are respectively https://parentdata.org/ask-a-question/ and https://parentdata.org/articles/

[34] For confidentiality reasons, we do not observe which questions these users asked.

Figure 10: Treatment Effect on AI Engagement Measures



Notes: The figure plots the share of subjects in each treatment choosing to receive the link to the chatbot over the link to parenting articles, and the share who actually use the chatbot after the experiment (defined as asking at least one question within 3 weeks after the experiment). Sample sizes are $n = 451$ for *Reasonable* and $n = 454$ for *Unreasonable*. p-values of two-sided tests of proportions are reported.

accurate belief about AI performance.

6.3.3 Results

To ensure effects are not driven by specific conversations or remaining differences in usefulness, all results include appropriate controls and conversation fixed effects when applicable.

We first look at beliefs in performance and trust. Figure 9 reports results consistent with Prediction 4: both beliefs and trust drop after failures, but the drop observed is significantly larger in *Unreasonable* for both beliefs (80% larger) and trust (75% larger). Appendix E.2 presents OLS regression estimates, with strong positive effect of reasonableness on beliefs and trust.

Engagement behavior reflects patterns observed for beliefs and trust. Figure 10 plots our measures of engagement: on the left panel, the share of subjects choosing to receive the chatbot link over the link to parenting articles; on the right panel, the share of subjects actually engaging with the AI (defined as asking at least one query within 3 weeks after the experiment). Both measures display a significant positive effect of reasonableness. Actual usage rates are unconditional, computed as the share of the entire sample who end up using the AI. Engagement rates conditional on requesting the AI link show qualitatively similar patterns (6.5% in *Reasonable* vs. 2.3% in *Unreasonable*; p-value = 0.048; see Appendix E.2), suggesting that even among participants who are sufficiently optimistic or curious about the chatbot to request a link, observing less humanly-reasonable failures decreases actual usage. As pre-registered, we perform an anal-

Table 4: Chatbot Engagement

	Engagement Measures			
	Chatbot Link		Actual Usage	
	(1)	(2)	(3)	(4)
Unreasonable	-0.094***	-0.095***	-0.021**	-0.022**
	(0.033)	(0.033)	(0.010)	(0.011)
Controls	No	Yes	No	Yes
Conversation FE	Yes	Yes	Yes	Yes
R^2	0.030	0.036	0.013	0.022
Observations	905	894	905	894

Notes: This table reports OLS estimation results, with dummy dependent variables for choosing the chatbot link, and using the chatbot within 3 weeks post-experiment. Independent variable is the treatment group. Robust standard errors in parentheses. Sample sizes are $n = 451$ for *Reasonable* and $n = 454$ for *Unreasonable*. We include conversation fixed effects, and usefulness as control. Other controls include age, gender, income, and familiarity with AI.

ysis excluding the small share of subjects who reported prior familiarity with ParentData.org: results are qualitatively similar, and even more pronounced for actual engagement (39% vs. 49%, p=0.005 for link choice; 0.9% vs. 3.4%, p=0.010 for actual usage; see Appendix E).

Finally, we estimate the following specification:

$$\text{Engagement}_i = a_0 + a_1 \text{Unreasonable}_i + \mathcal{C}_i + \mathcal{X}_i + \epsilon_i. \tag{4}$$

where *Engagement* is the binary decision to request the AI link, *Unreasonable* is the treatment dummy, and we include conversation fixed effects $\mathcal{C}_i$ and individual controls $\mathcal{X}_i$. Table 4 presents estimation results, which imply a strongly negative effect of humanly unreasonable AI answers on the likelihood of engagement.

Results taken together suggest users project human similarity onto the AI chatbot, as the human reasonableness of its errors—i.e., their human similarity to useful answers—strongly affects inferences and engagement behavior. While reasonableness is highly informative when making inferences about humans, it is less so in the case of AI: despite making un-human-like errors, Dewey is on average highly accurate and its answers are deemed very useful by users (see Figure 37 in Appendix E).

7 Conclusion

In this paper, we study how people evaluate AI performance and how it informs their adoption and usage decisions. We formalize and document Human Projection: the tendency to rely on human-relevant task features when forming expectations about AI performance. Observing a human fail at a human-easy task, or provide an answer that is highly humanly-dissimilar to the correct answer, strongly decreases expectations and trust because it is very diagnostic of the human's (low) ability. The same logic is projected onto AI, even though what is difficult (or similar) for humans may not be for AI. Inferences made under projection can then be misguided, leading to misspecified beliefs in performance and suboptimal adoption decisions. We provide three complementary pieces of experimental evidence for Human Projection. The first two are in the lab, to precisely measure projection and its equilibrium effects. The third is in the field, to show ecological consequences for user engagement with AI chatbots.

Our findings have three types of practical implications. First, we show (in Appendix A) that in dynamic settings where AI technology improves over time, HP first delays adoption and, once adoption occurs, leads to over-adoption. Second, we find a drawback to the type of AI "anthropomorphism" used by leading LLM companies like OpenAI or Anthropic. While human-like features increase user trust (Chugunova and Sele, 2022), they also increase the degree of Human Projection and de-align user expectations and AI capabilities. Third, our findings inform the design of human-AI interactions. In the Reinforcement Learning with Human Feedback (RLHF) phase of LLM training, human coders' ranking of possible LLM outputs may include the task's (human) difficulty, or the (human) reasonableness of the output.[35] Then, simple disclaimers regarding patterns of AI performance may improve the accuracy of users' perceptions.

Despite the misperceptions we documented, we do not take a stance on whether HP can be considered as an error or a type of projection "bias." The typical user has very little information on how a given AI functions, especially on whether the tasks they give to AI resemble those on which the AI was trained. Under such constraints, relying on a well-known prior—how humans would perform—seems natural and could be entirely rational. An alternative model, ignoring human features and updating task-to-task based on observed performance, would reduce misspecification but might converge more slowly, hindering beneficial AI adoption. In other words, there is a tradeoff for mental models between misspecification and speed of convergence: our findings highlight that the tradeoff depends on the actual correlation between AI performance and human features—such as difficulty—being projected.

How correlated is AI performance with human difficulty (or other features)? The evidence we put forth in this paper is not meant to be comprehensive, but rather provide a symptomatic example of a deeper misalignment between human and AI performance, which is still noted today for the best-performing models (Mialon et al., 2023; Xie et al., 2024). Our findings have con-

[35] RLHF, the last training step of modern LLMs, includes a series of questions to which the model provides several possible answers, which are then ranked by humans according to their quality. Some models only use a ranking (Ouyang et al., 2022), while some also include a measure of distance between ranks (Touvron et al., 2023), but they mostly ignore the human features we highlight in this work.

sequences for LLM's evaluation procedures: beyond maximizing average model performance on benchmarks of human-difficult tasks, reducing the variance of this performance and increasing its robustness on tasks that are straightforward for humans are important goals in developing human-centered models that align with users' expectations.

Finally, while our empirical evidence focuses on AI, nothing in our theoretical framework is AI-specific. Human Projection provides a simple structure to study human learning from features of observed performance, which leads to misspecification whenever the *perceived* covariation of performance with features departs from the actual one. Beyond new technologies, this framework may therefore be applied to humans. For example, basic linguistic mistakes from a non-native speaker may be incorrectly perceived as highly diagnostic of low intelligence or education, as they would be for a native speaker. Applying the concept of projection to human learning about unknown times, places or cultures opens up promising possibilities for future research.

References

Acland, Dan and Matthew R Levy, "Naiveté, projection bias, and habit formation in gym attendance," *Management Science*, 2015, *61* (1), 146–160.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining human expertise with artificial intelligence: Experimental evidence from radiology," Technical Report, National Bureau of Economic Research 2023.

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart, "Narratives about the Macroeconomy," 2023.

_ , Philipp Schirmer, and Johannes Wohlfart, "Mental models of the stock market," 2023.

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

Barron, Kai and Tilman Fries, "Narrative persuasion," 2023.

Bick, Alexander, Adam Blandin, and David Deming, "The Rapid Adoption of Generative AI," Technical Report, NBER Working Paper 32966 2024.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *The Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond, "Generative AI at work," Technical Report, National Bureau of Economic Research 2023.

Bushong, Benjamin and Tristan Gagnon-Bartsch, "Failures in Forecasting: An Experiment on Interpersonal Projection Bias," *Management Science*, 2024.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang et al., "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, 2024, *15* (3), 1–45.

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.

Danz, David, Lise Vesterlund, and Alistair J Wilson, "Belief elicitation: Limiting truth telling with information on incentives," Technical Report, National Bureau of Economic Research 2020.

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.

Enke, Benjamin, "What you see is all there is," *The Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

_ , Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen Van De Ven, "Cognitive biases: Mistakes or missing stakes?," *Review of Economics and Statistics*, 2023, *105* (4), 818–832.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh, "Replication: Belief elicitation with quadratic and binarized scoring rules," *Journal of Economic Psychology*, 2020, *81*, 102315.

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.

Fudenberg, Drew and Giacomo Lanzani, "Which misspecifications persist?," *Theoretical Economics*, 2023, *18* (3), 1271–1315.

Gabriel, Iason, "Artificial intelligence, values, and alignment," *Minds and machines*, 2020, *30* (3), 411–437.

Gagnon-Bartsch, Tristan and Antonio Rosato, "Quality is in the eye of the beholder: taste projection in markets with observational learning," 2022.

_ **and Benjamin Bushong**, "Learning with misattribution of reference dependence," *Journal of Economic Theory*, 2022, *203*, 105473.

_ **, Matthew Rabin, and Joshua Schwartzstein**, *Channeled attention and stable errors*, Harvard Business School Boston, 2018.

Grable, John, Ruth Lytton, and Barbara O'Neill, "Projection bias and financial risk tolerance," *The Journal of Behavioral Finance*, 2004, *5* (3), 142–147.

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

Green, Ben and Yiling Chen, "The principles and limits of algorithm-in-the-loop decision making," *Proceedings of the ACM on Human-Computer Interaction*, 2019, *3* (CSCW), 1–24.

Gruber, Jonathan, *Risky behavior among youths: An economic analysis*, University of Chicago Press, 2009.

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.

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

Imas, Alex, Minah H Jung, Silvia Saccardo, and Joachim Vosgerau, "The Impact of Joint versus Separate Prediction Mode on Forecasting Accuracy," Technical Report, National Bureau of Economic Research 2022.

Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang et al., "Ai alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2023.

Johnson-Laird, Philip Nicholas, *Mental models: Towards a cognitive science of language, inference, and consciousness* number 6, Harvard University Press, 1983.

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

Kulms, Philipp and Stefan Kopp, "More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation," in "Proceedings of mensch und computer 2019" 2019, pp. 31–42.

Lai, Vivian, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan, "Human-ai collaboration via conditional delegation: A case study of content moderation," in "Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems" 2022, pp. 1–18.

Lanzani, Giacomo, "Dynamic concern for misspecification," *Available at SSRN 4454504*, 2022.

Lichtenstein, Sarah and Baruch Fischhoff, "Do those who know more also know more about how much they know?," *Organizational behavior and human performance*, 1977, *20* (2), 159–183.

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.

Martinez-Plumed, Fernando and Jose Hernandez-Orallo, "Dual indicators to analyze ai benchmarks: Difficulty, discrimination, ability, and generality," *IEEE Transactions on Games*, 2018, *12* (2), 121–131.

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.

Natarajan, Manisha and Matthew Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in "Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction" 2020, pp. 33–42.

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

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray et al., "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, 2022, *35*, 27730–27744.

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

Shankar, Shreya, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo, "Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences," in "Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology" 2024, pp. 1–14.

Terry, Michael, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris, "AI alignment in the design of interactive AI: Specification alignment, process alignment, and evaluation support," *arXiv preprint arXiv:2311.00710*, 2023.

Terwiesch, Christian, "Would chat GPT3 get a Wharton MBA," *A prediction based on its performance in the operations management course*, 2023.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

Troshani, Indrit, Sally Rao Hill, Claire Sherman, and Damien Arthur, "Do we trust in AI? Role of anthropomorphism and intelligence," *Journal of Computer Information Systems*, 2021, *61* (5), 481–491.

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.

Wallach, Hanna, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow et al., "Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge," *arXiv preprint arXiv:2502.00561*, 2025.

Wang, Dakuo, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini, "How much automation does a data scientist want?," *arXiv preprint arXiv:2101.03970*, 2021.

Wang, Zhichao, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur et al., "A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more," *arXiv preprint arXiv:2407.16216*, 2024.

Waytz, Adam, Joy Heafner, and Nicholas Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of experimental social psychology*, 2014, *52*, 113–117.

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.

You, Sangseok, Cathy Liu Yang, and Xitong Li, "Algorithmic versus human advice: does presenting prediction performance matter for algorithm appreciation?," *Journal of Management Information Systems*, 2022, *39* (2), 336–365.

ONLINE APPENDIX

A Theoretical Appendix

This section provides the models' details and formal statements of experimental predictions presented in the main text. All proofs are presented at the end of this section.

A.1 Human Projection Framework

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 indexed by $k \in \{1, , K\}$ such that $(t_1, ..., t_K) \in \{0,1\}^K$ denote possible outcomes (success or failure) for each task of a random draw from some joint distribution over $\{0,1\}^K$. The vector $\boldsymbol{s} = (s_1, ..., s_K)$ 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 $k \in \{1, , K\}$. These success rates are unknown to the principal.

We assume that the principal entertains the following mental model: the agent has a unidimensional type denoted $\theta \in \Theta \subseteq \mathbb{R}$, and is 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. We denote the difficulty of problem t_k by $\delta(t_k)$, and a problem with a level of difficulty δ by t^δ. The probability that an agent with ability θ succeeds in solving a given problem t^δ is given by $p : \Theta \times \Delta \to [0,1]$. We assume the mapping $p(.,.)$ satisfies the properties of Assumption 1, which constitutes the Ability Model. 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 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. The other extreme of the distribution, i.e., tasks so difficult that even very high ability agents have a null success rate, would be captured by a MLRP assumption with reversed inequalities between ratios. We also 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 $\delta(t)^A$ but receives a noisy signal of difficulty centered around the truth $s_t = \delta(t)^A + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^s)$. Since human difficulty is a natural anchor for the principal when trying to assess AI difficulty, we assume that the principal's prior is centered around the human difficulty: $\delta_0^A \sim \mathcal{N}(\delta^H, \sigma^H)$. The principal then combines the signal with his prior of difficulty and obtains a posterior mean $\tilde{\delta}$ for AI. The mean takes the usual "shrinkage" form:

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

Taking the expected value of this expression we get the expression for $\tilde{\delta}^A(t)$ from Assumption 2. Our assumption thus follows directly, assuming the principal plugs the mean posterior difficulty into the mapping p^A. Notice that λ depends positively both on the precision of its prior (σ^H) and the imprecision of the signal on AI difficulty (σ^s).

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.

The only type of projection we assume is that of the human feature relevant for inferences about performance (here, task difficulty). We let priors over ability ($G^H \neq G^A$) differ, which would happen if one has higher uncertainty about AIs, or thinks machines are generally better than humans in the relevant domain. We also let the mappings themselves (p^A and p^H) be agent-specific.

The Ability Model only imposes a light structure on the principal's updating problem. It implicitly assumes a consistent performance ranking among agents across all tasks and among tasks for all agents, which for humans follows directly from the definition of difficulty, based on average performance, and for AI is at least sometimes verified (Martinez-Plumed and Hernandez-Orallo, 2018). This assumption is natural when the principal is faced with a large number of tasks to make inferences on: keeping track of task-specific success rates is both cognitively costly and impractical, if the number of signals of performance they can observe is limited. In other words, the Ability Model 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.

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

Proposition 1. *The predicted success rate is decreasing in δ^H for both humans and AI.*
Formally, $\forall i \in \{H, A\}$:

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

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

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

Proposition 2 holds for posterior rates on any task, and for any difference in human difficulty on the observed task. 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).

The types of extrapolations highlighted by Proposition 2 is only an issue for belief misspecification if the *perceived* co-variation of performance and task features departs from the *actual*

co-variation. In this case, if AI performance does not vary with (human) task difficulty the way that human performance does. 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.

A.2 Long-Run Equilibrium Adoption under HP

A.2.1 Baseline Model

Setup. A principal (e.g., a firm) manages two types of tasks: an easy task (t_E) and a difficult task (t_D), where $\delta^H(t_E) < \delta^H(t_D)$. Tasks can be performed by humans or AI with success rates denoted $\boldsymbol{p} = (p_E, p_D) \in [0,1]^2$. The human agent has known ability θ^H, determining success rates via $p^H(\theta^H, \delta^H)$. The AI success rates, denoted by $\boldsymbol{p}^A = (p_E^A, p_D^A)$, are instead unknown to the principal.

The principal believes that AI has ability $\theta^A \sim G^A$ over Θ, and that its success rates follow the human mapping: $p(\theta^A, \delta_k)$.

The Berk-Nash equilibrium (BkNE) for beliefs and actions includes the following elements:

1. **Actions (X):** Principal chooses AI or human for each task. X is the power set of $\{e, d\}$.
2. **Consequences ($Y \in \{0,1\}^2$):** Observed binary performance (success/failure) tasks.
3. **Payoff ($R : Y \to \mathbb{R}$):** Increases with Y elements.
4. **Objective distribution ($Q(\cdot|x)$):** True distribution given an action x. Bivariate uncorrelated Bernoulli with success rates based on the chosen agent for each task according to x.
5. **Subjective distribution ($Q_\theta(\cdot|x)$):** Distribution given an action x and a belief about AI ability θ. Bivariate independent Bernoulli with success rates $(\hat{p}_D(\theta; x), \hat{p}_E(\theta; x))$.
6. **Subjective expected payoff ($E_\theta(R|x)$)**: Expected payoff given action x and belief θ.

Each adoption action $x \in X$ induces an objective distribution over consequences $Q(\cdot \mid x) \in \Delta(Y)$, which is bi-variate uncorrelated Bernoulli with success rates $\tilde{p}_E(x)$, and $\tilde{p}_D(x)$ where $\tilde{p}_k(x)$ is the success rate of the agent chosen to perform task k according to action x.[36]

Adoption actions also influence the subjective distribution over consequences $Q_\theta(\cdot|x)$, but through the prism of Human Projection: the principal believes that AIs have ability θ^A distributed G^A over Θ. For simplicity, we assume full projection ($\lambda = 1$) and identical subjective mappings: AI success rates are given by $p(\theta^A, \delta_k)$.[37] Notice that whenever the AI success rate

[36] Using the notation of Esponda and Pouzo (2016), we have $\Omega = [0,1]$, $\omega \sim U(0,1)$, and

$$y = f(x, \omega) = (\mathbf{1}(\omega \geq 1 - \tilde{p}_E(x)), \mathbf{1}(\omega \geq 1 - \tilde{p}_D(x))$$

where $\mathbf{1}(\cdot)$ is an indicator function.

[37] Partial projection can still be accommodated at the group level, assuming a fraction of principals fully projects, while the remainder does not.

pair cannot be explained by some ability θ, i.e., whenever $\boldsymbol{p}^A \notin \{(p(\theta, \delta_E), p(\theta, \delta_D)) : \theta \in \Theta\}$, HP is a form of model misspecification. For technical reasons, we assume that for any success rate in a particular task can be rationalized by some ability level, i.e., for any $p' \in [0,1]$ there exists $\theta^A \in \Theta$ such that $p(\theta^A, \delta_k) = p'$ for $k \in \{E, D\}$.[38]

Distance to true model. In equilibrium, the principal only entertains subjective distributions that "best explain" observed consequences. This notion is given by minimizing the Kullback-Leibler divergence between the objective and subjective distributions. This divergence between distributions $Q(\cdot \mid x)$ and $Q_\theta(\cdot \mid x)$ is given by

$$K(x, \theta) = E_{Q(\cdot|x)}\left[\ln \frac{Q(y \mid x)}{Q_\theta(y \mid x)}\right] = \sum_{\hat{y} \in \{0,1\}^2} Q(\hat{y} \mid x) \ln \frac{Q(\hat{y} \mid x)}{Q_\theta(\hat{y} \mid x)}$$

Berk-Nash Equilibrium. A (pure) *Berk-Nash* equilibrium is an action $x^* \in X$ such that there exists $\theta(x^*) \in \Theta$ for which:

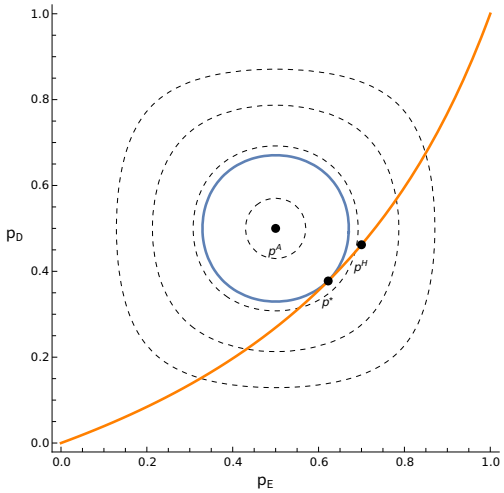
(i) x^* maximizes $E_{\theta(x^*)}(R \mid x)$.

(ii) $\theta(x^*)$ minimizes $K(x^*, \theta)$.

In words, the Berk-Nash equilibrium requires actions to be optimal with respect to beliefs, and beliefs to minimize the KL divergence from the objective distribution over consequences induced by the actions. Figure 11 shows an example of the success rates implied by KL-minimizing θ^A when $x =$ full adoption. In this example, the AI has an advantage in the difficult tasks, and a disadvantage in the easy tasks. The KL-minimizing AI ability is $\theta^* < \theta^H$. This process then leads to the "triangle" of AI success rates sustained as BkNE depicted in Figure 6 in the main text. Theorem 1 in the main text then characterizes the Berk-Nash equilibrium under Human Projection.

[38] This assumption is equivalent to part (iii) in Assumption 1 in Esponda and Pouzo (2016) which guarantees the existence of a Berk-Nash equilibrium.

Figure 11: KL Minimization



Notes: An example of KL minimization for $x =$ full adoption. The orange curve represents the subjective possibility curve: the set of possible points under HP, where p^H is the known human success rate. The dashed lines represent sets of points that have the same KL divergence from the truth (p^A). p^* is the KL-minimizing point, where the KL-equidistant curve (in blue) is tangent to the subjectively possible curve.

A.2.2 Extension: Dynamics of Adoption

We develop an extension where we allow technology to develop over time, and characterize the adoption path under HP, relative to a well-specific model benchmark.

We analyze a continuous-time model, where in each period $\tau \in [0, \overline{\tau}]$, there is a convex set of feasible AI success-rate pairs $\boldsymbol{p}$, denoted by $\mathcal{P}(\tau)$. For concreteness, we assume $\mathcal{P}(0) = \{(0,0)\}$ and $\mathcal{P}(\overline{\tau}) = [0,1]^2$. We assume that technology evolves monotonically and continuously. Monotonicity means that that (i) for any $\tau' > \tau$ we have $\mathcal{P}(\tau) \subset \mathcal{P}(\tau')$, and (ii) for any $p \in \mathcal{P}(\tau)$ if p' is dominated by p then $p' \in \mathcal{P}(\tau)$. Continuity means that the correspondence $\mathcal{P}(\tau)$ is continuous (in particular, upper and lower hemicontinuous) with respect to τ. Last, we denote the *technological frontier*, (i.e., the set of undominated success-rate pairs) in period τ by $\mathcal{P}^*(\tau)$ and assume that for any $\tau > 0$ we have $|\mathcal{P}^*(\tau)| > 1$.

Under this minimal structure on the evolution of technology, the following theorem shows that optimal adoption always starts with no adoption, followed by partial adoption, eventually converging to full adoption.[39]

Theorem 2 (Optimal Path). *Let $\boldsymbol{p}^H \in (0,1)^2$. Then there exist τ_1, τ_2 such that*

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

[39] Notice that after the first period where partial adoption is optimal, the economy could, in principle, alternate between different types of partial adoption and full adoption, but it will eventually converge to full adoption.

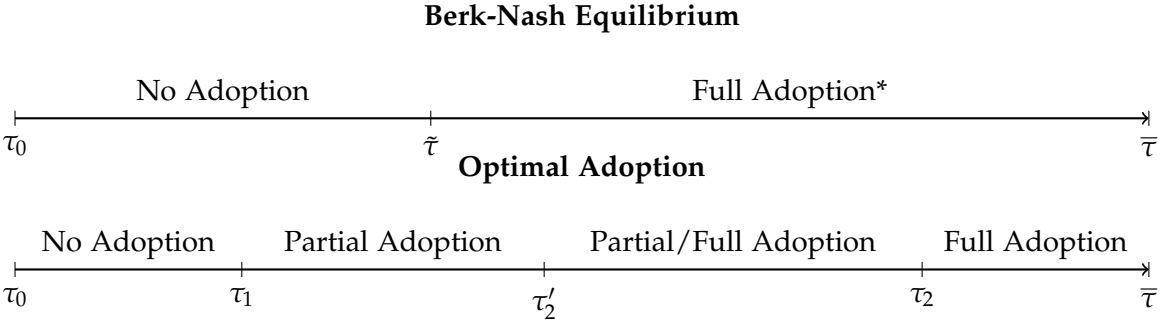
We compare this benchmark to the adoption path under HP, presented in the following result:

Theorem 3 (HP Path)**.** *Under HP, there exist $\tilde{\tau} \in (\tau_1, \tau_2]$ such that*

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

This theorem shows that (i) compared to the optimal path, adoption is delayed, but that (ii) at the early stages of adoption, over-adoption arises. In particular, for $\tau \in [\tau_1, \tilde{\tau}]$, some adoption is optimal, and yet no adoption is the unique BkNE. Second, it is typically the case that under HP, during the early stages of adoption we have *over adoption*: for $\tau \in [\tilde{\tau}, \tau_2',]$, full adoption is dominated by partial adoption, and yet it is a BkNE.[40] Therefore, conditional on adoption being an equilibrium, the firm delegates suboptimally too many tasks to the AI at first. Figures 12 illustrates the optimal and HP adoption paths.

Figure 12: Adoption Path under HP vs. Optimal Adoption Path



Notes: This timeline represents the different stages of adoption under HP relative to the optimal action. "Full Adoption*" represents the stage where full adoption is a Berk-Nash equilibrium, but recall that by Theorem 1, no-adoption is always a Berk-Nash equilibrium..

A.3 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.[41]. 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 $S(a, a) = 1$. S captures the *human* similarity of answers, depending on various factors such as semantic overlap, contextual proximity, shared meaning, etc. Then define the *reasonableness*

[40] Notice that outside a knife-edge case where $\tilde{\tau} = \tau_2'$, the length of this segment is strictly positive.

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

of answer a_j to question q_i by $r_{ij} = \max_{a \in A_i} S(a_j, a)$. 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 thus simply a certain way to measure similarity to useful answers: it is the most "forgiving" way, because it considers the maximum possible similarity of the (useless) answer to a useful answer.

We make the following assumption, which is the counterpart of the MLRP part of Assumption 1 in the main text.

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

The relative probability of giving a *more* reasonable answer is higher if the agent has higher ability. As before, we denote the user's posterior over the agent's ability by $G_{|x}$, given data $x = (r_{ij}^1, \ldots, r_{ij}^T)$. Notice that the reasonableness of a question-answer pair is a sufficient statistic to make inference on θ. The subjective probability of providing a useful answer is then $p(q_i \mid x) \equiv E_{G_{|x}} \left(\sum_{a_j \in A_i} \tilde{p}(\theta, 1) \right)$.

We thus can the following result about belief updating, which is the formal statement of Prediction 4 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 induces stronger negative inference over θ, which in turn decreases the perceived chance of providing a useful answer as $p(.|x)$ is increasing in θ.

Further assuming that trust in the agent and willingness to ask questions to them are increasing functions in perceived likelihood of providing a useful answer, we obtain the last two parts of Prediction 4 in the main text.

A.4 Proofs

Proposition 1. By assumption, $p(.,.)$ is decreasing everywhere in δ, and $\partial \tilde{\delta}^A / \partial \delta^H \geq 0$ for any $\lambda \in [0, 1]$, which gives the result. □

Proposition 2. We prove the first part, the proof of the second part is symmetric. Since $p(\theta, \delta)$ is monotonically increasing in θ, it is enough to show that $G_{|t^{\bar{\delta}}=1} \underset{FOSD}{\succ} G_{|t^{\underline{\delta}}=1}$. Let $\Theta = [\underline{\theta}, \bar{\theta}]$. Then

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

$$= \frac{\int_{\underline{\theta}}^{\theta} g(s) p(s, \delta) ds}{\int_{\underline{\theta}}^{\theta} g(u) p(u, \delta) du + \int_{\theta}^{\bar{\theta}} g(u) p(u, \delta) du}$$

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

Focusing on the second expression in the denominator:

$$\frac{\int_{\theta}^{\overline{\theta}} g(s)p(s,\delta)ds}{\int_{\underline{\theta}}^{\theta} g(u)p(u,\delta)du}$$

$$= \int_{\theta}^{\overline{\theta}} \frac{g(s)p(s,\delta)}{\int_{\underline{\theta}}^{\theta} g(u)p(u,\delta)du} ds$$

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

Notice that $u \leq s$, and therefore by MLRP the derivative of $\frac{p(u,\delta)}{p(s,\delta)}$ w.r.t. δ is negative, which implies that the derivative of the whole expression (which is essentially a weighted average of the inverse of $\frac{p(u,\delta)}{p(s,\delta)}$) is positive. This then implies that $G(\theta \mid t^{\delta} = 1)$ is decreasing in δ, i.e., $\delta' > \delta \iff G(\theta \mid t^{\delta'} = 1) \underset{\text{FOSD}}{\succ} G(\theta \mid t^{\delta} = 1)$, and in particular, $G_{|t^{\overline{\delta}}=1} \underset{FOSD}{\succ} G_{|t^{\underline{\delta}}=1}$ as needed.

□

Proposition 3. We first show the subjective probability of providing a useful answer is increasing in θ. Notice that given q_i we have:

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

Since in the numerator we have $r_{ij} < 1$, by MLRP the expression decreases in θ.

By MLRP, using a similar argument as for Proposition 2, we have $r' > r \implies G_{|r'} \succ_{FOSD} G_{|r}$. Then, $\forall q_i \in \mathcal{Q},\ G_{|r'} \succ_{FOSD} G_{|r} \implies \mathbb{E}_{G_{|r'}}[q_i|r'] \geq \mathbb{E}_{G_{|r}}[q_i|r]$. This completes the proof. □

Theorem 1. Denote the highest achievable success rate for task $t \in \{E, D\}$ in period τ by $\overline{p}_t(\tau) \equiv \max\{p_t : (p_E, p_D) \in \mathcal{P}(\tau)\}$.

Since $\mathcal{P}(0) = \{(0,0)\}$, by the continuity of $\mathcal{P}(\tau)$, there is a range where no adoption is optimal. To see why, assume otherwise, i.e., for all $\tau > 0$, $\mathcal{P}(\tau)$ includes $p_t \geq p_t^H$ for some $t \in \{E, D\}$. Assume WLOG $t = E$. Take a sequence $(a_n)_{n=1}^{\infty}$ where $a_n = \frac{1}{n}$. Clearly $a_n \to 0$. Define the sequence $(b_n)_{n=1}^{\infty}$ where $b_n = (p_E^H, 0)$. Notice that $b_n \in \mathcal{P}(a_n)$ for all n and that $b_n \to (p_E^H, 0) \equiv b$. Upper hemicontinuity implies $b \in \mathcal{P}(0)$, but clearly $(p_E^H, 0) \notin \mathcal{P}(0) = \{(0,0)\}$, a contradiction.

Using a similar argument (alebeit this time using lower hemicontinuity) it is straightforward to see that since $(1,1) \in \mathcal{P}(\overline{\tau})$, there is a range where full adoption is optimal.[42] This proves

[42] Assume in contradiction otherwise, then for any $\tau < \overline{\tau}$ either $\overline{p}_E(\tau) < p_E^H$ or $\overline{p}_D(\tau) < p_D^H$. Take the sequence

parts (1) and (3).

To prove part (2), assume by contradiction that for all τ either no or full adoption is optimal. Let $\mathcal{P}_t(\tau)$ denote the set of feasible success rate for t-type problem at time τ (i.e., the first/second element in the pairs that are in $\mathcal{P}(\tau)$).

Define $\tau_t \equiv \min\{\tau : p_t^H \in \mathcal{P}_t(\tau)\}$, i.e., the first period where AI is at least as good as a human in task $t \in \{E, D\}$.

Notice that if $\tau_D < \tau_E$ then for $\tau \in [\tau_D, \tau_E]$ partial d-adoption is optimal. Similarly, if $\tau_E < \tau_D$ then for $\tau \in [\tau_E, \tau_D]$ partial e-adoption is optimal. The only case left is $\tau_E = \tau_D$. We will show that this can never be the case.

Assume $\tau_E = \tau_D \equiv \tau_H$. By the lemma proven below, p^H is at the frontier $\mathcal{P}^*(\tau_H)$. Since we assume that the frontier is never a singleton, there exists $p' \in \mathcal{P}^*(\tau_H)$, $p' \neq p^H$.

Since both p' and p^H are at the frontier, it must be the case that $p'_E > p_E^H$ or $p'_D > p_D^H$. Assume WLOG $p'_E > p_E^H$. Take an arbitrarily small ball around p', $B(p')$ such that for all $p \in B(p')$, $p_E > p_E^H$. Notice that the intersection of $\mathcal{P}(\tau_H)$ and $B(p')$ is non-empty.

By lower hemicontinuity, there exist $\tau < \tau_H$ (in a small enough neighborhood of τ_H) such that the intersection of $\mathcal{P}(\tau_H)$ and $B(p')$ is non-empty. However, since τ_H is the first period where $\mathcal{P}(\tau)$ includes p_E^H, by monotonicity, if $\tau < \tau_H$, $\mathcal{P}_E(\tau)$ cannot contain an element $p_E > p_E^H$, a contradiction.

The following lemma completes the proof by verifying that at τ_H, p^H is at the frontier:

Lemma 1. *At τ_H, p^H is at the frontier, i.e., $p^H \in \mathcal{P}^*(\tau_H)$.*

Proof. Assume that p^H is not at the frontier. Let $\tilde{p} \in \mathcal{P}^*(\tau_H)$ be a point at the frontier that strictly dominates p^H. Then there exists an arbitrarily small open neighborhood around $\tilde{p}$, denoted $N(\tilde{p})$, such that all elements in $N(\tilde{p})$ strictly dominate p^H. Notice that by monotonicity, if $p^H \notin \mathcal{P}(\tau)$, then $N(\tilde{p}) \cap \mathcal{P}(\tau) = \emptyset$.

By definition, for all $\tau' < \tau_H$ we have $p^H \notin \mathcal{P}(\tau')$, which implies $N(\tilde{p}) \cap \mathcal{P}(\tau)' = \emptyset$, so all open neighborhoods around τ_H contain points whose image does not intersect with $N(\tilde{p})$. This contradicts lower hemicontinuity. □

□

Theorem 2. The statements about $K(\cdot, \cdot)$ are proved below.

Starting with the first part, notice that under no-adoption, any belief about AI ability perfectly rationalizes the observed consequences, which are only a function of (known) human ability. Formally, since $\tilde{p}_t(\emptyset) = \hat{p}_t(\theta; \emptyset) = p_t^H$, we have $Q(\cdot \mid \emptyset) = Q_\theta(\cdot \mid \emptyset)$, and in particular $K(\emptyset, \theta) = 0$, for all θ. Therefore, if we choose any $\theta \leq \theta^H$, no adoption indeed maximizes profit.

$(a_n)_{n=1}^\infty$ where $a_n = \overline{\tau} - \frac{1}{n}$. Clearly $a_n \to \overline{\tau}$. Let $b = (1, 1)$. Since $b \in \mathcal{P}(\overline{\tau})$, by lower hemicontinuity there exists a subsequence $(a_{n_k})_{k=1}^\infty$ and a sequence $(b_k)_{k=1}^\infty$ such that (i) $b_k \in \mathcal{P}(a_{n_k})$ and (ii) $b_k \to (1, 1) = b$. This is clearly a contradiction since if we assume that full adoption is not optimal for all $\tau < \overline{\tau}$, there is a lower bound on the distance between all $b_k \in \mathcal{P}(a_{n_k})$ and $(1, 1)$ and so $b_k \not\to (1, 1) = b$.

Moving to the second part, to arrive at a contradiction assume that d-adoption is a Berk-Nash equilibrium (the proof about e-adoption is symmetric).

Let $\theta_D = p_D^{-1}(p_D^A)$, i.e., the ability level that explain the AI observed success rate in the difficult task.[43] We have

$$\begin{aligned}K(d,\theta) =& p_E^H p_D^A \ln \frac{p_E^H p_D^A}{p_E^H p_D(\theta)} + (1-p_E^H)(1-p_D^A)\ln \frac{(1-p_E^H)(1-p_D^A)}{(1-p_E^H)(1-p_D(\theta))} \\ &+ (1-p_E^H)p_D^A \ln \frac{(1-p_E^H)p_D^A}{(1-p_E^H)p_D(\theta)} + (1-p_D^A)p_E^H \ln \frac{(1-p_D^A)p_E^H}{(1-p_D(\theta))p_E^H}\end{aligned}$$

Since $p_D(\theta_D) = p_D^A$ by definition, $K(d,\theta_D) = 0$ and in particular θ_D is the unique KL minimizer. For d-adoption to be a Berk-Nash equilibrium, we must have $p_D(\theta_D) > p^H$. However, this implies $\theta_D > \theta^H$, i.e., $p_E(\theta_D) > p^H$, and so partial adoption is not optimal under θ_D, a contradiction.

Moving on to the last part, to simplify notation assume WLOG $p_E(\theta) = \theta$. This simply recasts ability as the success rate in the easy task, and allows us to keep track of a single variable. Notice that adoption is optimal iff $\theta > p_E^H$. We then have

$$\begin{aligned}K(\{e,d\},\theta) =& (p^A)^2 \ln \frac{(p^A)^2}{\theta p_D(\theta)} + (1-p^A)^2 \ln \frac{(1-p^A)^2}{(1-\theta)(1-p_D(\theta))} \\ &+ (1-p^A)p^A \ln \frac{(1-p^A)p^A}{(1-\theta)p_D(\theta)} + (1-p^A)p^A \ln \frac{(1-p^A)p^A}{(1-p_D(\theta))\theta}\end{aligned}$$

The first-order condition implies

$$\frac{(1-p_D(\theta))p_D(\theta))(p_E^A-\theta)}{\theta(1-\theta)(p_D^A-p_D(\theta))} = -p_D'(\theta)$$

Which we can also write as

$$p_E^A = \theta + p_D(\theta)\frac{p_D'(\theta)\theta(1-\theta)}{(1-p_D(\theta))p_D(\theta))} - p_D^A \frac{p_D'(\theta)\theta(1-\theta))}{(1-p_D(\theta))p_D(\theta))}$$

or

$$p_E^A = \alpha(\theta) - \beta(\theta)p_D^A$$

Let θ^* be the value solving the equation above for a given pair p^A. Clearly, θ^* is increasing in p_E^A and p_D^A.

[43] Remember that we explicitly assumed that this value exist.

Notice that since it must be the case that

$$p_E^H = \alpha(\theta^H) - \beta(\theta^H) p_D^H$$

Therefore θ^H defines a linear function, or a hyperplane, L, such that for all points (p_E, p_D) that lie below L, $\theta^* < \theta^H$, and for all points that lie above L, $\theta^* > \theta^H$. Notice that full adoption is indeed a Berk-Nash equilibrium iff $\theta^* > \theta^H$, i.e., if the KL-minimizing θ is strictly higher than the human ability level θ^H. This completes the proof. □

Theorem 3. First, by Theorem 1, for $p = (0,0)$ no adoption is the unique Berk-Nash equilibrium. Using the same arguments from the proof of Theorem 2, there is a range where for any $p \in \mathcal{P}^*(\tau)$, no-adoption is the unique Berk-Nash equilibrium. Similarly, for $p = (1,1)$, full adoption is a Berk-Nash equilibrium, and again using similar arguments there is a range where full adoption is an equilibrium for any $p \in \mathcal{P}^*(\tau)$.

Next, if for some τ there exists $p \in \mathcal{P}(\tau)$ that lies above the hyperplane L, then by monotonicity we must have $p \in \mathcal{P}(\tau')$ for all $\tau' > \tau$. This proves the existence of $\tilde{\tau}$.

We now need to show that $\tilde{\tau} \in (\tau_1, \tau_2]$. Using the notation the result from the proof of Theorem 2, we must have either $\tau_E < \tau_D$ or $\tau_E > \tau_D$. Assume WLOG the latter. Then $\tau_D = \tau_1$ i.e., the first period where no adoption is dominated by d-adoption. Notice that we must have $\overline{p}_E(\tau_2) < p_E^H$ and therefore all points in $\mathcal{P}^*(\tau_2)$ (the frontier at τ_2) must lie strictly below the hyperplane L, which implies $\tilde{\tau} > \tau_1$.

Next, assume $\tilde{\tau} > \tau_2$. Notice that $\mathcal{P}(\tau_2)$ includes points that dominate p^H, and in particular, above the hyperplane L. However, by definition $\tilde{\tau}$ is the first point where $\mathcal{P}^*(\tau)$ crosses the hyperplane, a contradiction.

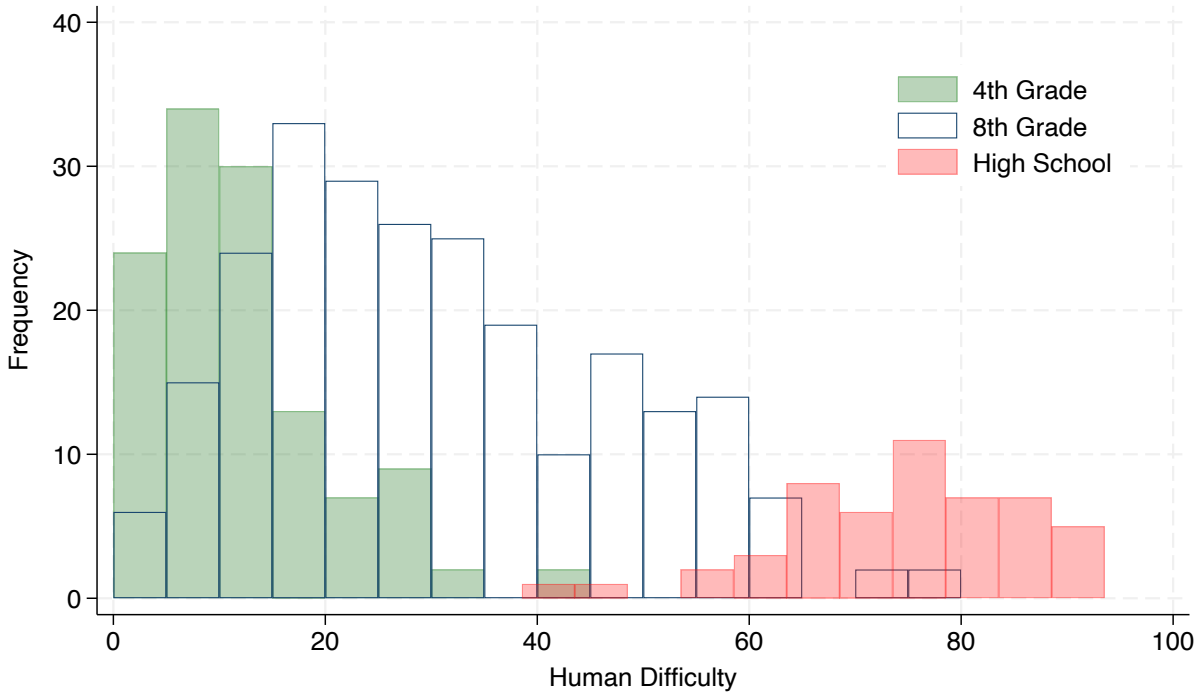
□

B Domain of Mathematical Tasks

B.1 Details on Task Dataset

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. We manually re-transcribed all items into the same format, while trying as much as possible to preserve wording, tables, and symbols used. We excluded items with components which could not be processed in textual form by ChatGPT, e.g., with visual charts or geometric shapes. 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.

Figure 13: Overlaid Histograms (By Grade Level) of Item Difficulty



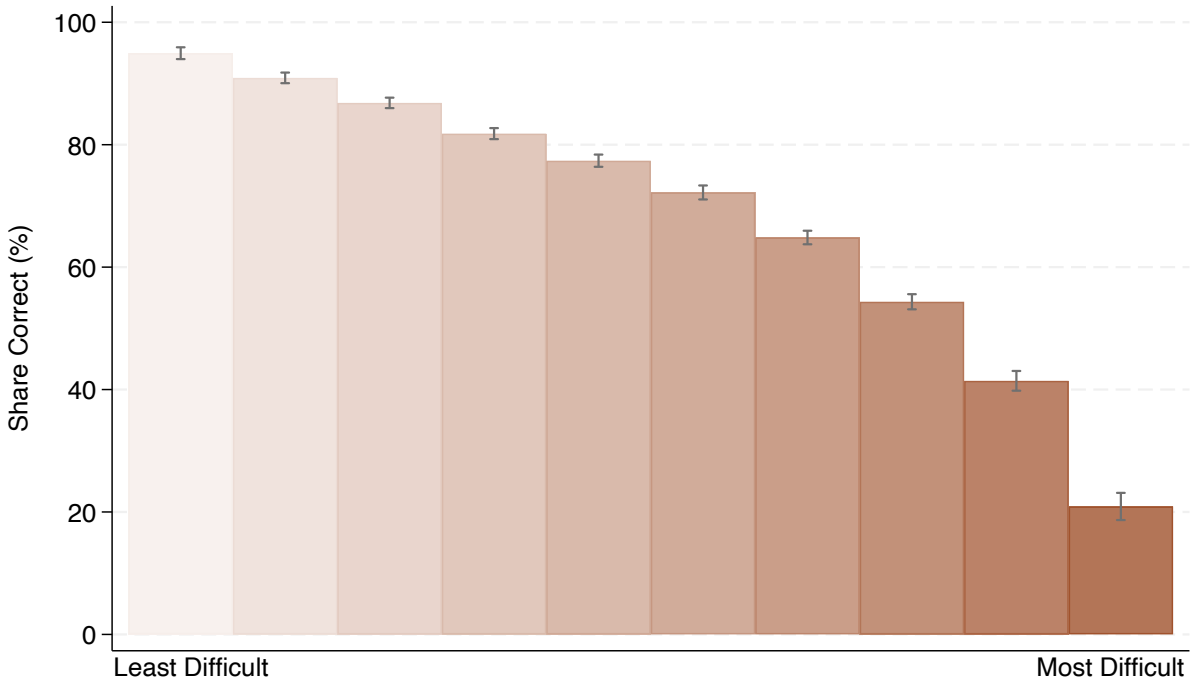
Notes: The figure plots three separate histograms, one for each grade level, of item difficulty. Frequency (number of questions) on the y-axis, and human difficulty on the x-axis. Those histograms are overlaid; for a total histogram see Figure 15 in Appendix.

B.2 Human Performance

Descriptive Results. 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 94. A total of 29 items have a difficulty level of more than 75, which is more than random guessing would imply. This is consistent with the TIMSS item guidelines, which encourage the use of plausible incorrect response options (called "distractors") when designing multiple-choice items.[44] The full distribution of task difficulty is presented in Figure 15. Figure 14 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%. Figure 13 shows the origin of questions for each difficulty level. As one would expect, most of the easier items are from either 4^{th}-grade or 8^{th}-grade tests, while most of the harder ones are of High School level.

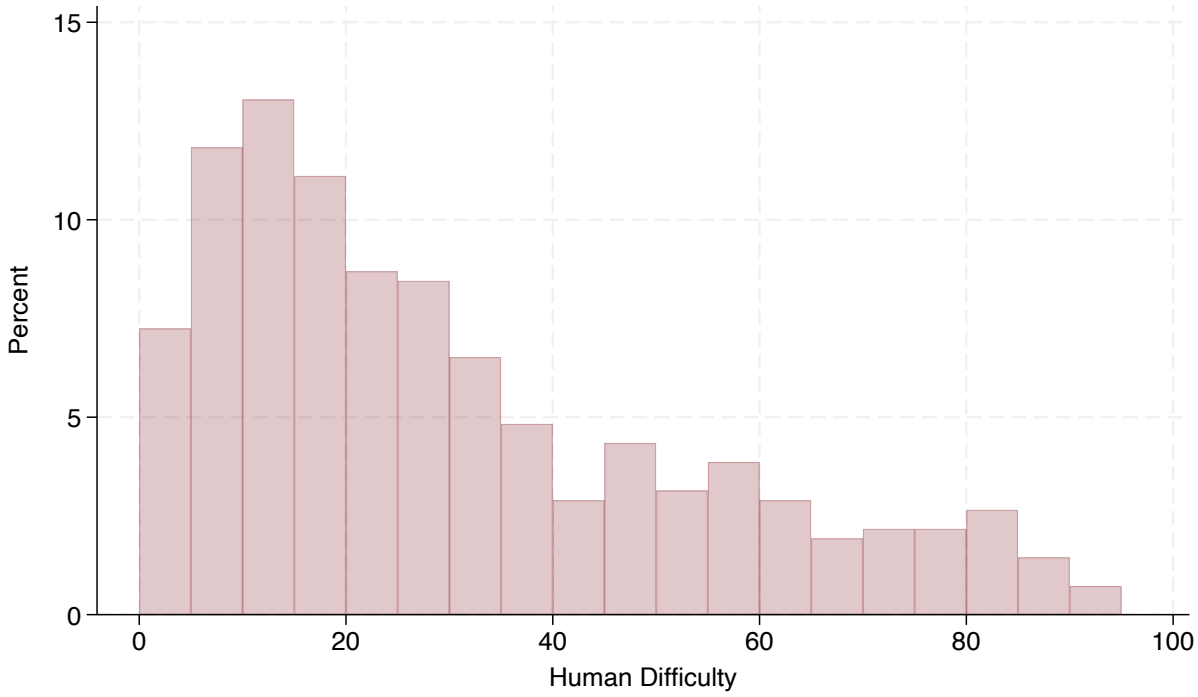
[44] See "Plausibility of Distractors" section in guidelines for the TIMSS 2019 edition.

Figure 14: Average Human Performance by Difficulty Deciles



Notes: The figure plots the average share of subjects answering items correctly within each decile of human difficulty. Each decile, constructed with our measure of item difficulty, is composed of around 41 items.

Figure 15: Histogram of All Items Over Difficulty



Notes: Histogram of all available items (414), with the share (%) of all items plotted on the y-axis, and human difficulty (defined as share of incorrect answers) on the x-axis.

Table 5: Correlation of AI Performance With Item Difficulty

	Dep. Var: AI Performance		
	ChatGPT	**GPT-4**	**Bard**
Difficulty	-0.001	-0.001*	-0.003***
	(0.001)	(0.001)	(0.001)
Constant	0.849***	0.980***	0.743***
	(0.031)	(0.015)	(0.037)
R^2	0.002	0.009	0.023
Observations	414	414	414

Notes: The table presents results from OLS regression of binary AI performance on the human difficulty level of the item. Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B.3 Examples of Tasks

Table 6: Example 4^{th} grade Items

In Toshi's class there are twice as many girls as boys. There are 8 boys in the class. What is the total number of boys and girls in the class? A. 12 B. 16 C. 20 D. 24
There are 9 boxes of pencils. Each box has 125 pencils. What is the total number of pencils? A. 1025 B. 1100 C. 1125 D. 1220 E. 1225
Which of these has the same value as 342? A. $3000 + 400 + 2$ B. $300 + 40 + 2$ C. $30 + 4 + 2$ D. $3 + 4 + 2$
It takes Chris 4 minutes to wash a window. He wants to know how many minutes it will take him to wash 8 windows at this rate. He should: A. multiply 4×8 B. divide 8 by 4 C. subtract 4 from 8 D. add 8 and 4

This chart shows temperature readings made at different times on four days.

	6 a.m.	9 a.m.	Noon	3 p.m.	8 p.m.
Monday	15°C	17°C	20°C	21°C	19°C
Tuesday	15°C	15°C	15°C	10°C	9°C
Wednesday	8°C	15°C	14°C	13°C	15°C
Thursday	8°C	11°C	14°C	17°C	20°C

When was the highest temperature recorded?

A. Noon on Monday
B. 3 p.m. on Monday
C. Noon on Tuesday
D. 3 p.m. on Wednesday

Table 7: Example 8th grade Items

A bowl contains 36 colored beads all of the same size: some blue, some red, some green, and the rest yellow. A bead is drawn from the bowl without looking. The probability that it is blue is $\frac{4}{9}$*. How many blue beads are in the bowl?*

A. 4
B. 8
C. 16
D. 18
E. 20

A circular pond has a radius of 10 meters. There is an average of 2 frogs per square meter in the pond. Approximately how many frogs are in the pond? π *is approximately 3.14.*

A. 120
B. 300
C. 600
D. 2400

In Zedland the original price of a coat was 120 zeds. During a sale the price of the coat was 84 zeds. By what percentage was the price of the coat reduced?

A. 25
B. 30
C. 35
D. 36

There were m boys and n girls in a parade. Each person carried 2 balloons. Which of these expressions represents the total number of balloons that were carried in the parade?

A. $2(m + n)$
B. $2 + (m + n)$
C. $2m + n$
D. $m + 2n$

Look at this table:

4^1	4^2	4^3	4^4	4^5	4^6
4	16	64	256	1,024	4,096

Use the table to express the value of $256 \times 4,096$ *as a power of 4.*

A. 4^{10}
B. 4^{16}
C. 4^{20}
D. 4^{24}

Table 8: Example High School Items

How many points with integer coordinates are there on the graph of the function $y = \frac{12}{x}, x > 0$ *?*

A. 2
B. 4
C. 6
D. infinitely many

What is $\int e^{1+4x} dx$ *?*

A. $\frac{1}{4}e^{1+4x} + C$
B. $e^{1+4x} + C$
C. $4e^{1+4x} + C$
D. $e^{x+2x^2} + C$

If $x = -1 + \frac{1}{2}i$, *which of the following is equal to* $\frac{5}{x}$ *?*

A. $-5 + i$
B. $-4 - 2i$
C. $-4 + 2i$
D. $4 + 2i$

Which of the following lines is perpendicular to the line $6x + 2y = 4$ *and goes through the point* $(-6,5)$ *?*

A. $3x - y = -23$
B. $3x - 7 = 13$
C. $3x - 9y = 9$
D. $x - 3y = -7$
E. $x - 3y = -21$

What is the sum of this geometric series?

$$\left(-\frac{1}{2}\right)^0 + \left(-\frac{1}{2}\right)^1 + \left(-\frac{1}{2}\right)^2 + \ldots$$

A. 2
B. $\frac{3}{2}$
C. $\frac{2}{3}$
D. $-\frac{1}{3}$

C Beliefs Experiment

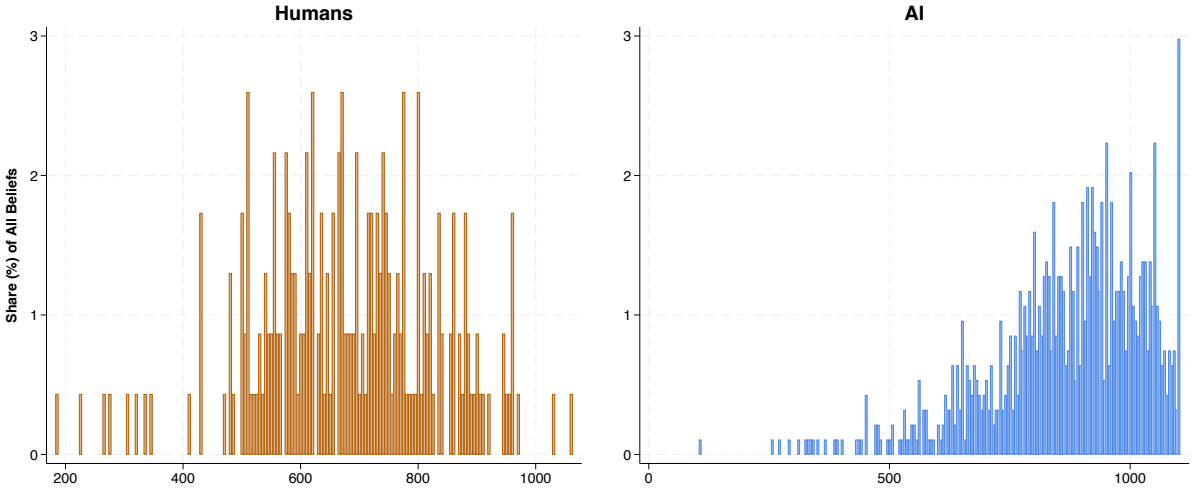
C.1 Additional Results

Table 9: Demographics of Participants in Beliefs Experiment

	Treatment	
Share (%)	*Humans*	*AI*
Gender		
Female	46.8	50.1
Male	50.5	48.5
Race		
White	73.4	78.6
Black	13.8	12.8
Highest Education		
HS or more	98.9	99.6
College or more	62.8	58.9
Highest Math Class Taken		
High School or more	98.9	99.6
College or more	46.3	47.3
Graduate level or more	1.1	2.7
AI Familiarity		
Only heard of AI	n.a.	98.3
Some understanding, but never used	n.a.	64.0
Interacted with AI in some capacity	n.a.	24.0
Mean Age	40.9	42.9
Observations	231	809

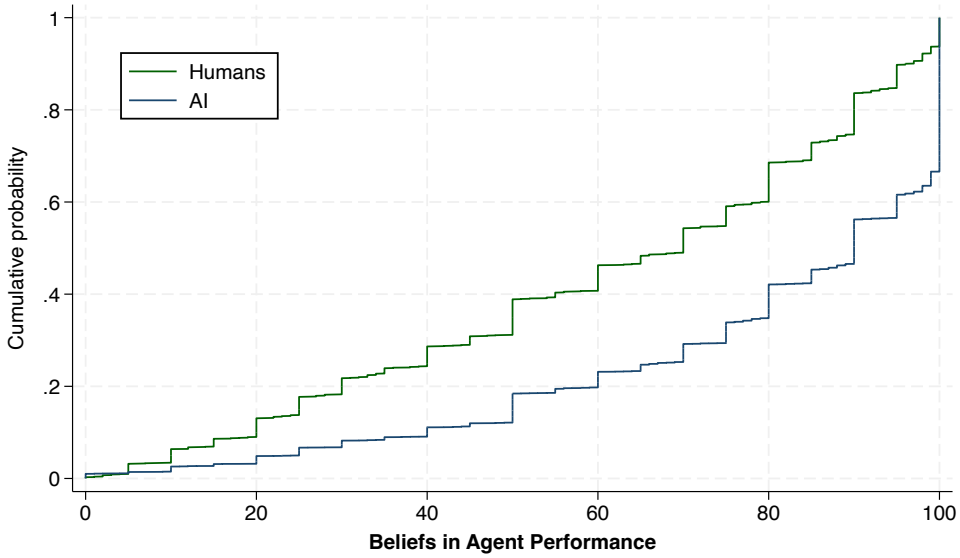
Notes: This table provides demographics of participants in the beliefs experiment, run in October 2023. The AI familiarity question was only asked in the *AI* treatment. The sample is for belief updating results: it excludes subjects failing comprehension checks and the updating comprehension question.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

Figure 16: Histograms of Subjects' Sum of Beliefs in Performance



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

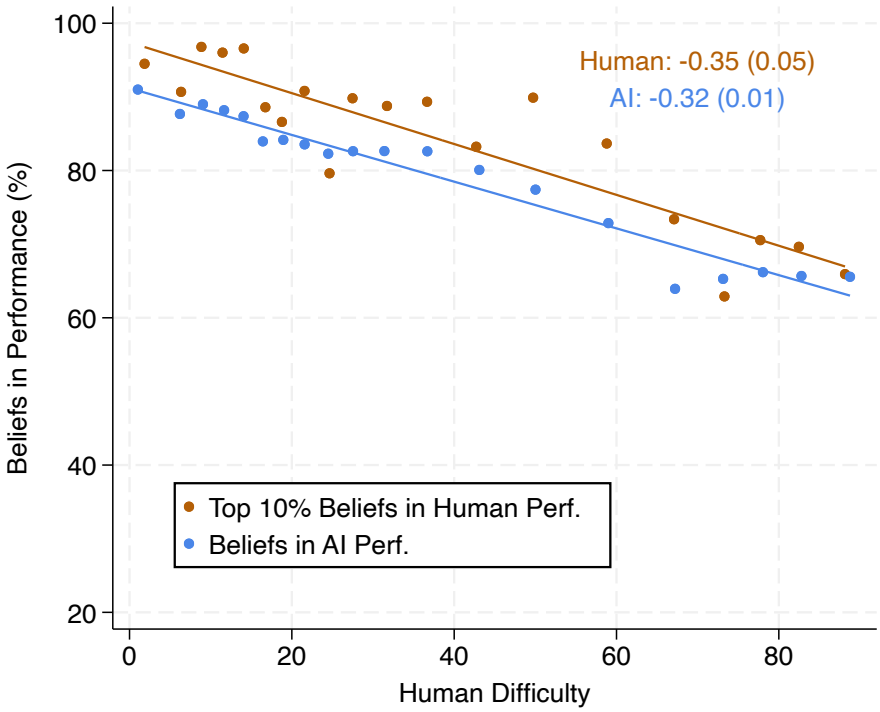
Figure 17: CDF of Prior Beliefs in Performance



Notes: The figure plots cumulative distribution functions of prior beliefs elicited in the second part of the experiment. On x-axis is the belief in %.

Controlling for ability priors. In Figure 18 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. The Figure thus compares subjects that are the most optimistic when predicting humans, with all subjects predicting AI. Figure 19 presents an alternative exercise, matching sub-samples based

Figure 18: 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.

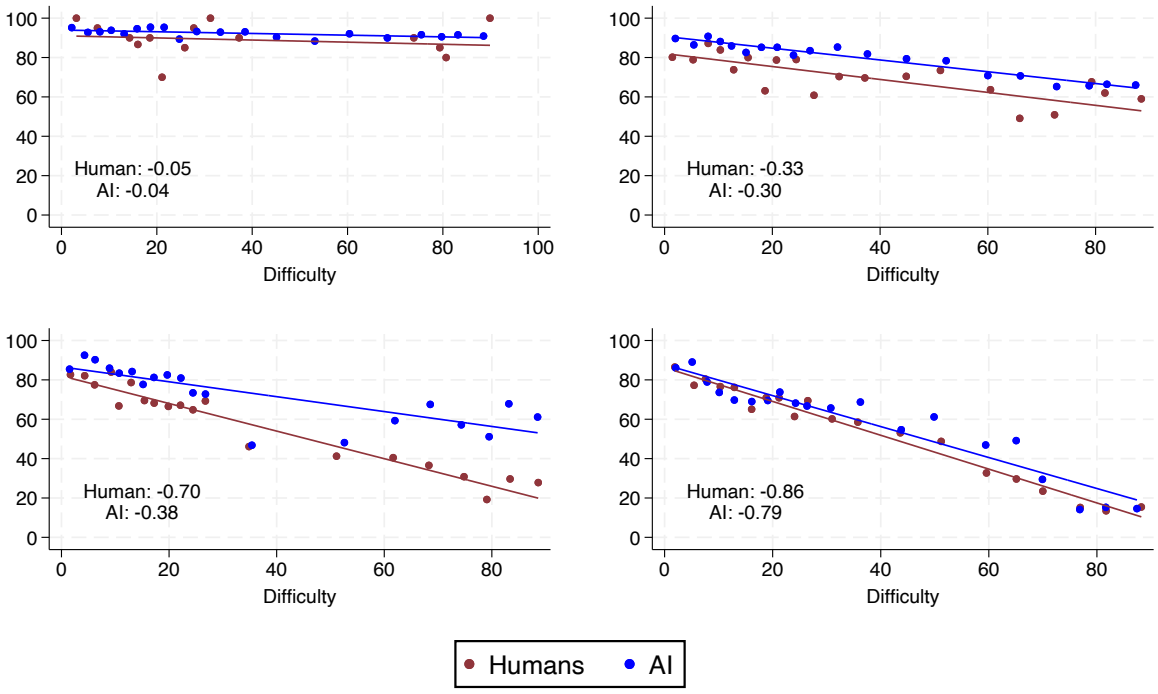
on average predicted performance in the two highest deciles of task difficulty. The top left panel replicates Figure 4 for all subjects who report an average belief of 75% to 100% on tasks within the top two deciles of difficulty. We do the same for average beliefs of 50% to 75% (top right panel), 25% to 50% (bottom left), and 0% to 25% (bottom right). Overall, the figure further suggests that fixing the prior mean ability, subjects in *Human* and *AI* seem to have a similar slope with respect to human difficulty. This is again consistent with the idea that the difference between *Human* and *AI* noted in Figure 4 is partly driven by different priors on ability. Slopes are most similar for the more extreme beliefs (below 25% and above 75%), since those are more indicative of an extreme (high or low) prior on ability.

Belief Accuracy. To complement results on AI performance presented in Section 3, we can quantitatively assess the accuracy of subjects' beliefs by regressing actual performance on elicited beliefs. We estimate the following specification:

$$Y_t = c_0 + c_1 \text{Beliefs}_{it} + d_i + \epsilon_{it}, \tag{5}$$

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).

Figure 19: Matching Beliefs on Top Quintile of Difficulty - Highest Average Beliefs (Top Left) to Lowest (Bottom Right)



Notes: This figure compares subsets of subjects in both treatments who report similar beliefs for the top two deciles of difficulty. Top left: take all subjects who report an average belief of 75% or more on tasks belonging to the top two deciles of difficulty, and plot all their beliefs. Top right is average belief between 50%-75%, bottom left is 25%-50%, bottom right is 0%-25%.

However, there is stark contrast in predictive power: the R^2 is 0.391 for humans, while only 0.002 for AI. As subjects seem to strongly rely on human difficulty to predict (mostly uncorrelated) AI performance, their answers carry very little predictive power overall.

Table 10: Predictive Power of Beliefs on Actual Performance

	Dep. Var: Actual Performance	
	Humans	**AI**
Belief in Perf	0.584***	0.061***
	(0.014)	(0.015)
R^2	0.391	0.002
Observations	2541	10340

Notes: The table presents OLS coefficients. The dependent variable for humans is the share (%) answering the item correctly, while for AI it is binary performance. Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Belief Updating - Effect of Signal Difficulty on Beliefs

	Dep. Var: Belief Movement			
	Humans		**AI**	
	Success	**Failure**	**Success**	**Failure**
	(1)	(2)	(3)	(4)
Prior	-0.281***	-0.305***	-0.532***	-0.243***
	(0.080)	(0.086)	(0.039)	(0.033)
Task Difficulty	0.211***	0.145***	0.190***	0.073**
	(0.050)	(0.044)	(0.022)	(0.030)
Controls	Yes	Yes	Yes	Yes
Prediction Task FE	Yes	Yes	Yes	Yes
R^2	0.305	0.309	0.505	0.153
Observations	117	114	386	425

Notes: Belief movement is the difference between posterior and prior (positive for successes and negative for failures). The main independent variable is signal task difficulty. Controls include socio-demographic variables, and familiarity with AI (only in *AI*). Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C.2 Beliefs Experiment Instructions

Figure 20: Instructions Screen - Initial Test

Thank you for participating in our survey! You'll be solving 30 math problems of varying difficulty. Here's what to expect regarding the types of questions:

- **Multiple Choice:** Some questions are multiple-choice, with answers labeled as A, B, C, D, etc.
- **Open-ended:** For these, simply provide the final answer. If an answer format is specified, please follow it.
- **Show Your Work:** If a question specifies this, share the main steps or equations leading to your answer. A detailed write-up isn't required.

On the next screen, you'll begin the first of three blocks, each containing 10 questions (a total of 30 questions).

Taking the Test:

- Approach this as you would a real-life school math test.
- Answer to the best of your ability, but be time-conscious. Some questions might need more thought, but don't spend too long on any single one. If unsure after a reasonable attempt, give your best guess and proceed.
- Please **do not seek outside help** during the test.

Compensation: For every correctly answered question, you'll earn a bonus of **5 cents**.

Feedback: Once our data collection is finished in 1-2 weeks, we'll:

- Inform you about how your performance compares to others.
- Deliver your earned bonus.

Stay tuned and good luck!

Figure 21: Instructions Screen - AI Treatment

We will now describe the main task in more detail. On the next screens you'll see 10 math questions: these have been randomly sampled from a pool of several hundred questions taken from different standardized math tests.

Instructions:

Solving the Math Questions:

- Approach this as you would a real-life school math test.
- All questions are **multiple-choice**, with only one correct answer.
- Answer to the best of your ability, but be time-conscious. Some questions might need more thought, but don't spend too long on any single one. If unsure after a reasonable attempt, give your best guess and proceed.
- Please **do not seek outside help** (you may use a pen and paper).

Predicting ChatGPT's success:

- Each question you see has also been presented to ChatGPT (with minimal formatting changes to help process tables and symbols).
- After each question, you'll be asked: "What do you think is the % chance ChatGPT answered correctly?". Consider the likelihood that ChatGPT correctly answered the question and provide a percentage (from 0 to 100) to represent this belief.

Two Ways to Earn Bonus Payments:

- **Solving:** For each math question you correctly solve, you will earn **5 cents**.
- **Predicting:** for every prediction you can earn up to **10 cents**. Your earnings depend both on your answer and on the actual performance of ChatGPT. If ChatGPT answered the question correctly, you have a better chance of earning the reward if you gave a higher % chance . Conversely, if ChatGPT answered incorrectly, you have a better chance of earning the reward if you gave a lower % chance. The bottom line: Your potential reward will be maximized if you report what you genuinely believe is the % chance that ChatGPT answered the question correctly.

You can click "Show Details" to understand details of our payment rule regarding your predictions. **Show Details**

Feedback on Your Test Performance: Stay Tuned!

- Once data collection is completed, in 1-2 weeks, we will inform you about how your test performance compares to other participants
- We will deliver your total bonus payment (up to **$1.8**) at the same time

Figure 22: Prior Beliefs Screen - AI Treatment

Which number is five hundred four and seven-tenth?

A. 54.7
B. 504.7
C. 547
D. 5004.7

	Your answer A B C D	**What do you think is the % chance ChatGPT answered correctly?** Enter a number between 0 and 100
This question:	○ ○ ○ ○	

Which of the following is true when a, b, and c are different real numbers?

A. $a - b = b - a$
B. $a(b - c) = b(c - a)$
C. $b - c = c - b$
D. $ab = ba$
E. $ab - c = ac - b$

	Your answer A B C D E	**What do you think is the % chance ChatGPT answered correctly?** Enter a number between 0 and 100
This question:	○ ○ ○ ○ ○	

Figure 23: Belief Updating Screen - Revealing Performance (Easy Fail)

Reference Question:

Which is the largest fraction?

A. $\frac{1}{2}$

B. $\frac{1}{3}$

C. $\frac{1}{4}$

D. $\frac{1}{5}$

ChatGPT's Answer:

To determine the largest fraction among the options, we need to compare their values. The smaller the denominator, the larger the fraction.

Among the options:

A. 1/2
B. 1/3
C. 1/4
D. 1/5

The largest fraction is:

B. 1/3

Since 3 is the smallest denominator among the options, 1/3 is the largest fraction.

Evaluation: This answer is **Incorrect** (the correct answer is "A. 1/2").

Figure 24: Belief Updating Screen - AI Treatment

Prediction Question:

If $x > 0, y > 0$, *and* $x \neq y$ *then* $\frac{1}{\sqrt{x}-\sqrt{y}}$ *is equal to:*

A. $\frac{\sqrt{x}+\sqrt{y}}{x-y}$

B. $\frac{\sqrt{x}-\sqrt{y}}{x-y}$

C. $\frac{1}{x-y}$

D. $\frac{1}{\sqrt{x}} - \frac{1}{\sqrt{y}}$

E. $\frac{\sqrt{x}+\sqrt{y}}{x^2-y^2}$

Given what you saw, what do you think is the % chance that **ChatGPT** answered this question correctly? (Enter a number between 0 and 100)

D Medium-Run Adoption

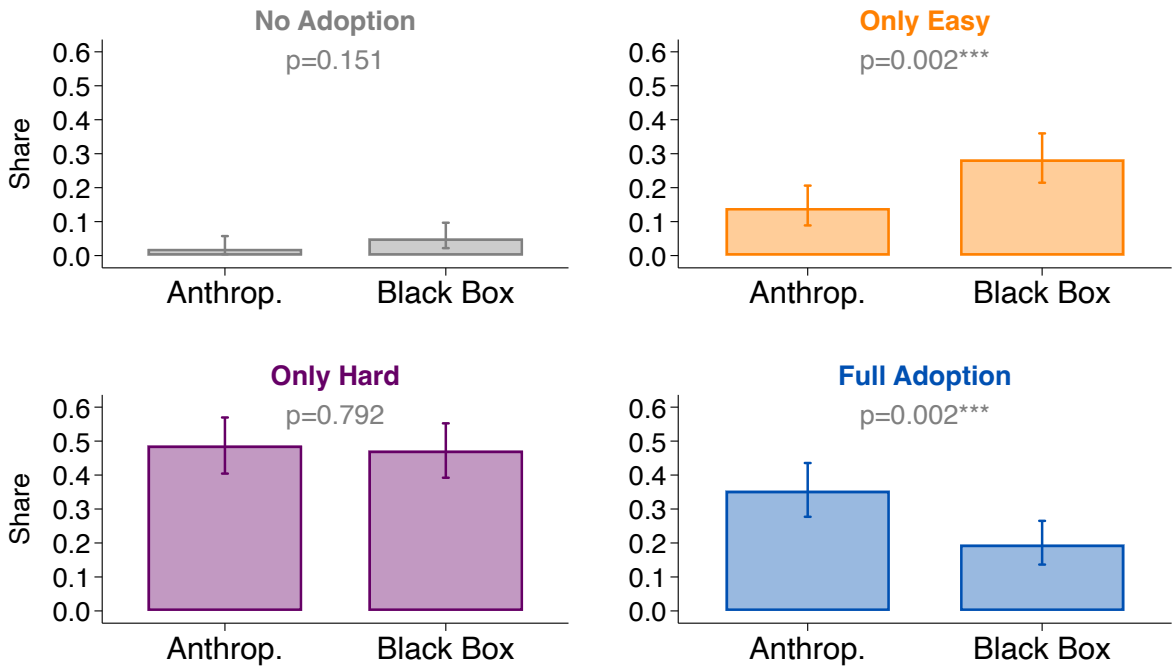
D.1 Additional Results

Table 12: Demographics of Participants in Adoption Experiment

	Treatment	
Share (%)	*Black Box*	*Anthropomorphic*
Gender		
Female	54.2	48.3
Male	44.1	51.7
Race		
White	64.4	74.1
Black	32.2	22.4
Highest Education		
HS or more	98.3	100.0
College or more	64.4	53.4
Highest Math Class Taken		
High School or more	98.3	100.0
College or more	64.4	46.6
Graduate level or more	3.4	3.4
AI Familiarity		
Heard of AI	100.0	100.0
Some understanding, but never used	78.0	79.3
Interacted with AI in some capacity	45.8	39.7
Mean Age	37.7	39.1
Observations	59	58

Notes: This table provides demographics of participants in the adoption experiment. Averages are consistent with overall demographics of Prolific participants.

Figure 25: Shares of Adoption by Treatment



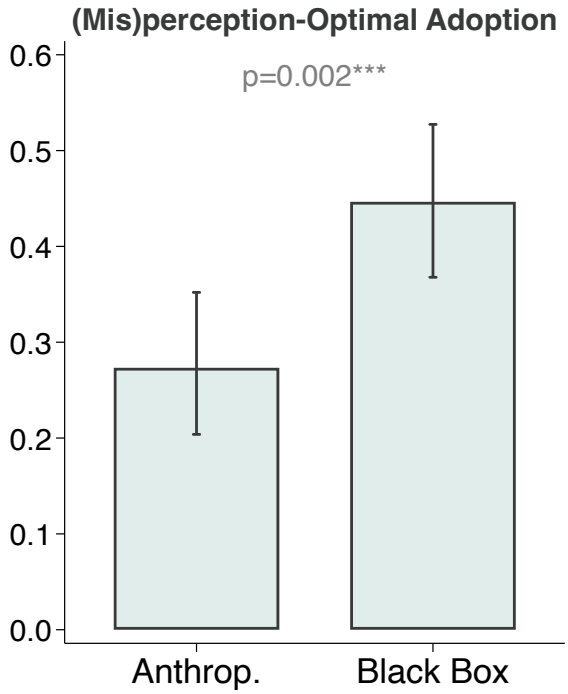
Notes: The figure presents the shares of each type of adoption, for the Black Box and Anthropomorphic treatments, for the first sample collected (which includes misperceptions of the human difficulty of tasks, see Figure 26). Sample sizes are $n = 150$ for Anthropomorphic and $n = 159$ for Black Box. Confidence intervals at the 95% level are included, along with p-values for two-sided tests of proportions.

Table 13: Correlations between Beliefs

	Dep. Var: Beliefs about AI on Hard Tasks					
	Anthropomorphic			**Black Box**		
	Prior	**Interim**	**Posterior**	**Prior**	**Interim**	**Posterior**
Prior AI (Easy)	0.449***			0.151**		
	(0.132)			(0.070)		
Inter. AI (Easy)		0.346***			0.137*	
		(0.089)			(0.074)	
Post. AI (Easy)			0.258***			0.068
			(0.078)			(0.081)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.302	0.180	0.116	0.079	0.100	0.082
Observations	147	147	147	157	157	157

Notes: This table reports OLS estimates, where the dependent variable is reported belief (0-100% scale) in AI performance on hard tasks (prior, interim and posterior), and independent variable is the corresponding belief in AI performance on easy tasks. Controls include age, gender, education, income, and AI familiarity.

Figure 26: (Mis)perception-Optimal Adoption



Notes: The figure presents the shares of subjects choosing the type of partial adoption that is optimal given their perception of the relative human difficulty of tasks. This means choosing Only Hard if the subject perceives difficulty correctly, and Only Easy if incorrectly. perception is assessed on the basis of reported prior beliefs about human performance: those reporting a higher human success rate on hard tasks than on easy tasks are labeled as perceiving incorrectly, and correctly otherwise. 95% confidence intervals, and a p-value for a two-sided test of proportions are included.

D.2 Adoption Survey Details

Initial instructions.

Thank you for taking part in our survey! We study people's perceptions of standardized test questions. On the next page, we will detail the content of the survey and what you will be asked to do. Note that your decisions can earn a bonus payment.

It is important to pay attention throughout this survey: in addition to helping us obtain valuable data for our research, this will help you earn a bonus payment!

Transparency and Integrity: Please note that there is no deception involved in this experiment. Everything we state and all the information you will see is truthful and accurate. Your participation and understanding are important to us, and we are committed to ensuring a transparent experience.

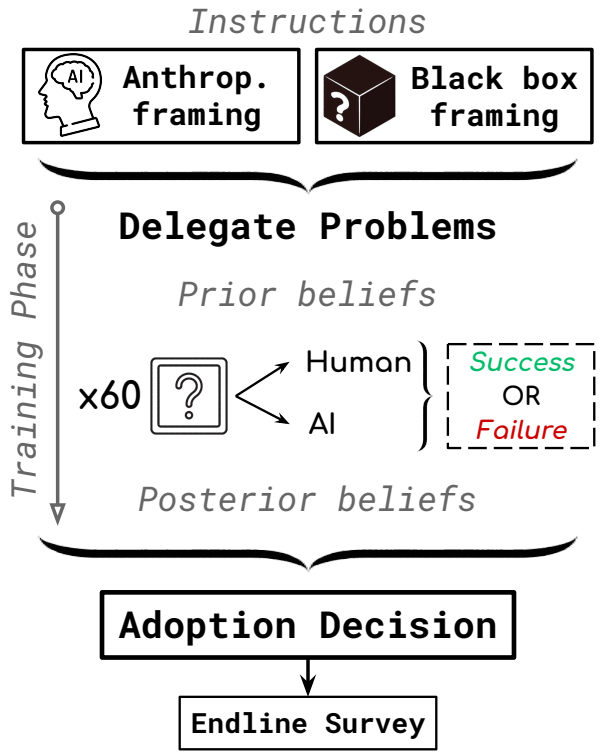
General structure: You'll be presented with a math question and choose whether to delegate it to a Prolific respondent or to a black box. You'll find out if your chosen pick solved it successfully. Each correct solution earns you a 3-cent bonus. You will repeat this process 60 times, each time with a different question.

Table 14: Training Phase Behavior

	Anthropomorphism				Black Box			
	Easy		Hard		Easy		Hard	
Panel A:	*Delegation Choices*							
# given to AI	14.22	(1.21)	22.50	(.72)	9.53	(.9)	21.17	(.63)
# Mistakes Seen	4.66	(.42)	7.36	(.28)	3.14	(.29)	7.46	(.23)
Share Mistakes Seen	31.86	(1.56)	32.16	(.77)	38.19	(2.83)	35.46	(.61)
Panel B:	*Beliefs in AI performance (%)*							
Priors	88.95	(1.88)	78.53	(2.41)	75.20	(2.74)	62.97	(2.71)
Interim	74.05	(2.38)	69.60	(2.31)	63.97	(2.8)	62.63	(2.29)
Posteriors	70.97	(2.62)	70.21	(1.9)	61.12	(3.13)	64.86	(2.33)
Panel C:	*Beliefs in Human performance (%)*							
Priors	74.71	(2.13)	52.72	(2.75)	74.14	(2.28)	49.34	(2.88)
Interim	70.31	(2.45)	49.41	(2.59)	71.20	(2.28)	45.47	(3.14)
Posteriors	71.79	(2.27)	47.67	(2.86)	68.66	(2.94)	41.71	(2.9)

Notes: This table presents subject behavior during the "training phase," which contains a total of 60 problems. There are 30 easy and 30 hard problems, which subjects delegate to either humans or AI. Performance is revealed: either ChatGPT's performance on the problem, or a random draw of a Bernoulli with human success rates. Beliefs are elicited on a 0-100% scale, at the start (priors), middle (interim) and end (posteriors) of the training phase.

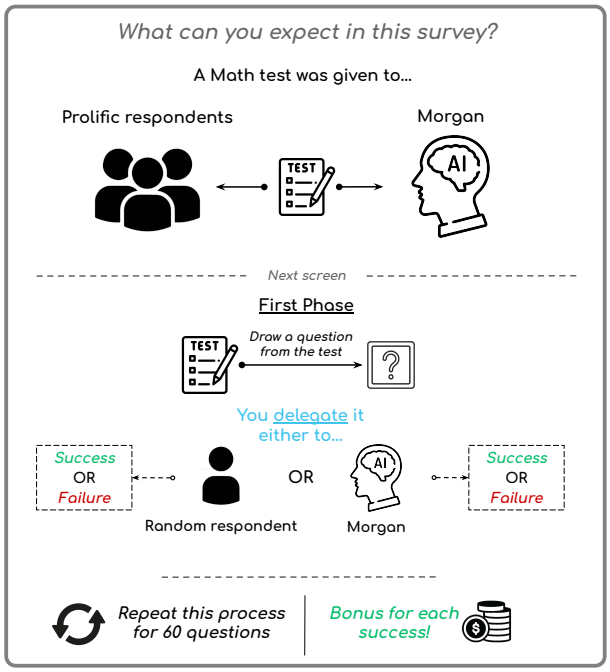
Figure 27: Adoption Experiment Design Flowchart



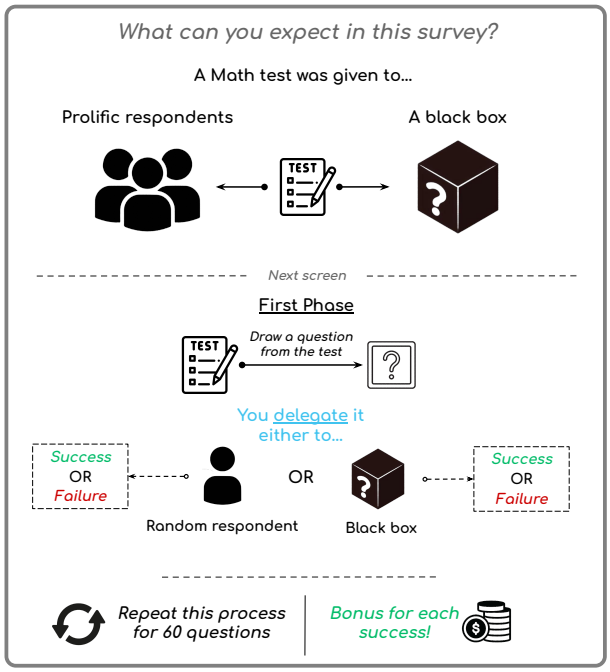
Notes: The treatment variation is contained within instructions. The training phase is composed of 30 problems of each type, which subjects delegate to either humans or AI and see corresponding signals of binary performance. Endline survey elicits basic demographics, as well as prior familiarity with AI and most advanced math class taken.

Figure 28: Initial Instructions Figures

(a) Anthropomorphic Framing



(b) Black Box Framing



Notes: Initial instruction visuals describing the structure of the experiment. Alongside these figures are presented details about the math test, human respondents, and bonus payment rules.

D.2.1 Survey Screenshots

Figure 29: Example Prompt Screen - Anthropomorphic

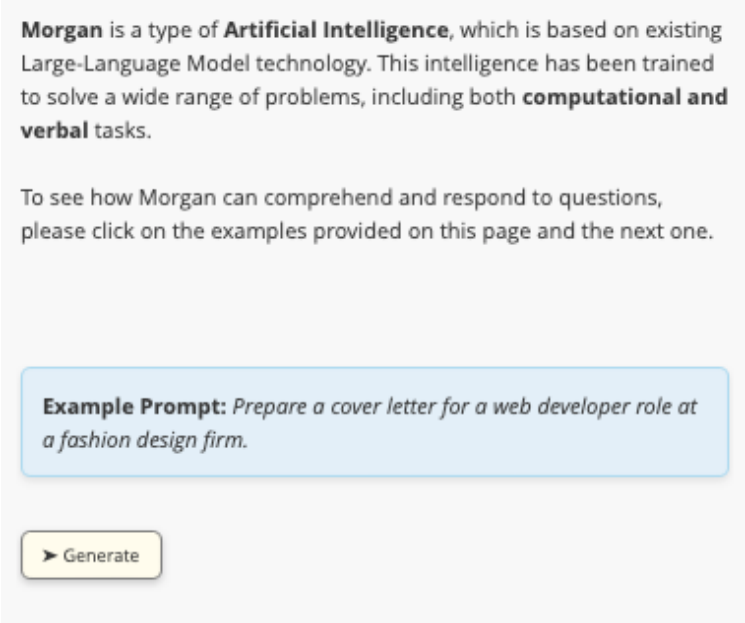


Figure 30: Delegation Screen (Blue Task) - Anthropomorphic

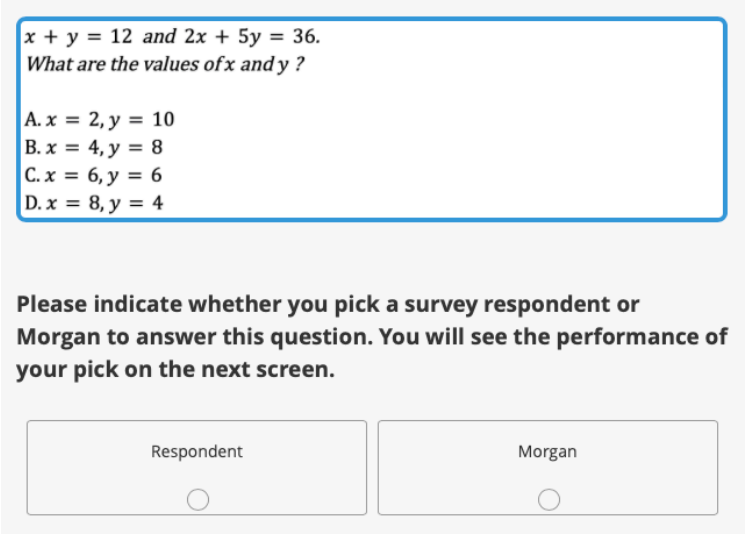


Figure 31: Delegation Screen (Green Task) - Anthropomorphic

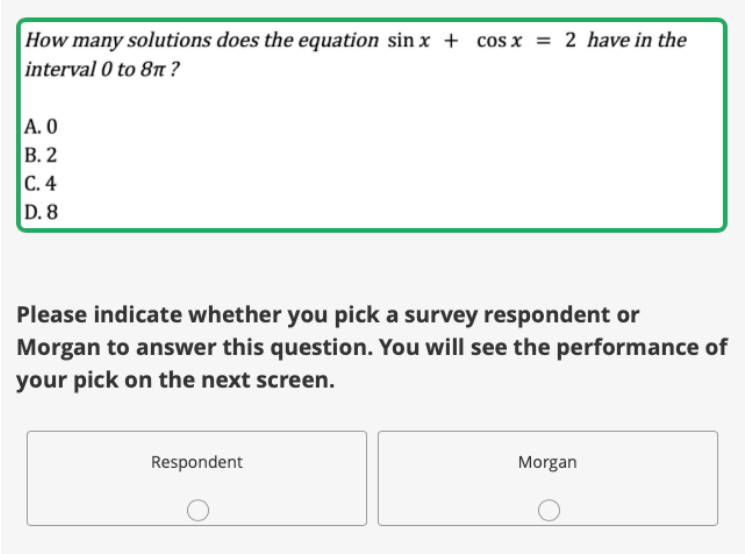


Figure 32: Performance Reveal Screen

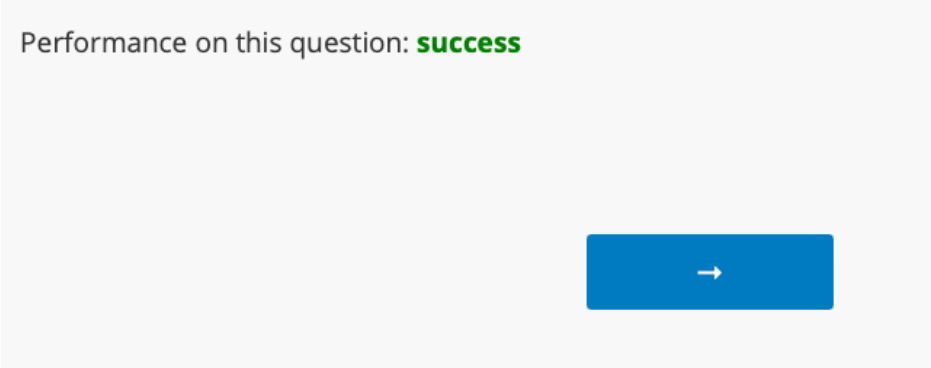
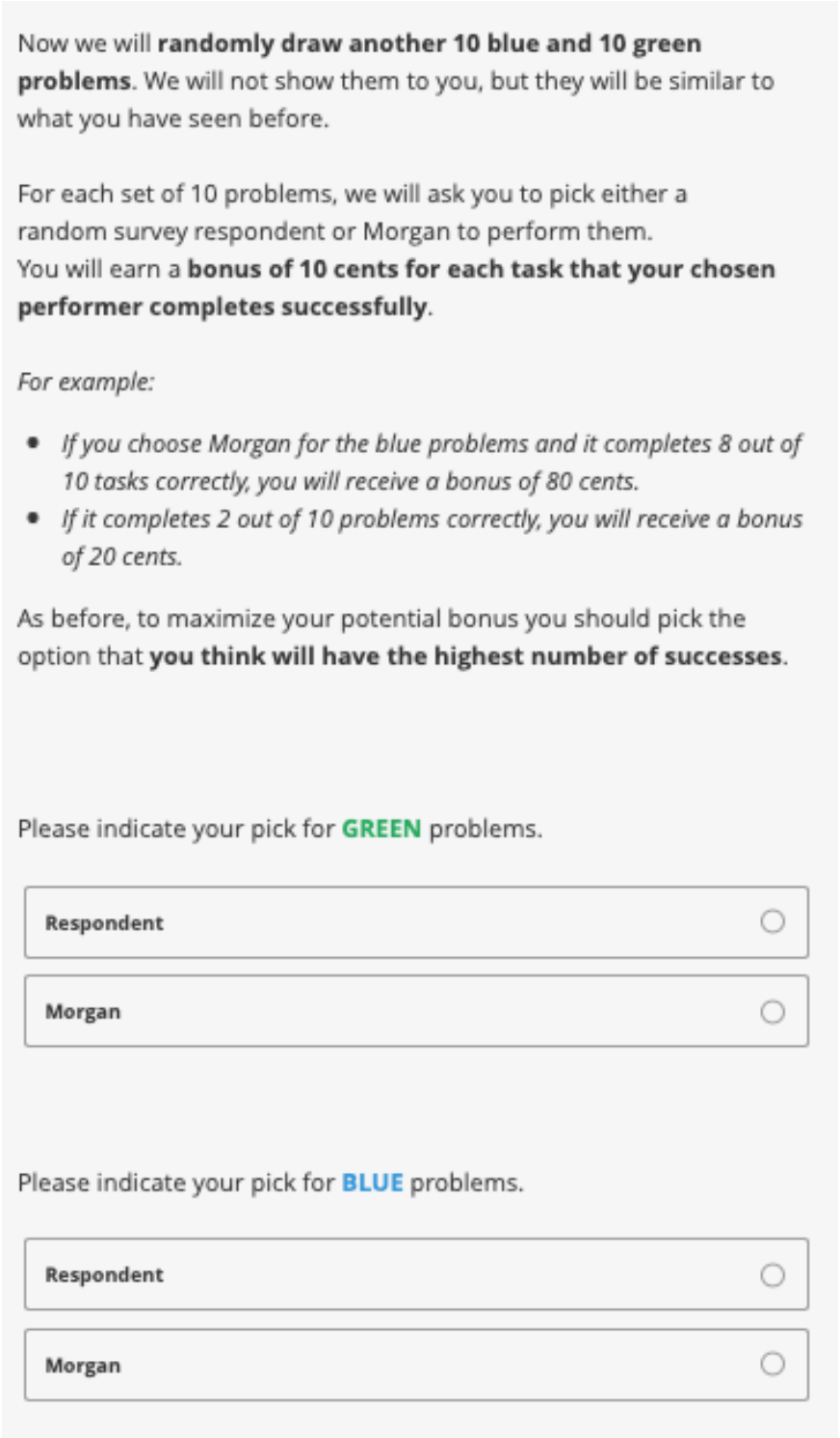


Figure 33: Prior Beliefs Screen - Anthropomorphic



Figure 34: Final Adoption Screen - Anthropomorphic



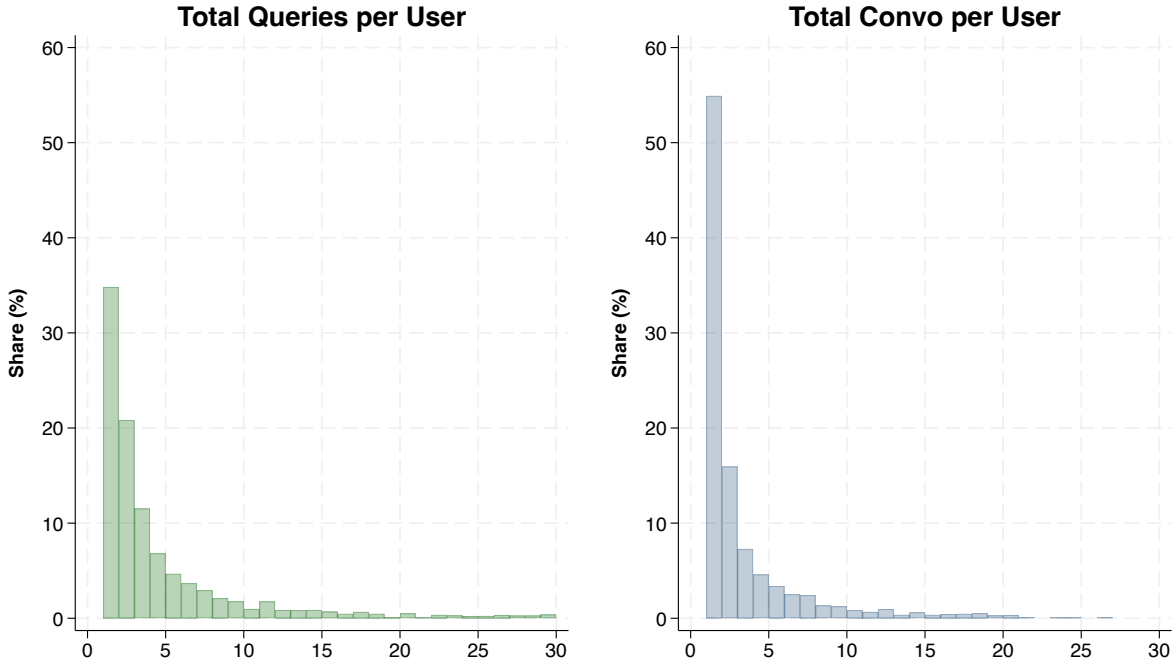
E ParentData

E.1 Descriptive facts

Conversation Data. Over the summer of 2024, the AI (Dewey) had around 7000 unique users on a monthly basis (around 230 daily), asking a total of 25000 queries.We were given access to a snapshot of around 40000 queries asked between December 31st 2023 and April 30th 2024. These queries are asked in 24089 unique conversations, from 16422 unique users.

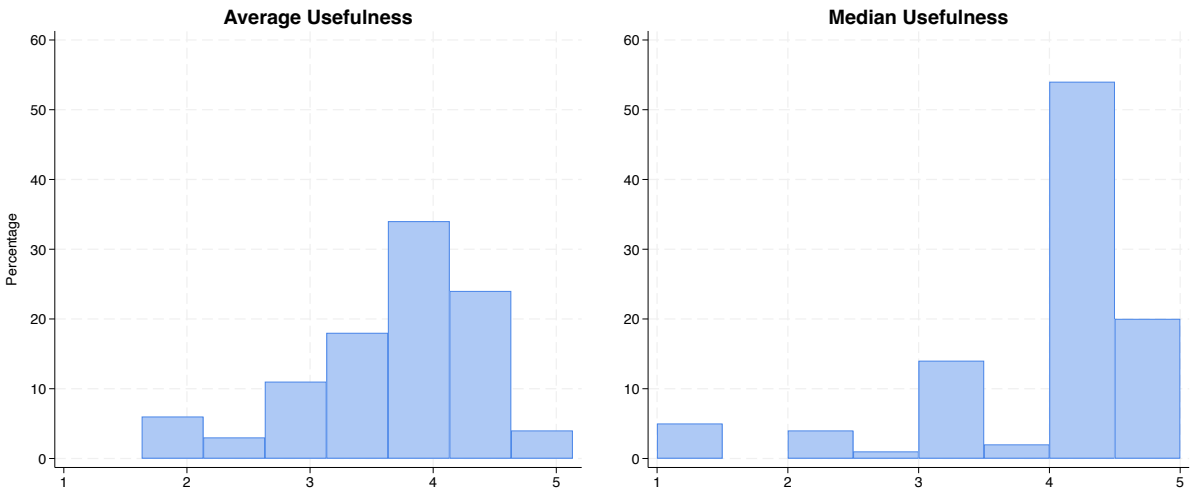
Dewey. Dewey is presented as an "AI librarian," and is free to use. It is a LLM-based chatbot, which has summarized ParentData.org's articles into a series of short answers, which were then human-vetted by the ParentData.org team. Upon receiving a user's query, Dewey matches it to all the premade questions in its database, using a confidence score. This score is a custom approach to a cosine similarity measure using vector embeddings (further details were not disclosed). Dewey ranks premade questions by score, and displays the top match's associated answer if above a certain confidence threshold. Below this threshold, Dewey either displays some question suggestions or a generic message telling the user this question has not been answered yet.

Figure 35: Histograms of queries and conversations per user



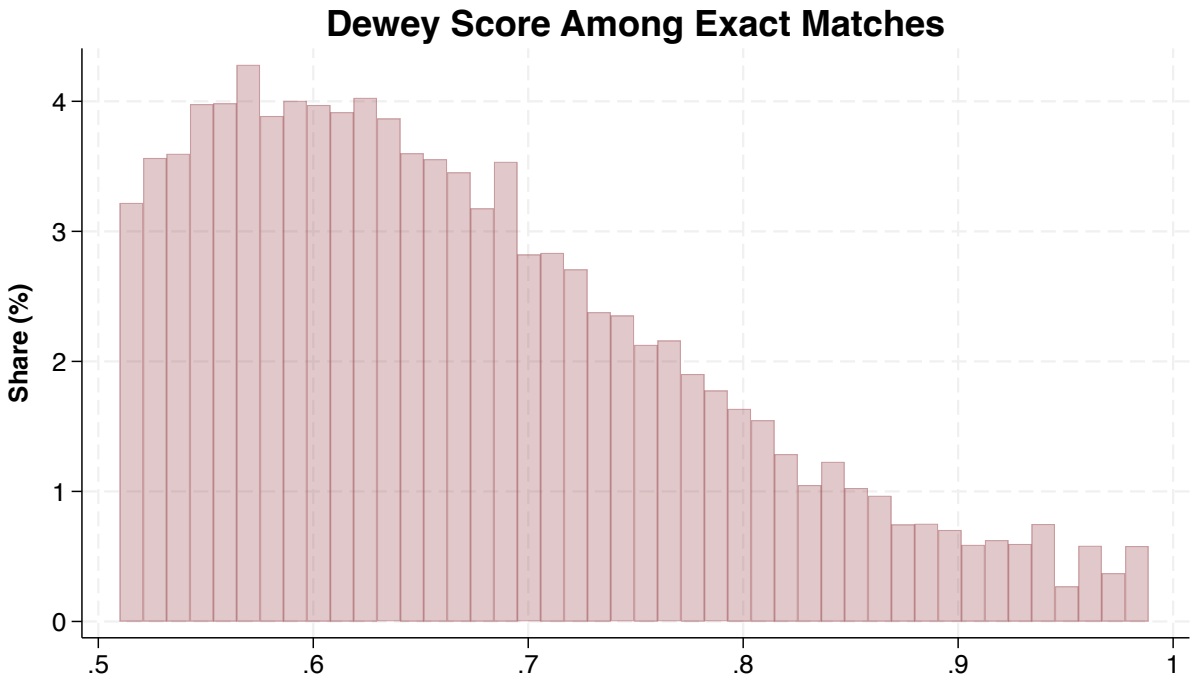
Notes: The figure plots the distribution of the number of queries and conversations per unique user, truncated at 30 for each. A conversation refers to a "session," and can contain multiple queries (unlike the definition of conversation in the main text).

Figure 37: Dewey Performance: Histograms of Answer Usefulness



Notes: The left panel plots the histogram of mean answer usefulness, while the right panel plots the histogram of median usefulness. Usefulness is elicited on a scale of 1-5. On the y-axis are the percentages of each bin.

Figure 36: Histogram of match score



Notes: Upon receiving a user query, Dewey matches it to all questions in its database based on this score, and displays the top-ranked match if above a confidence threshold (of 0.51 to display the answer, and of 0.4 to display suggestions of questions). The figure plots the distribution of AI confidence scores among answers displayed to users, which represents around 80% of the total number of queries in the dataset.

Table 15: Most frequent Dewey matches

Description	Frequency
Is Tylenol safe during pregnancy?	170
Is it OK to drink alcohol while breastfeeding?	169
Is it safe to sleep on your back during pregnancy?	168
Can I eat raw fish, like sushi, during pregnancy?	132
Can I dye my hair while pregnant?	131
How long can breast milk be stored in the fridge?	130
Is it okay to have a glass of wine during pregnancy?	118
Can I use retinol while breastfeeding?	110
Should I get my child's tongue tie released?	105
Can I have deli meats, like ham and turkey, while I'm pregnant?	104
Can I go in a hot tub while pregnant?	97
How much caffeine is ok when breastfeeding?	93
Do I need to sterilize bottles?	88
Is it safe to use white noise machines for my baby's sleep?	88
How long does morning sickness last?	88
How long can I leave a bottle of formula sitting out?	87
Can eating dates help induce labor during pregnancy?	83
Can I get botox while breastfeeding?	81
Can I drink raspberry leaf tea when pregnant?	78
Can I eat prosciutto while pregnant?	75
Is cold medicine safe in pregnancy?	73
Can I take a bath while I'm pregnant?	73
What helps induce labor?	68
How much coffee is safe to drink during pregnancy?	66
Can I eat rare steak while pregnant?	66
Can I take magnesium when pregnant?	65
Is it OK to drink during the two-week wait?	65
Do I need to give my baby vitamin D drops?	63
Can I take melatonin while pregnant?	62
Is sunscreen safe while pregnant?	62
Can breast milk be refrozen after it thaws?	61
Can I sleep on my stomach in the first trimester?	59
Can I eat cheese while pregnant?	57
Does sex induce labor?	57
Can I eat smoked ham when pregnant?	56
Total	3148

Notes: The table presents the most frequent question matches made by Dewey. These are premade questions which were interpreted by Dewey as the best possible match to the user's query. To each premade question corresponds one premade answer, which gets displayed to the user if the match score exceeds a confidence threshold.

We performed a preliminary engagement analysis based on AI confidence score—a proxy for

answer quality. Results (Table 16) are consistent with basic intuitions regarding answer quality and user engagement. The score of the first query asked is strongly negatively correlated with a follow-up query within the same conversation: most users start a conversation with a specific question in mind, and are likely to move on once it has been properly answered. This score is then strongly positively correlated with the total number of conversations created by the user: if they received a good answer on their first use, users are more likely to use the chatbot again.

The absence of any information on users (especially paid subscription status and history of prior engagement) prevents any credible observational analysis of repeated engagement, so we resort to a field experiment as described in the main text.

Table 16: Probability of user return

	(1)	(2)	(3)
	Follow-up query	Start other conv	Start other conv
Score 1st query	-1.006***	0.463***	
	(0.178)	(0.170)	
Avg Score 1st conv			0.420**
			(0.177)
Constant	-0.723***	-1.713***	-1.683***
	(0.120)	(0.117)	(0.122)
Pseudo R^2	0.002	0.000	0.000
Observations	16422	16422	16422

Standard errors in parentheses

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table presents results from OLS estimations. Dependent variables are binary: whether the user asks a follow-up query within the same first conversation (column 1) and whether the same user starts other conversations at a later time (columns 2 and 3). Independent variables are the AI confidence score of the first query, and the mean AI score of queries within the first conversation. Robust standard errors in parentheses.

E.2 Additional Experimental Results

Table 17: Demographics of Participants in Engagement Experiment

	Treatment	
Share (%)	*Unreasonable*	*Reasonable*
Gender		
Female	85.0	83.8
Male	14.1	16.0
Race		
White	72.7	74.7
Black	18.1	18.8
Highest Education		
HS or more	99.8	99.3
College or more	59.3	59.6
Parenting Status		
Trying to conceive	24.2	22.2
Currently expecting	6.8	7.3
Have children below 18	79.1	77.6
AI Familiarity		
Only heard of AI	98.5	97.8
Some understanding, but never used	77.1	76.3
Interacted with AI in some capacity	38.1	36.1
ParentData Familiarity		
Knew of the website	3.1	2.9
Mean Age	34.3	34.0
Observations	454	451

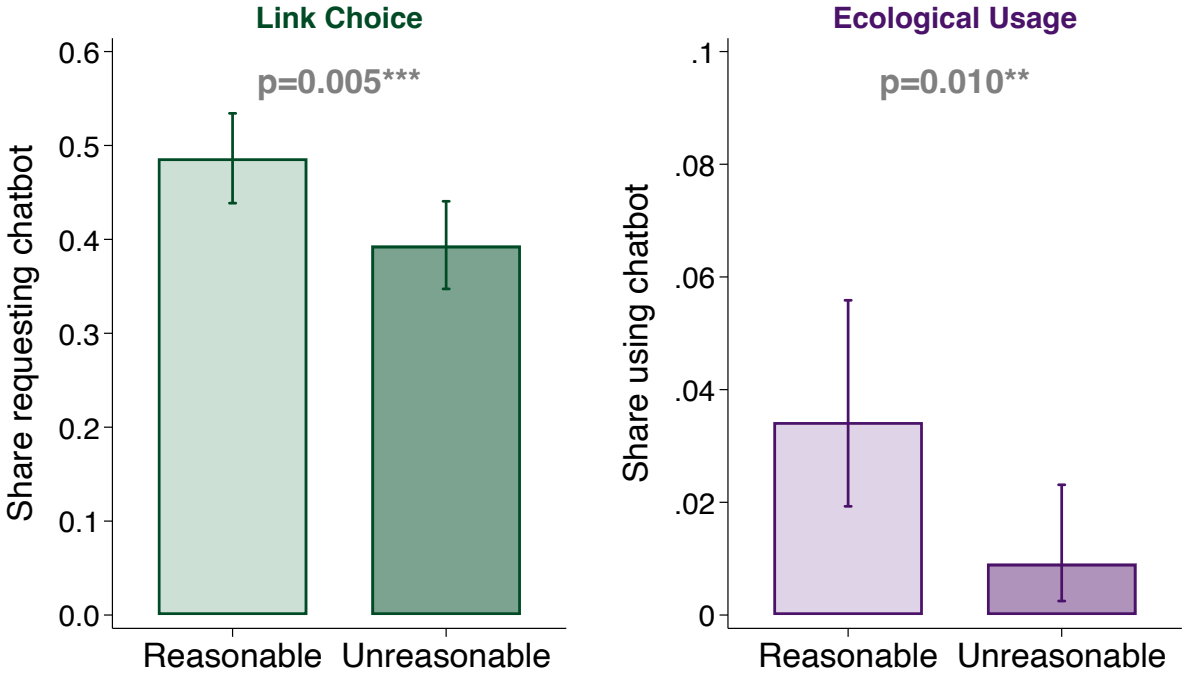
Notes: This table provides demographics of participants in the engagement experiment, run in August 2024. Participants were either trying to conceive, expecting, or parents to young children (percentages add up to over 100 as statuses may overlap). Participants were screened for parenting status, age (between 18 and 45), and gender (quota of 85% women). Samples recruited to rate the usefulness and reasonableness of conversations are distinct but highly similar in terms of demographics.

Table 18: Beliefs and Trust

Panel A:	***Beliefs in Chatbot Performance***			
	Post 4th Conv		**Post 5th Conv**	
	(1)	(2)	(3)	(4)
Unreasonable	-13.831***	-13.877***	-9.799***	-10.505***
	(2.149)	(2.110)	(1.883)	(1.849)
Post 4th Conv			0.549***	0.546***
			(0.031)	(0.030)
Controls	Yes	Yes	Yes	Yes
Conversation FE	No	Yes	No	Yes
R^2	0.061	0.099	0.361	0.389
Observations	894	894	894	894
Panel B:	***Trust in Chatbot***			
	Post 4th Conv		**Post 5th Conv**	
	(1)	(2)	(3)	(4)
Unreasonable	-0.761***	-0.760***	-0.604***	-0.640***
	(0.125)	(0.123)	(0.107)	(0.106)
Post 4th Conv			0.564***	0.563***
			(0.030)	(0.030)
Controls	Yes	Yes	Yes	Yes
Conversation FE	No	Yes	No	Yes
R^2	0.055	0.084	0.379	0.402
Observations	894	894	894	894

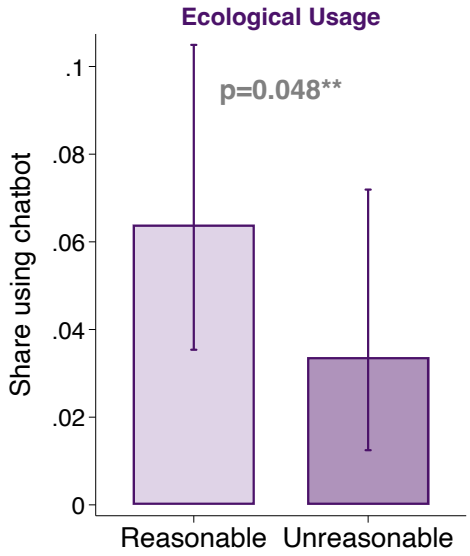
Notes: The dependent variables are posterior beliefs and trust, after seeing the fourth and fifth conversations (the first three conversations are successes, the last two are failures). Reasonable is the treatment dummy, and we include the 4th posterior as control. We include conversation fixed effects. Controls include conversations' usefulness and demographics: age, gender, income, and prior familiarity with AI.

Figure 38: Treatment Effect on Engagement Measures - Excluding Subjects Familiar with ParentData



Notes: The figure plots the share of subjects in each treatment choosing to receive the link to the chatbot over the link to parenting articles, and the share who actually use the chatbot after the experiment (defined as asking at least one question within 3 weeks after the experiment). We exclude subjects who reported being familiar 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 39: Treatment Effect on Engagement Measures - Conditional Engagement



Notes: The figure plots the (conditional) share of subjects in each treatment who actually use the chatbot after the experiment (defined as asking at least one question within 3 weeks after the experiment), computed among those who choose to receive the chatbot link. We exclude subjects who reported being familiar with ParentData. Sample sizes are $n = 213$ for *Reasonable* and $n = 173$ for *Unreasonable*. p-value of two-sided test of proportions is reported.

E.3 Conversations

Details on elicitation. Two distinct samples of current or expecting parents were recruited, using the same screening criteria as in the engagement experiment. Each conversation was presented on its own page, using a layout that replicates the ParentData.org webpage. Instructions urged participants to read conversations carefully, and to provide their honest assessment of AI's answers. Usefulness is on a 1-5 scale, while reasonableness is on a 0-100 scale. Each conversation was rated on average by 26 parents for usefulness, and 35 for reasonableness.

E.3.1 Pairs of useless conversations

Intent: Is COVID dangerous for infants?	
Low Reasonableness	**High Reasonableness**
Query: My 6 month old has Covid. Will it be ok? *Answer:* The COVID-19 vaccine is approved for infants from 6 months up. This is based on studies showing safety and showing efficacy in producing antibodies.	*Query:* My baby has Covid. Will she be ok *Answer:* Most data suggests that COVID-19 does not cause birth defects or other problems in the baby if the mother gets infected while pregnant. There is some evidence to suggest that getting infected in the first trimester might be a bit more concerning than getting infected later in pregnancy, but even then, the risks seem quite low. The biggest risk in this situation is probably to you, as pregnant women who get COVID-19 are more likely to be hospitalized and to have severe illness. The best advice is to get vaccinated, and to monitor symptoms if you do get sick.

Notes: Statistics are for low and high conversations respectively. Median usefulness: 2 & 2. Mean usefulness: 1.76 & 2. Median reasonableness: 10 & 20. Mean reasonableness: 19.86 & 29.70.

Intent: What is the best car seat to buy for my infant?	
Low Reasonableness	**High Reasonableness**
Query: What's the best car seat to buy? *Answer:* When it comes to choosing the best baby food brand, it really depends on what you're looking for. There are so many options out there, from organic brands like Plum Organics and Earth's Best to homemade baby food services like Yumi. Each brand has its own unique offerings and ingredients to consider. You may want to think about factors like convenience, price, and whether you prefer purees or baby-led weaning options. Ultimately, the best brand is the one that aligns with your values and meets your baby's nutritional needs. It's always a good idea to read labels, compare ingredients, and consult with your pediatrician for personalized advice. Happy feeding!	*Query:* Which is the best car seat *Answer:* When it comes to installing a car seat, the safest place is in the back seat. According to the National Highway Traffic Safety Administration (NHTSA), the back seat is generally the safest place for children under the age of 13. In fact, research shows that children are 70% less likely to be injured in a crash when properly restrained in the back seat. It's important to follow the specific guidelines provided by the car seat manufacturer and the NHTSA to ensure proper installation and positioning.

Notes: Statistics are for low and high conversations respectively. Median usefulness: 1 & 1. Mean usefulness: 1.24 & 1.41. Median reasonableness: 0.5 & 13. Mean reasonableness: 17.07 & 23.3.

Intent: Is getting COVID while pregnant dangerous for the fetus?	
Low Reasonableness	**High Reasonableness**
Query: I am 33 weeks pregnant and have tested positive for COVID 19. What are the risks to my baby? *Answer:* At 37 weeks, your baby is considered 'early term' rather than 'full term.' While there is a slightly increased risk of complications compared to babies born later, the majority of these complications are minor, and NICU stays tend to be short if they are needed. One study found that the rate of NICU admission at 37 weeks was 9.4%, compared to 5.8% at 38 weeks and 5% at 39 weeks.	*Query:* How does covid affect the fetus during pregnancy? *Answer:* COVID-19 can have an impact on fertility, but the overall effect is still being studied. Some research suggests that the virus may affect sperm count and motility in men, while others indicate potential disruptions to the menstrual cycle in women. However, it's important to note that these effects appear to be temporary and reversible. Additionally, the risk of severe illness from COVID-19 may be higher during pregnancy, so it's crucial to take necessary precautions. If you're planning to conceive, it's best to consult with your healthcare provider for personalized advice and guidance. Keep yourself safe and healthy!

Notes: Statistics are for low and high conversations respectively. Median usefulness: 1 & 1. Mean usefulness: 1.65 & 1.79. Median reasonableness: 6 & 19. Mean reasonableness: 16.11 & 30.03.

Intent: Can you eat chicken salad while pregnant?	
Low Reasonableness	**High Reasonableness**
Query: Can you eat chicken salad sandwich when pregnant *Answer:* Chicken liver is a nutritious food that is high in iron and other essential nutrients. However, when it comes to eating chicken liver during pregnancy, it's important to exercise caution due to the risk of foodborne illness. Chicken liver should be cooked thoroughly to kill any potential bacteria or parasites, such as salmonella or toxoplasma. It's recommended to avoid consuming undercooked or raw chicken liver during pregnancy. While it's always best to consult with your doctor for personalized advice, it's generally safer to opt for fully cooked poultry to ensure the health and safety of you and your baby.	*Query:* Can you have chicken salad if you're pregnant? *Answer:* Absolutely, salad can be a healthy and nutritious choice during pregnancy! Leafy greens like spinach and kale are packed with important vitamins and minerals. Just make sure to wash the greens thoroughly to remove any potential bacteria. If you're concerned about foodborne illnesses, it's good to know that there is a very low risk of contracting them from properly washed and prepared vegetables.

Notes: Statistics are for low and high conversations respectively. Median usefulness: 2 & 2. Mean usefulness: 2.12 & 2.69. Median reasonableness: 3 & 23. Mean reasonableness: 19.02 & 32.54.

Intent: How soon is it safe to get pregnant after giving birth?	
Low Reasonableness	**High Reasonableness**
Query: How soon can you get pregnant after giving birth *Answer:* After ovulation, the egg is available for fertilization for a couple of days. If a sperm meets the egg during this time and fertilization occurs, pregnancy can happen. But the chances of getting pregnant decrease significantly as time passes after ovulation. So, to maximize your chances of pregnancy, it's best to have sex on the day of ovulation or the day before. Timing is everything!	*Query:* How soon after having a baby is it safe to get pregnant again? *Answer:* There is no set waiting period for resuming sex after giving birth. Physically, if you have had tearing, it is important to wait until the perineum is healed. Depending on the severity of the tearing, this could happen much before six weeks, or it could take longer. The standard advice of waiting six weeks appears to have been invented by doctors so husbands wouldn't ask for sex. Your doctor will check this at your first postpartum checkup (which is, in fact, around six weeks), but you may be able to tell if you've healed before that. In addition to physical readiness, emotional readiness is also important. You need to want to have sex, and both partners need to be ready. It is important to take it slow and use lubrication if needed, as breastfeeding can cause vaginal dryness.

Notes: Statistics are for low and high conversations respectively. Median usefulness: 2 & 2. Mean usefulness: 2.15 & 2.53. Median reasonableness: 12 & 24.5. Mean reasonableness: 26.29 & 29.84.

E.3.2 Useful conversations

Intent: Is cold medicine safe during pregnancy? (Usefulness: 4 median; 3.96 mean)
Query: Is cold medicine safe to take during pregnancy?
Answer: When it comes to cold medicine during pregnancy, it depends on the medication and the phase of pregnancy you are in. Acetaminophen (Tylenol), for example, is generally considered safe to use for fever and pain relief throughout pregnancy, whereas others, like Sudafed, are considered safe in the second and third trimester.

Intent: Is biking safe during pregnancy? (Usefulness: 5 median; 4.71 mean)
Query: Can I bike while pregnant?
Answer: Riding a bike while pregnant can be a great way to stay active and enjoy the outdoors. The good news is that cycling is generally considered safe during pregnancy, as long as you take certain precautions. It's important to listen to your body and make adjustments as needed. Avoid rough terrains and opt for smoother paths. Pay attention to your balance and avoid sudden movements. Also, make sure your bike is properly adjusted to ensure a comfortable and

safe ride. As always, it's a good idea to consult with your healthcare provider before starting or continuing any exercise routine during pregnancy. Happy cycling!

Intent: Are white noise machines safe to help a baby sleep? (Usefulness: 5 median; 4.53 mean)
Query: Is it safe to use white noise machines for my baby's sleep?
Answer: White noise machines are intended to help babies sleep, but there are concerns about whether they can damage infant hearing. Noise exposure greater than 85 decibels for eight hours can damage hearing, and a study of 14 baby-sleep white noise machines found that some of them produced noise levels above 85 decibels when placed in the crib. Therefore, to be safe, parents should try to generate a level of 70 to 75 decibels for the white noise they use and turn it down once the baby is asleep. Using a phone app to measure the decibel level of the machine can help parents find the right level.

E.4 Experimental instructions: ParentData

E.4.1 Reasonableness Elicitation

Welcome to our survey!

In this study, you will be asked to evaluate the reasonableness of different misunderstandings.

You will read several (real) conversations between a parent (or prospective parent) and an artificial intelligence (AI) chatbot. The parent asks questions about parenting (dos and don'ts during pregnancy, breastfeeding, etc), and the AI answers these questions by looking up topics in a comprehensive parenting textbook, which is based on currently available scientific evidence.

Unfortunately, the AI sometimes misunderstands the question and looks up the wrong topic. As a result, the answer might not be helpful given the question asked.

In all conversations you will see, the AI misunderstood questions and gave unhelpful answers. We are interested in your assessment of how likely you think it is that a reasonable human would misunderstand the question in the way the AI did.

Put differently, we are NOT asking you to say whether the answer is correct or incorrect. We are only interested in how reasonable is the AI's misunderstanding, in your opinion. There is no "right" answer, we simply want your honest opinion.

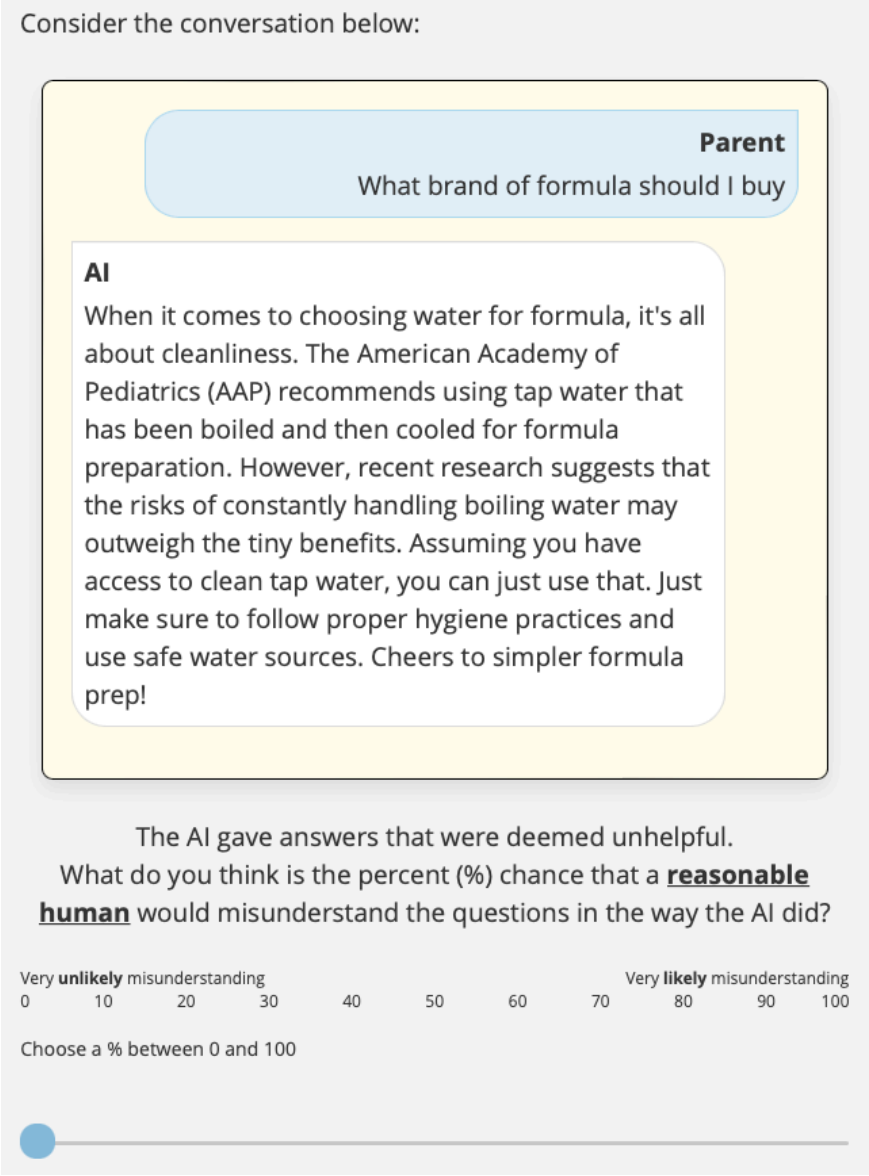
Here is an example of what the decision screen looks like: [Screenshot of decision]

If you provide focused and sincere answers, you will have a chance to earn $1 as bonus payment. Once data collection is complete we will draw half of the participants who passed data quality checks (attention and comprehension questions) and deliver them the bonus. The next screen contains comprehension checks.

Elicitation. 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. [slider]

Figure 40: Screenshot of Reasonableness Elicitation



E.4.2 Usefulness Elicitation

Welcome to our survey!

We are conducting academic research on the human use of artificial intelligence.

In this study, you will read several (real) conversations between a parent (or prospective parent) and an artificial intelligence (AI) chatbot. The parent asks questions about parenting (dos and don'ts during pregnancy, breastfeeding, etc), and the AI answers these questions by looking up topics in a comprehensive parenting textbook, which is based on currently available scientific evidence.

The quality of these answers may vary: sometimes the AI misunderstands questions and gives unhelpful answers.

For each question, your task is to assess the usefulness of the answer, on a 1-5 scale. "Usefulness" = does the answer's content provide a useful response to that specific question. Do NOT conclude an answer is useful only because it is on the same theme/topic as the question: some answers may appear useful at a quick glance, even though they are not! It is therefore crucial for us that you read both question and answer carefully. Try putting yourself in the parent's shoes, who had a specific question in mind and expected a precise answer.

There is no right or wrong answer to this question. We simply ask that you give the answer an attentive read and provide us with your honest opinion.

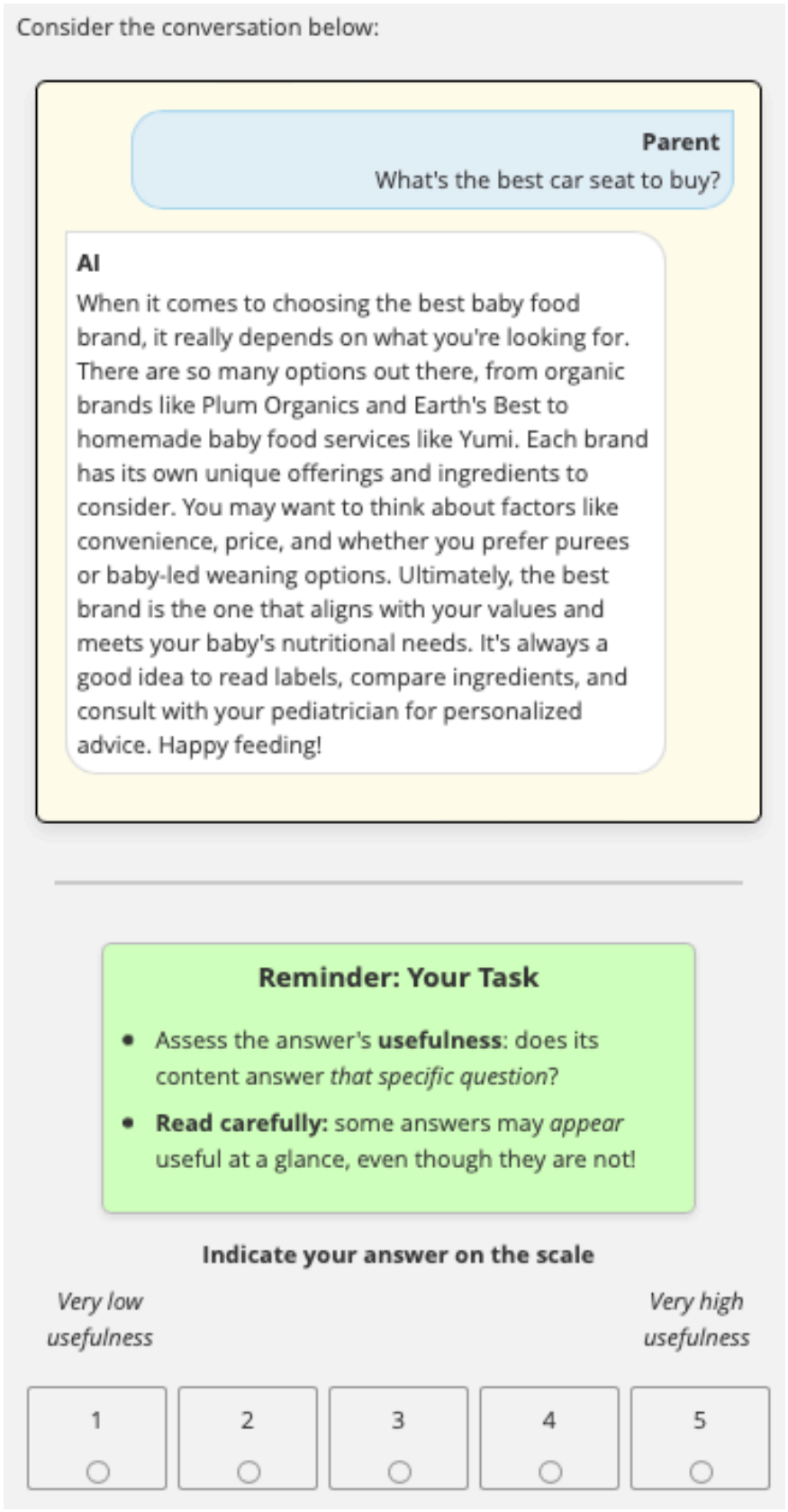
You will evaluate a total of 10 conversations. Some may contain technical medical terms or acronyms: you are welcome to google those terms to help your understanding of conversations!

If you provide focused and sincere answers, you will have a chance to earn $1 as bonus payment. Once data collection is complete we will draw half of the participants who passed data quality checks (attention and comprehension questions) and deliver them the bonus. The next screen contains comprehension checks.

Good luck!

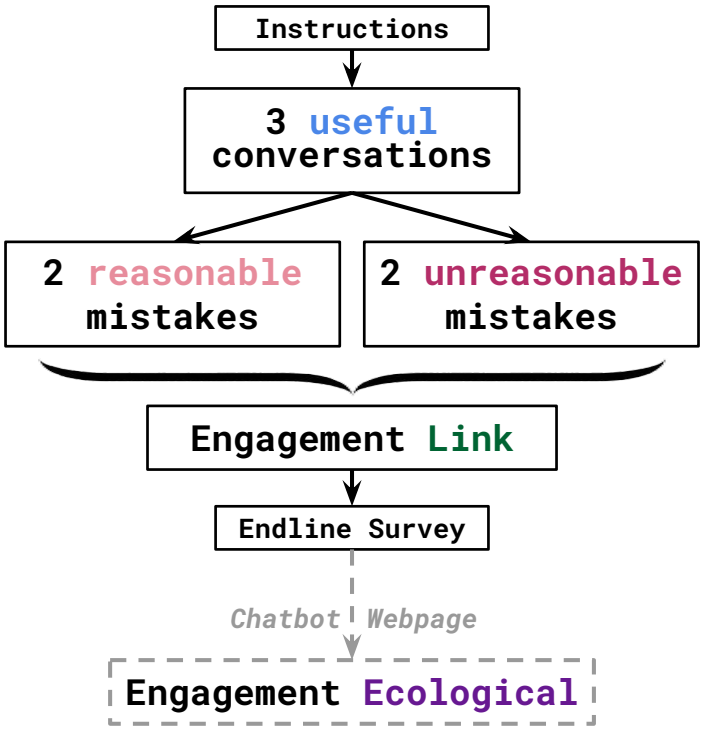
Elicitation. Reminder: Your Task Assess the answer's usefulness: does its content answer that specific question? Read carefully: some answers may appear useful at a glance, even though they are not! Indicate your answer on the scale [1-5 scale]

Figure 41: Screenshot of Usefulness Elicitation



E.4.3 Engagement Experiment

Figure 42: Flowchart of Engagement Experiment



Notes: Useful conversations are held fixed across treatments, and are presented in a random order. Endline survey contains usual demographic questions, as well as measures of prior familiarity with AI and with ParentData.org. The link chosen in the engagement decision (chatbot or articles) opens in a new tab upon exiting the survey.

Instructions. Welcome to our survey! In this study, you will interact with an artificial intelligence (AI) chatbot who specializes in answering parenting questions. This is a highly specific chatbot, it is different from famous chatbots you may have heard of before.

In what follows you will see a total of 5 questions that are typically asked by current or prospective parents. After reading each question, you will be able to generate an answer to the question using the chatbot. We will ask you your impressions after each conversation, including your beliefs about the chatbot's performance. If your answer is accurate (within 10 percentage points of the success rate) you will get 10 cents as bonus!

The quality of these chatbot-generated answers may vary (they can be more or less adequate given the questions asked). Please read those answers carefully.

After this, you will have the opportunity to receive different links to parenting resources. When you are ready, click on the button below to start. Good luck!

Belief and Trust Elicitation.

What do you think is the percent (%) chance that the chatbot correctly answers a random parenting question? Indicate your answer on a scale from 0% to 100% [Scale]

How much do you trust the chatbot? Indicate your answer on a scale from 1 to 7 [Scale]

Eliciting willingness to engage. Thanks for completing the previous task! Now, you may have your own parenting questions you would like good answers to.

You can choose to receive at the end of the survey a link to either the chatbot or to a large list (500+) of good-quality parenting articles. Both options are completely free, and of course you can use them for as little or as much as you like!

At the end of the survey I would like to receive: [Link to chatbot/Link to articles]

Figure 43: Screenshot of Belief and Trust elicitation

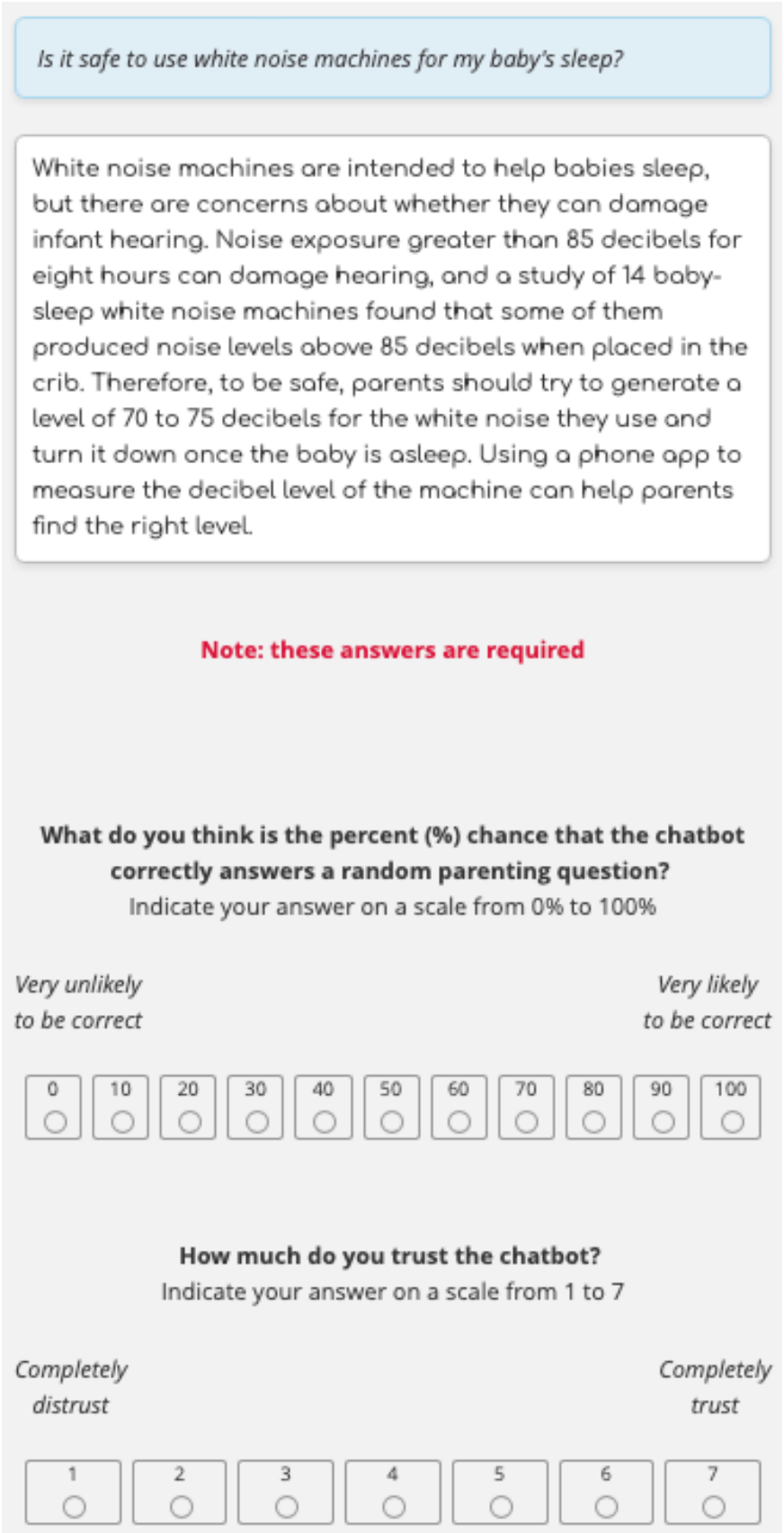


Figure 44: Screenshot of Engagement Decision

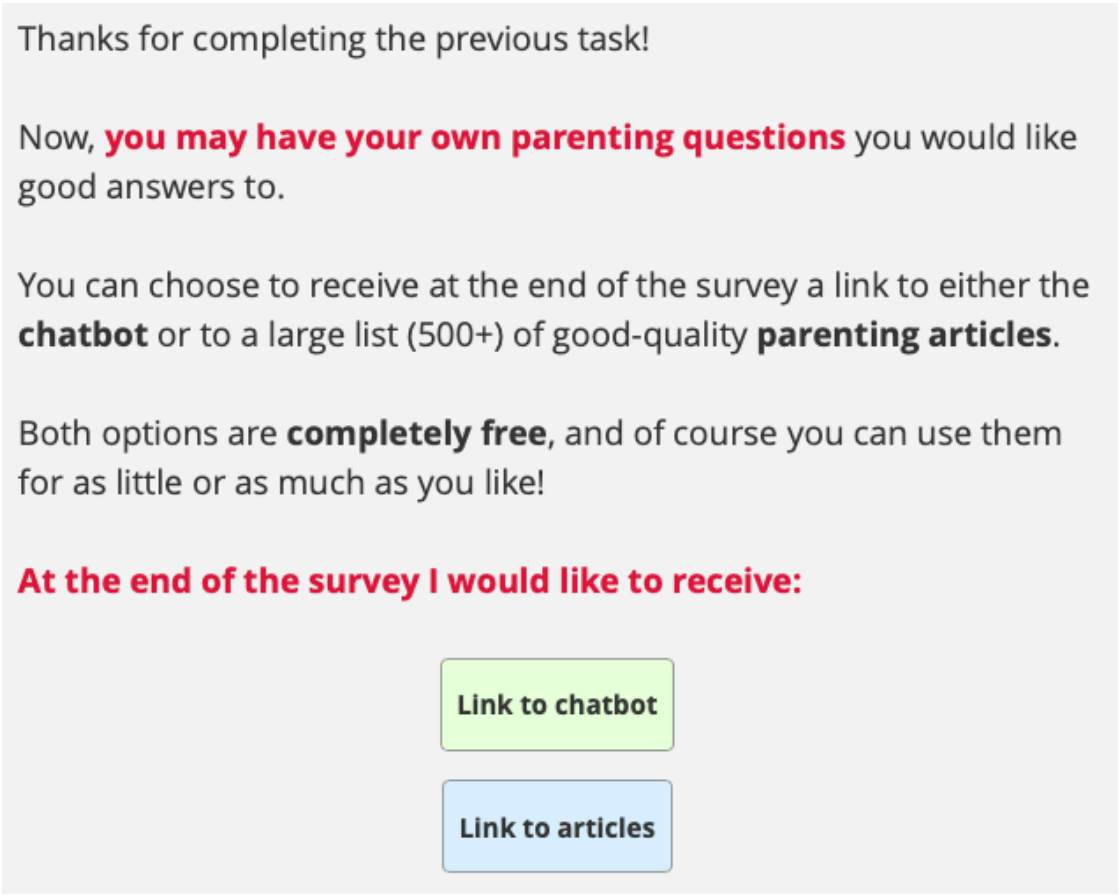


Figure 45: Screenshot of Dewey's Webpage

