Who is More Bayesian: Humans or ChatGPT?*

Tianshi Mu, Tsinghua University
Pranjal Rawat, Georgetown University
John Rust, Georgetown University †
Chengjun Zhang, Georgetown University
Qixuan Zhong, Georgetown University

December 22, 2025

Abstract

We compare human and artificially intelligent (AI) subjects in classification tasks where optimal choices are defined via Bayes' Rule. Experimental studies reach mixed conclusions about whether human beliefs accord with Bayes' Rule. We reanalyze landmark experiments using a behavioral model of decision making and show that decisions can be nearly optimal for specific reward functions even when beliefs are not Bayesian, though optimal behavior for all possible reward functions requires Bayesian beliefs. Using an objective measure of "decision efficiency," we find that humans are 96% efficient despite the fact that only a minority have Bayesian beliefs. We replicate these same experiments using three generations of ChatGPT as subjects. Using the reasoning provided by GPT responses to understand its "thought process," we find that GPT-3.5 ignores the prior and is only 75% efficient, whereas GPT-4 and GPT-4o use Bayes' Rule and are 93% and 99% efficient, respectively. Most errors by GPT-4 and GPT-4o are algebraic mistakes in computing the posterior, but GPT-4o is far less error-prone. GPT performance increased from sub-human to super-human in just 3 years. By version 4o, its beliefs and decision making had become nearly perfectly Bayesian.

Keywords: Bayes' Rule, decision making, statistical decision theory, win and loss functions, learning, Bayes' compatible beliefs, noisy Bayesians, classification, machine learning, artificial intelligence, large language models, ChatGPT, maximum likelihood, heterogeneity, mixture models, Estimation-Classification (EC) algorithm, binary logit model, structural models

*We dedicate this study to the memory of David Grether and Daniel Kahneman for their important contributions to economics and psychology, and particularly their pathbreaking work in experimental and behavioral economics and the study of individual decision making that inspired this paper. We thank Mahmoud El-Gamal and David Grether as well as Charles Holt and Angela Smith for providing the human experimental data that were reanalyzed in this paper. Rust is grateful for financial support from his University Professorship at Georgetown University. We are also grateful for helpful feedback from Tracy Xiao Liu and Yucheng Liang and seminar participants at the Universities of Texas at Austin, Chicago, and Toronto.

†**Corresponding Author:** Department of Economics, Georgetown University, `jr1393@georgetown.edu`.

1 Introduction

We compare the performance of human and artificially intelligent (AI) decision makers in binary classification tasks where optimal decisions are defined in terms of Bayes' Rule. AI algorithms such as support vector machines or neural networks can be trained to closely approximate optimal Bayesian decision rules for these tasks. Machine learning methods have been extended to more difficult real-world classification problems where the covariates used to classify outcomes can be very high dimensional (e.g., using mammograms to detect breast cancer). A number of studies have shown these specially trained classifiers can perform at superhuman levels, see, e.g., Yoen and Chang (2023).

It is not surprising that humans, whose brains consume only about 20 watts of power, do not outperform special-purpose machine learning algorithms trained with large volumes of data to approximate optimal decision rules for specific tasks. Kühl et al. (2022) note that humans possess *general intelligence* whereas, prior to the advent of large language models (LLMs), most special-purpose machine learning algorithms were designed to work in narrow domains and will not necessarily make sensible decisions in a huge variety of different (and often unexpected) situations that humans do. As Hutchinson and Meyer (1994) noted, *"From a broader perspective, however, one can argue that optimal solutions are known for a relatively small number of similar, well-specified problems whereas humans evolved to survive in a world filled with a large and diverse set of ill-specified problems. Our 'suboptimality' may be a small price to pay for the flexibility and adaptiveness of our intuitive decision processes."*

Rapid recent improvements in large language models and generative AI suggest that we may be close to the advent of *Artificial General Intelligence* (AGI) where general-purpose algorithms equal or exceed human performance in solving a wide range of problems even though the algorithms were not specifically trained to do well in specific narrow domain tasks. Generative AI models such as ChatGPT are deep neural networks with billions of parameters that have been trained to predict text and images using vast databases obtained from the web and other sources. The progress in this area has been breathtaking, and now a variety of LLMs have demonstrated a capability to compete with humans on a wide range of intellectual tasks.[1]

[1]In the paper, we will use interchangeably the abbreviations LLM and GPT (for Generative Pretrained Transformer), though the latter is a subset of the former.

Despite the rapid improvements, the consensus is that LLMs still lack full rationality, including the capability to reason and think creatively the way humans do, and other features associated with intelligence including "consciousness". The review by Maslej et al. (2024) concludes that *"AI has surpassed human performance on several benchmarks, including some in image classification, visual reasoning, and English understanding. Yet it trails behind on more complex tasks like competition-level mathematics, visual commonsense reasoning and planning."*

The rationality of human subjects and the conformance of their choice and elicited beliefs in simple binary classification problems has been extensively studied in numerous experiments in economics and psychology. The consensus is that humans are not Bayesian and make suboptimal decisions due to systematic biases, including "framing" and contextual effects that might be caused by reliance on heuristics to reduce cognitive burden, see, e.g., Tversky and Kahneman (1974) and the survey by Benjamin (2019).

However, this conclusion is controversial due to the use of "real world" scenarios to test decision making (e.g., asking whether a person is more likely to be an engineer or a lawyer based on a description of their appearance), since it provides extraneous information that amplifies the potential for framing effects and stereotyping to distort judgments. Grether (1978) noted that Kahneman and Tversky's experiments had "features that make the applicability of the findings to economic decisions doubtful" due to the "difficulty of controlling the information given when verbal descriptions or situations are presented. Both of these difficulties could be taken care of by the use of actual balls in urns or book-bag poker chip setups." (p. 71-72).[2]

Motivated by this logic and to give humans the "best shot" we reanalyze the experiments of El-Gamal and Grether (1995), who showed subjects random samples of balls drawn with replacement from one of two bingo cages, A or B, with different proportions of red and blue balls. A credible random procedure (e.g., dice throw) was used to select the cage to draw the sample, providing an "objective prior" that differed across trials. Based on the prior and sample outcome, subjects chose the cage they thought was more likely to have been selected to draw the observed sample. We also reanalyze experiments

[2] Cosmides and Tooby (1996) also argued that experiments framed in frequentist terms are more likely to generate behavior that conforms to Bayes' Rule since "our inductive reasoning mechanisms do embody aspects of a calculus of probability, but they are designed to take frequency information as input and produce frequencies as output."

by Holt and Smith (2009) who used a similar design but asked subjects to directly report their self-assessed probability that the sample was drawn from cage A.

Our conclusions are based on a behavioral model of decision-making that enables us to infer subjective beliefs using binary data on the chosen cage as well continuous elicited beliefs. Previous models, such as the cutoff rule model of El-Gamal and Grether (1995), focused on subjects' behavior (decision rules) rather than their beliefs: our model captures both. The model can also be interpreted as a flexible but parsimoniously parameterized two layer neural network that fits the data better than previous models. The first layer captures how the subject forms a belief based on the available information, and the second "output layer" captures the subject's choice of cage A or B given their belief. The model incorporates unobserved stochastic shocks that account for errors in processing the information and making a final choice, and is able to represent a wide range of biased or distorted beliefs. It nests a model of *noisy Bayesian* decision making where beliefs coincide with Bayes' Rule but choices reflect additional "decision noise", and perfectly Bayesian decision making when there is no decision noise.

We strongly reject the hypothesis that the choices of human subjects is governed by Bayes' Rule. We find substantial heterogeneity in subjects' beliefs with many exhibiting biased beliefs that include representativeness (i.e., overweighting the sample) and conservatism (overweighting the prior). Only half of the subjects can be described as noisy Bayesians. Of these nearly half are affected by substantial levels of decision noise that cause their choices to frequently deviate from the optimal Bayesian decision rule.

A central contribution of this study is to provide a simple objective measure of performance, *decision efficiency,* that allows us to compare human and AI subjects when their choices are governed by decision rules that may reflect biases in beliefs and different types of random mistakes. An *optimal decision rule,* i.e., one that maximizes the probability of selecting the correct bingo cage, can be defined in terms of Bayes' Rule. Decision efficiency is the ratio of expected payoffs under the subject's decision rule to optimal expected payoffs under Bayes' Rule. An important insight from the behavioral model is that *a decision rule can be optimal even if the subject's beliefs are not Bayesian.* We show that decision noise and belief biases have different effects on performance, and a subject with biased beliefs can outperform a noisy Bayesian. However this conclusion is dependent on the use of a symmetric payoff function to reward subjects in these experiments. We show that a decision rule based on subjective beliefs cannot be optimal for all possible payoff functions unless the beliefs coincide with Bayes Rule and choices are not affected by extraneous "decision noise".

Section 2 introduces the relevant statistical decision theory and defines the *win function,* i.e. the probability of selecting the correct cage under any given decision rule, and from this we define our scalar efficiency measure, the ratio of the win probability implied by the subject's decision rule to the optimal win probability implied by Bayes Rule. We argue that efficiency is a better measure of performance than the commonly used notion of *accuracy,* i.e., the fraction of a subject's choices that coincide with Bayesian choices. Our efficiency measure differentiates between "hard cases" (where the Bayesian posterior probability is close to 1/2) and "easy cases" (those where the posterior probability is close to 0 or 1). Among two subjects with equal accuracy the subject whose choices disagree with Bayes' Rule mostly on the easy cases will have a lower expected payoff and thus lower efficiency.

A surprising finding from our reanalysis of the experimental data is that despite the prevalence of biases and decision noise, human subjects are remarkably good predictors overall, with an average efficiency of 96%. A minority of the best-performing human subjects have small levels of decision noise and beliefs that are nearly Bayesian, and thus achieve nearly 100% efficiency.

Next, we compare the behavior and performance of human and GPT subjects using the same behavioral model and experimental design used in the human experiments. We find rapid improvement in the efficiency of GPT subjects over successive versions of the GPT software. The earliest version, GPT-3.5 (released in 2022), displayed distinctly sub-optimal behavior, with efficiency 12% lower than humans in the binary choice experiments of El-Gamal and Grether (1999) and 22% lower in the more challenging experiments of Holt and Smith (2009).[3] However, the subsequent version, GPT-4 (released in 2023), has decision efficiency that is comparable to human subjects, and the most recent version we analyzed GPT-4o (released in 2024) behaves as a "noisy Bayesian" but with less noise than human subjects, and as a result it surpasses human efficiency in both experiments.[4]

[3]These experiments were more difficult because subjects were asked to report the probability that cage A was used to draw the observed sample using a complicated second stage gamble known as the BDM mechanism (described in Section 4.2) to incentivize accurate reporting of the posterior.

[4]We also conducted a limited analysis using ChatGPTo1, and found that its efficiency is close to 100%, with behavior that closely approximates a perfect Bayesian decision maker.

An advantage of AI subjects is that they provide step-by-step reasoning with each answer that gives a unique window into their "thought process" that allows us to isolate where they make their errors. Analyzing the textual responses, we find that the sub-optimal performance of GPT-3.5 is due to ignoring prior information, so its beliefs are not Bayesian. GPT-4 generally recognizes the applicability of Bayes' Rule but makes algebraic errors in the process of computing the posterior. GPT-4o is a remarkable improvement: it understands the applicability of Bayes' Rule and makes far fewer algebraic errors in transforming the formula for Bayes' Rule to obtain numerical values for the posterior probabilities.

Why should we care whether GPT is more "Bayesian" than humans? Bayes' Rule captures an essential principal of logical reasoning in probabilistic environments by encoding how a rational decision maker should update their beliefs in response to new information. Optimal decision making also depends on having accurate understanding of the probabilistic environment and the consequences or payoffs taking different possible actions. Decision makers who are better at recognizing and formulating the underlying decision problem and processing new information in ways that more closely approximate the correct application of Bayes Rule will make objectively better decisions, as shown by Frick et al. (2024), who note that decision noise and learning biases "can lead to inefficient choices in many important economic problems, from career choices to financial investment decisions and voting" (p. 1612). We have shown that GPT has rapidly evolved to become more rational and Bayesian, outperforming human subjects who already behave close to optimally in a particular classification task. To the extent that LLMs are now able to outperform highly trained, professional human decision makers in a wide variety of more complicated real world decision making environments has serious implications for "human replacement" even in relatively high skilled occupations.

One such example is the problem of making optimal *differential diagnoses* (DDx) that involve classifying which of several alternative diseases or medical problems most likely caused by a set of observed symptoms in a patient. Recent studies (e.g., Yang et al. (2025), Goh et al. (2024), McDuff et al. (2023)) have demonstrated that LLMs and GPTs can outperform human physicians in the quality and accuracy of their differential diagnoses.[5]

[5]Studies have also evaluated the performance of human physicians who were allowed to use AI as a diagnostic

However, unlike our setting, these studies lack an objective definition of a "correct" diagnosis or optimal decision, and instead rely on the majority opinion of a panel of experts. Even experts frequently disagree due to ambiguity stemming from the incomplete state of medical science and differences in subjective assumptions about the costs and benefits of different diagnoses and treatment decisions. By contrast, laboratory experiments can eliminate or reduce this ambiguity by controlling the payoffs and specifying the probabilistic mechanism governing the signals subjects receive. This allows us to provide an objective definition of an optimal or correct decision and distinguish easy from hard cases in order to identify where subjects make their most costly mistakes.

We also contribute to the growing literature comparing the performance and rationality of humans and LLMs in economic decision making (Chen et al., 2023; Mei et al., 2024; Raman et al., 2024; Raman et al., 2025; Fish et al., 2025). Much of this work evaluates LLM "rationality" either through internal-consistency measures based on GARP and the Afriat Theorem, or through accuracy on large batteries of multiple-choice questions. However, internally consistent choices need not correspond to optimal decision rules, and accuracy measures weight errors by question categories rather than by their economic consequences. By focusing on a narrowly defined urn-and-ball environment where we can mathematically characterize the optimal decision rule, we provide a setting with an objective ground truth and evaluate behavior using decision efficiency, a welfare-based metric grounded in statistical decision theory. This framework allows for a sharper comparison between humans and GPT models by quantifying not only whether decisions are consistent, but how close they come to maximizing expected payoffs under the true data-generating process.[6]

assistant, with the argument being that humans may be superior in processing hard to quantify contextual information from patients, whereas AI may be superior in processing available digital information. Ironically, Goh et al. (2024) found that allowing such collaboration does not improve performance: "This randomized clinical trial found that physician use of a commercially available LLM chatbot did not improve diagnostic reasoning on challenging clinical cases, despite the LLM alone significantly outperforming physician participants." See also Agarwal et al. (2024) who reach a similar conclusion in their comparison of the performance of human radiologists with a special purpose AI algorithm trained using to detect various pathologies in chest X-rays.

[6] Caplin et al. (2025) and Kovach et al. (2025) compare the accuracy of human and AI subjects in a classification problem with an objectively correct solution but an unknown data generating process: they asked subjects to judge whether individuals in photographs are over 21 years old. They show that a special-purpose AI face classifier significantly outperforms humans in this task, and even humans who are also shown the choice of the AI classifier. Though the true "likelihood function" is unknown in this problem, the AI classifier could be viewed as providing a non-parametric approximation to it, and thus the optimal Bayesian decision rule.

2 Statistical Decision Theory Background

This section introduces relevant statistical decision theory to provide an objective metric for comparing the performance of human and AI subjects. The human subject data we reanalyze in Section 4 were gathered from four separate experiments: 1) 257 student subjects from four different universities in California reported in El-Gamal and Grether (1995), 2) 79 student subjects at the University of Wisconsin reported in El-Gamal and Grether (1999), 3) 22 subjects at the University of Virginia, and 4) 24 subjects who participated in web-based experiments, both reported in Holt and Smith (2009).

These experiments used "binomial designs" that require subjects to choose one of two bingo cages, labeled A and B, each containing the same known number of balls of two types. A credible random mechanism was used to select one of the two cages (e.g., selecting one of the cages based on a toss of a die or a random number generator) though subjects were not shown which cage was selected. A random sample of D balls with replacement was drawn from the selected cage and shown to the subjects.

In experiments 1 and 2 subjects were asked to choose the cage they believed was most likely to have been used to draw the sample. In experiment 1 a subset of subjects received a \$10 bonus if they selected the actual cage used to draw the random sample of balls for a randomly selected trial out of the total trials they participated in, and in experiment 2 all subjects received a \$20 bonus for each correct response in 3 randomly selected trials.[7] In experiments 3 and 4 subjects were asked to report the probability that cage A was the one from which the sample was drawn using an incentive-compatible procedure introduced by Becker et al. (1964) known as the *BDM mechanism* which involves a second stage lottery whose payoff depends on the probability the subject reports.[8]

The problem of selecting the cage from which the observed sample was most likely

[7] In all experiments, incentive payments were made *after* all trials were completed. Beyond an initial description of the bingo cage setup and a single demonstration of how it works at the start of the experiment, *none of the subjects received any feedback on whether they had selected the correct cage after each trial in the experiment.* This was evidently an intentional feature of the experimental design, to reduce the possibilities of non-stationarity in subjects' decision rules during the experiment due to "learning-by-doing" that is enhanced by real-time feedback. We tested for learning by doing effects (simply due to repeated participation even without sequential feedback on whether their choices were correct) by comparing performance on the first third of trials with the last third. We find small learning by doing effects even for the non-incentivized subjects, even in the absence of sequential feedback about whether they had selected the correct cage after each trial. However, the effect is sufficiently small that we omit it from further analysis in this paper.

[8] The BDM lottery is designed so that the payoff maximizing report is the subjective posterior probability of cage A. We will describe the BDM mechanism in more detail in Section 4.2 below.

to have been drawn is an elementary *statistical decision problem* whose optimal solution is given by Bayes' Rule. Let d denote the number of balls in the sample of D balls that have a designated type (i.e., balls marked N in experiments 1 and 2, light balls in experiment 3 or red balls in experiment 4). Though d is a sufficient statistic for the full random sample, subjects were shown the full sample outcomes. Let p_A and p_B be the probabilities of selecting the designated type of ball from each cage. The probabilities equal the fractions of the total number of balls in each cage of the designated type. Let $f(d|p_A, D)$ and $f(d|p_B, D)$ be the probabilities of observing d balls of the designated type in the random sample of D balls for cages A and B. These are binomial distributions with parameters (p_A, D) and (p_B, D), respectively. Finally, let $\pi \in (0, 1)$ denote the credible *objective prior probability* that cage A was selected to draw the random sample of D balls.

The behavior of subjects in the experiments can be summarized by a *decision rule* which is a function $\delta(d, \pi, p_A, p_B, D)$ that maps the information provided to subjects in the experiments into a choice of cage A or B. Following El-Gamal and Grether (1995) we do not assume all subjects use the same decision rule, and our analysis will attempt to identify different *types* of subjects who use similar decision rules, using finite mixture methods that are closely related to their *Estimation-Classification* (EC) algorithm.

Our analysis of human and AI subject data also allows for probabilistic decision rules (i.e. "mixed strategies") as well as pure strategies that appear to be probabilistic because the subject's choice depends on idiosyncratic stochastic psychological "decision noise" that is not observed by the experimenter. To allow for this we define a decision rule as a conditional probability of selecting cage A.

Definition D1. Decision Rule: *Any conditional probability $P(A|d, \pi, p_A, p_B, D)$ of selecting cage A as a function of the publicly observable information (d, π, p_A, p_B, D).*

Note that P is also referred to as a *conditional choice probability* (CCP) — it is the probability that a subject chooses cage A, *not* the subjective probability that the sample came from cage A. An arbitrary decision rule δ need not be derivable from subjective beliefs about the likelihood the sample came from A. For example, a variety of machine learning algorithms such as support vector machines or neural networks can be trained to have nearly optimal decision rules, but they do not require or make use of subjective posterior beliefs about A. In Section 3 we introduce a behavioral model that allows subject choices to depend on their subjective posterior beliefs, Π_s. In contrast, Bayes'

Rule provides a formula for the true or *objective* beliefs, Π, that guide the decisions of rational decision maker who maximizes expected payoff.

Definition D2. Bayes' Rule: *The conditional probability that cage A was selected given the information* (d, π, p_A, p_B, D) *given by*

$$\Pi(A|d, \pi, p_A, p_B, D) = \frac{\pi f(d|p_A, D)}{\pi f(d|p_A, D) + (1-\pi) f(d|p_B, D)}. \tag{1}$$

Define two binary random variables, $\tilde{W}_P$ and $\tilde{L}_P$, implied by decision rule P by $\tilde{W}_P = 1$ if the subject selects the correct cage from which the sample was drawn, and 0 otherwise. Thus, $\tilde{W}_P$ is an indicator for a "win" i.e., a correct prediction or classification. $\tilde{L}_P$ is the indicator for a loss, i.e., an incorrect prediction. It follows that with probability 1 we have $1 = \tilde{W}_P + \tilde{L}_P$, and so we can define an optimal decision rule as one that maximizes the probability of a win or conversely one that minimizes the probability of a loss. Following the standard terminology from the literature on statistical decision theory, we define

Definition D3. Loss Function *The loss function is the conditional probability of a loss,*

$$\begin{aligned} L_P(d, \pi, p_A, p_B, D) &= E\{\tilde{L}_P|d, \pi, p_A, p_B, D\} \\ &= P(A|d, \pi, p_A, p_B, D)[1 - \Pi(A|d, \pi, p_A, p_B, D)] \\ &+ [1 - P(A|d, \pi, p_A, p_B, D)]\Pi(A|d, \pi, p_A, p_B, D). \end{aligned} \tag{2}$$

Definition D4. Win Function *The win function is the conditional probability of a win, i.e., selecting the correct cage,*

$$W_P(d, \pi, p_A, p_B, D) = E\{\tilde{W}_P|d, \pi, p_A, p_B, D\} = 1 - L_P(d, \pi, p_A, p_B, D). \tag{3}$$

Using equation (2) or (3) the optimal decision rule is the pure strategy (4) defined in terms of Bayes' Rule in Lemma L1.

Lemma L1. *The optimal decision rule for a statistical experiment with a binomial design* *can be defined in terms of Bayes' Rule by*

$$\delta^*(d,\pi,p_A,p_B,D) = \begin{cases} A & \textit{if } \Pi(A|d,\pi,p_A,p_B,D) \geq 1/2 \\ B & \textit{otherwise.} \end{cases} \tag{4}$$

Note that δ^* depends on the information (d,π,p_A,p_B,D) only via the Bayesian posterior, Π. In particular, it is a deterministic or *pure strategy* that does not depend on any other information. In the next section we define a class of potentially suboptimal decision rules P that depend on the payoff relevant information (d,π,p_A,p_B,D) via a *subjective posterior belief* $\Pi_s(A|d,\pi,p_A,p_B,D,\nu)$ that also reflects the effect of additional extraneous/irrelevant information denoted by ν that we refer to as *decision noise.* We will say that the mixed or probabilistic decision rule $P(A|d,\pi,p_A,p_B,D)$ is *generated by* Π_s if

$$P(A|d,\pi,p_A,p_B,D) = Pr\left\{\Pi_s(A|d,\pi,p_A,p_B,D,\nu) \geq 1/2\right\}. \tag{5}$$

The suboptimality of mixed decision rules P implied by subjective posterior beliefs comes from two sources: 1) "decision noise" ν and 2) "biased beliefs", i.e. any factor that causes subjective posterior beliefs Π_s to differ from the Bayesian posterior Π when decision noise is zero $\nu = 0$. It is easy to see that a decision rule based on subjective posterior beliefs Π_s will be optimal if and only if

1. Π_s is free from decision noise ν, and
2. Π_s is *Bayes-Consistent*

$$\Pi_s(A|d,\pi,p_A,p_B,D) \geq 1/2 \quad \text{if and only if} \quad \Pi(A|d,\pi,p_A,p_B,D) \geq 1/2. \tag{6}$$

Our empirical results demonstrate that human subjects who have non-Bayesian but Bayes-Consistent beliefs can behave nearly optimally when the level of decision noise is sufficiently small. Small shocks typically only change decisions when Π_s and Π are both close to 1/2, where human subjects and a perfect Bayesian decision maker are indifferent between choosing A or B. It follows that human behavior will differ from optimal behavior mostly for the "hard cases" where the loss from deviating from the optimal decision rule δ^* is small. As long as Π_s is not too close to 1/2, small shocks to Π_s will rarely lead to suboptimal decisions in the "easy cases" where the true Bayesian posterior is close to 0 or 1. But these are precisely the cases where the loss from deviating from the optimal decision rule δ^* is high.

It is important to note that the conclusion that non-Bayesian but Bayes-consistent beliefs can be optimal only holds for symmetric payoffs implicit in our definition of the loss function (2). We now show that the only way for a decision maker to behave optimally for *all possible payoff functions* is for their beliefs to be free of decision noise and coincide with Bayes Rule (1), i.e. $\Pi_s = \Pi$.

To see this, consider a binary decision problem where $a \in \{A, B\}$ denotes the decision maker's choice, $s \in \{A, B\}$ is an uncertain state of nature, and $u(s, a)$ denotes the *ex post* payoff the decision maker receives from taking action a when the realized state is s. Assume that u satisfies the restrictions $u(A, A) > u(B, A)$ and $u(B, B) > u(A, B)$. Note that win function (3) in Definition D4 corresponds to the symmetric payoff function where $u(A, A) = u(B, B) = 1$ and $u(A, B) = U(B, A) = 0$. Let $\delta^*(u, d, \pi, p_A, p_B, D)$ denote the optimal decision rule of a Bayesian decision maker who takes actions to maximize the expected payoff function u. Then it is not hard to see that δ^* takes the form of threshold rule given by

$$\delta^*(u, d, \pi, p_A, p_B, D) = \begin{cases} A & \text{if } \Pi(A|d, \pi, p_A, p_B, D) \geq \overline{p}(u) \\ B & \text{otherwise,} \end{cases} \tag{7}$$

where

$$\overline{p}(u) = \frac{u(B, B) - B(B, A)}{u(A, A) - u(A, B) + u(B, B) - u(B, A)} \in [0, 1]. \tag{8}$$

Let $\delta^*_s(u, \pi, p_A, p_B, D, \nu)$ denote the optimal decision rule of a subjective Bayesian decision maker who seeks to maximize the expected payoff function u *with respect to their subjective posterior beliefs.* Then it is easy to see that δ^*_s takes the same form as δ^* in equation (7) with the same threshold $\overline{p}(u)$ except that we replace the true Bayesian posterior Π with the subjective posterior Π_s. We will refer to δ^*_s and the implied mixed decision rule $P(A|d, \pi, p_A, p_B, D)$ generated by Π_s as *subjectively optimal* and the Bayesian decision rule δ^* in (7) as *objectively optimal.*

Lemma L2. *The only way for a subjectively optimal decision rule δ^*_s to be objectively optimal for all possible payoff functions, i.e. $\delta^*_s = \delta^*$ with probability 1 for all u, is for δ^*_s to be free of decision noise and $\Pi_s = \Pi$ for all possible values of (d, π, p_A, p_B, D).*

The proof of Lemma L2 is in Appendix A. See Frick et al. (2024) for a more general analysis of how different types of learning biases lead to inefficiency in terms of objective measures of welfare defined in terms of Bayes' Rule.

To compare the performance of different human and AI decision makers, it is convenient to have a single objective overall measure of *decision efficiency* namely, the ratio of the subject's expected win probability to the optimal expected win probability implied by Bayes' Rule, computed with respect to the distributions of the values of the experimental controls in the experiments and the possible values of d given (π, p_A, p_B, D) by

$$\begin{aligned} W_P(\pi, p_A, p_B, D) &= E\{\tilde{W}_P|\pi, p_A, p_B, D\} \\ &= \sum_{d=0}^{D} W_P(d, \pi, p_A, p_B, D)[f(d|p_A, D)\pi + f(d|p_B, D)(1-\pi)]. \quad (9) \end{aligned}$$

If $H(\pi, p_A, p_B, D)$ is the empirical distribution of the experimental control variables in all trials of the experiment, the overall expected win for a subject with decision rule P in this experiment is given by

$$W_P = E\{\tilde{W}_P\} = \int_\pi \int_{p_A} \int_{p_B} \int_D W_P(\pi, p_A, p_B, D) dH(\pi, p_A, p_B, D). \quad (10)$$

We will use W_P as a single summary statistic for the overall performance of decision rule P using our econometric estimates of P from our behavioral model of subject choice behavior discussed below. We can also define the corresponding optimal win probability in the same experimental design using the optimal decision rule implied by Bayes' Rule, W_{δ^*}. Then we can define an overall scalar efficiency metric ω_P equal to the ratio of the subject's expected win probability to the optimal win probability of a perfect Bayesian decision maker, $\omega_P = W_P/W_{\delta^*}$. Clearly, we have $0 \leq \omega_P \leq 1$.

3 Behavioral Econometric Model of Subject Responses

This section introduces a parsimonious yet highly flexible behavioral model of subject decision making that we refer to as the "structural logit model". It differs from the discrete threshold model introduced by El-Gamal and Grether (1995) (which we will describe in Section 4.1), but is related to and subsumes the "probability weighting" model used by Holt and Smith (2009) to analyze reported posterior and a "structural probit" model introduced by Grether (1980), and includes the optimal Bayesian decision rule as a special case as well. The model allows for various types of judgment biases studied in the psychology literature, as well as allowing for "decision noise" that can have separate effects on subjective beliefs and their ultimate binary choice via the random utility framework pioneered by McFadden (1974). The stochastic shocks can reflect "calculational errors" to determine a final numerical answer (which we show is present in responses of LLMs), as well what Enke and Graeber (2023) refer as *cognitive noise* that reflect "the difficulty of translating objective probabilities into decisions" by human subjects.[9]

The behavioral model also has an interpretation as a two layer neural network where the first input layer uses "transformed inputs" equal to the log-likelihood ratio and the log posterior odds ratio and the second output layer uses the subjective posterior probability output from the first layer as its input and includes it in a logistic "squashing function" that is a monotonic function of the difference between the subjective posterior and 1/2. Thus, it can also be viewed a flexible model for capturing a variety of other heuristics and rules of thumb subjects might use to make decisions that may not involve subjective posterior beliefs. However both human and LLM subjects do report numerical posterior probabilities in the data we re-analyze from Holt and Smith (2009), and a key advantage of the behavioral model is that provides a unifying framework that allows us to use subjects' beliefs when they are directly reported or to infer them in the experiments by El-Gamal and Grether where subjects only report the cage they believed to be more likely to have generated the observed sample d.

The conditional probability that a subject chooses cage A depends on the observed information (d, π, p_A, p_B, D) from the experiment and two additional stochastic shocks that are observed or experienced only by the subject, (ν, ε). We assume a subject makes their choice based on a *subjective posterior belief* $\Pi_s(A|d, \pi, p_A, p_B, D, \nu)$ that cage A was the one from which the observed sample was drawn. This depends on the public information from the experiment, as well as a private cognitive or calculational error ν that represents algebraic mistakes that a subject might make in mapping the information (d, π, p_A, p_B, D) into a numerical value of Π_s. Our specification for Π_s nests the true Bayesian posterior probability as a special case when $\nu = 0$. Specifically, we assume that

[9]See also Ba et al. (2025) who present a model of belief updating in which "limited attention leads people to form a distorted mental model or representation of the information environment, and limited processing capacity generates cognitive imprecision when using this representation to update beliefs." Their model assumes subjects receive a "cognitive signal" and then update their beliefs given a mental model, and "The noise in the cognitive signal induces randomness in the cognitive posterior."

subjects transform the experimental outcome data (d, π, p_A, p_B, D) into two "summary statistics" $\text{LPR}(\pi)$ and $\text{LLR}(d, p_A, p_B, D)$ where $\text{LPR}(\pi)$ is the log prior odds ratio and $\text{LLR}(d, p_A, p_B, D)$ is the log-likelihood ratio given by

$$\begin{aligned} \text{LPR}(\pi) &= \log(\pi/(1-\pi)) \\ \text{LLR}(d, p_A, p_B, D) &= \log(f(d|p_A, D)/f(d|p_B, D)). \end{aligned} \tag{11}$$

We allow for the possibility that human and AI subjects may make algebraic errors trying to evaluate the quantities $\text{LPR}(\pi)$ and $\text{LLR}(d, p_A, p_B, D)$. Let the scalar random variable ν equal the sum of these errors, so the log-posterior odds ratio that the subject would report if asked is given by

$$\log\left(\Pi_s(A)/(1-\Pi_s(A))\right) = \beta_0 + \beta_1 \text{LLR}(d, p_A, p_B, D) + \beta_2 \text{LPR}(\pi) + \nu. \tag{12}$$

In our empirical analysis below we assume $\nu \sim N(0, \eta^2)$. If we assume that subject reports cage A whenever $\Pi_s(A)(d, \pi, p_A, p_B, D) \geq 1/2$, the coefficients of the right hand side of equation (12) can be estimated using a binary probit model, and ν has a logistic distribution, then $(\beta_0, \beta_1, \beta_2)$ can be estimated with a binary logit model.

Solving equation (12) for $\Pi_s(A|d, \pi, p_A, p_B, D, \nu)$ results in the following logistic specification given by

$$\Pi_s(A|d, \pi, p_A, p_B, D, \nu) = \frac{\exp\{\beta_0 + \beta_1 \text{LLR}(d, p_A, p_B, D) + \beta_2 \text{LPR}(\pi) + \nu\}}{1 + \exp\{\beta_0 + \beta_1 \text{LLR}(d, p_A, p_B, D) + \beta_2 \text{LPR}(\pi) + \nu\}}. \tag{13}$$

Notice that the true Bayesian posterior $\Pi(A|d, \pi, p_A, p_B, D)$ given in equation (1) is a special case of (13) when $\beta = (0, 1, 1)$ and $\nu = 0$. For other values of β the subjective posterior can capture a number of well-known biases observed in past studies, including an outright bias for cage A or B if $\beta_0 \neq 0$ as well as *overconfidence* and *underconfidence* (i.e. β_1 and β_2 greater than 1 and less than 1, respectively) about the posterior probability of cage A. Additionally, there can be *base rate bias* ($\beta_2 < \beta_1$) resulting in behavior consistent with the representativeness heuristic (i.e., excessive weight on the data via LLR relative to the prior via $\text{LPR}(\pi)$), as well as conservatism ($\beta_2 > \beta_1$, i.e., putting excessive weight on prior information relative to sample information).

The subject's choice of cage A or B depends on their subjective expected reward from choosing either cage. Suppose the subject receives a reward R if they select the correct cage and 0 otherwise. In experiments where subjects were not paid for making a correct choice, R can be viewed as an internal "psychological reward" the subject receives from making a correct choice. The expected reward might also be affected by unobserved idiosyncratic preference shocks $\varepsilon = (\varepsilon(A), \varepsilon(B))$ that we assume are distributed independently of ν. In our empirical analysis below, we assume that ε has a bivariate Type 1 extreme value distribution with location parameter normalized to 0 and a common scale parameter σ. Normally, the subject should select cage A if $\Pi_s(A|d, \pi, p_A, p_B, D, \nu) > 1/2$ and cage B otherwise. However their choice might be affected by additional preference shocks ε that capture behavior such as simply guessing one of the cages, or other psychological factors or errors that cause the subject to choose a cage that does not have the higher subjective posterior probability. We observe such inconsistent choices among GPT subjects, whose textual responses reveal random errors in calculating the posterior for cage A (captured by the ν shock) but a final choice of the cage whose posterior probability is less than 1/2 (which can only be explained by the ε shocks). When both shocks are present, the subject's decision rule is given by

$$\delta(d, \pi, p_A, p_B, D, \nu, \varepsilon) = \begin{cases} A & \text{if} \quad R\Pi_s(A) + \sigma\varepsilon(A) \geq R\Pi_s(B) + \sigma\varepsilon(B) \\ B & \text{otherwise.} \end{cases} \tag{14}$$

As is well known from the discrete choice literature (see, e.g., McFadden (1974)) when ε has a bivariate Type 1 extreme value distribution, the probability that the subject chooses cage A is given by the mixed binomial logit formula

$$\begin{aligned} P(A|d, \pi, p_A, p_B, D) &= \Pr\{\delta(d, \pi, p_A, p_B, D, \nu, \varepsilon) = A|d, \pi, p_A, p_B, D\} \\ &= \int_\nu \frac{1}{1 + \exp\{R[1 - 2\Pi_s(A|d, \pi, p_A, p_B, D, \eta\nu)]/\sigma\}} \phi(\nu) d\nu, \end{aligned} \tag{15}$$

where ϕ is the standard normal density and η is the standard deviation of η. When $\Pi_s(A|d, \pi, p_A, p_B, D, \nu) = 1/2$ the subject is indifferent between choosing cage A or B and the noise terms $(\varepsilon(A), \varepsilon(B))$ determine the subject's choice, so $P(A|d, \pi, p_A, p_B, D, \nu)$ is also equal to 1/2. However, as $\Pi_s(A|d, \pi, p_A, p_B, D, \nu)$ approaches 0 or 1, the "strength of the evidence" reduces the role of the idiosyncratic shocks ε on the subject's choice. Thus, $P(A|d, \pi, p_A, p_B, D, \nu)$ increases to 1 when R is sufficiently large or σ is sufficiently small and $\Pi_s(A|d, \pi, p_A, p_B, D, \nu) > 1/2$, and conversely $P(A|d, \pi, p_A, p_B, D, \nu) \to 0$ as $R/\sigma \to \infty$ when $\Pi_s(A|d, \pi, p_A, p_B, D, \nu) < 1/2$.

As we noted, the behavioral model (15) can be interpreted as a two layer feedforward neural network that uses transformed inputs $\text{LPR}(\pi)$ and $\text{LLR}(d, p_A, p_B, D)$ and produces a single output from the "first layer", namely the subjective posterior probability of cage A. Then the second layer uses the output from the first layer to determine the probability of choosing cage A, i.e., the CCP. This network is fully determined by a total of 5 "weights" i.e., the three bias/input weights $(\beta_0, \beta_1, \beta_2)$ from the input layer and the weights η and R/σ at the output layer.[10] Maximum likelihood estimation of the behavioral model can be interpreted as training the neural network to behave like a human being. The ability of the behavioral model to fit a wide range of subject behaviors can be ascribed to the flexibility afforded by using a parsimoniously parameterized neural network to predict subject behavior.[11]

One might question the use of the log-likelihood ratio as a key input to the model as imposing "too much structure" via unrealistic assumptions about the cognitive capabilities of human subjects who might be untrained in probability theory, given that the likelihood ratio $\text{LLR}(d, p_A, p_B, D)$ involves binomial probabilities for cages A and B, respectively. As we already noted, our model does not presume that subjects are able to calculate LLR exactly: instead we assume subjects' responses are based on a potentially biased and noisy approximation to LLR. The key reason why we use $\text{LLR}(d, p_A, p_B, D)$ rather than simply the sample outcome d as the relevant covariate entering the model, is because LLR embodies the additional information (p_A, p_B, D) needed to determine whether the sample outcome d is more likely to have been drawn from cage A than B.

[10]The formula for the CCP in equation (15) does not include an output layer "bias term" reflecting the restriction that the probability of choosing A is 1/2 when the subjective posterior is 1/2. We can allow for decision rules that are biased for or against choosing cage A by incorporating an additional bias parameter in the output layer resulting in a total of 5 rather than 4 parameters. We did not include an output bias term in the behavioral model in equation (15) since a systematic bias for or against choosing cage A can be represented via "biased beliefs" via the input layer bias term β_0. It is not possible to empirically distinguish "biased beliefs" from "biased choices".

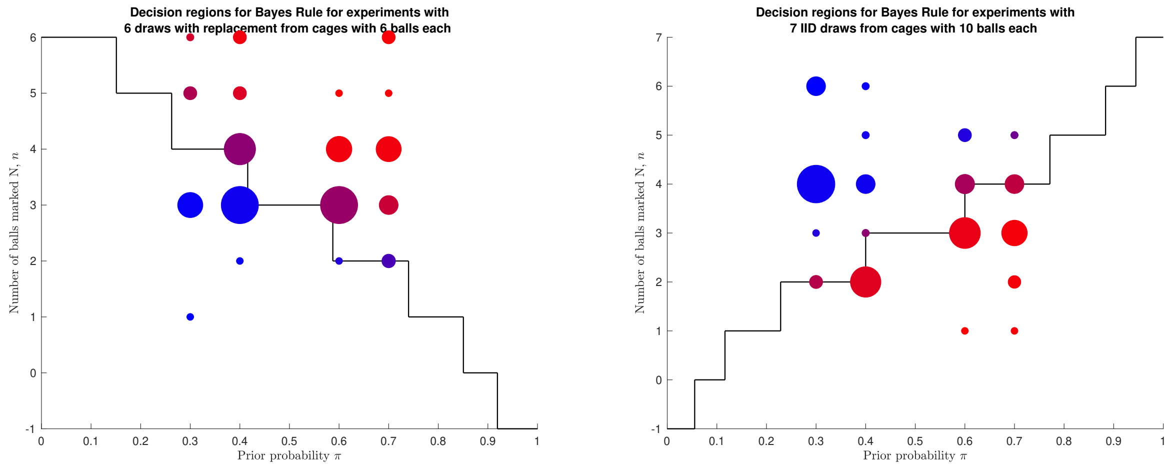
[11]It is not necessary to pre-transform the inputs (d, π, p_A, p_B, D) into $(\text{LPR}(\pi), \text{LLR}(d, p_A, p_B, D))$: additional layers can be added to the neural network so that the inputs to the deeper neural network can enter without any pre-transformation. Then the initial layers of this deeper neural network can be viewed as producing approximations to the transformed inputs $(\text{LPR}(\pi), \text{LLR}(d, p_A, p_B, D))$ that then feed into the two-layer neural net that used the transformed inputs to compute a subjective posterior probability and the top layer producing an output equal to the conditional probability of selecting cage A. These deeper networks require far more parameters, but do not result in substantially better predictions of subject behavior than our parsimonious 4 parameter two-layer neural network specification. Indeed, we can "train" the 4-parameter neural network specification to behave nearly identically to a perfect Bayesian decision maker using training samples with only a dozen observations.

In the original "California" experiments reported in El-Gamal and Grether (1995) there was no variation in (p_A, p_B, D) across subjects or trials: their values were fixed at $(2/3, 1/2, 6)$. In a subsequent set of experiments by El-Gamal and Grether (1999) carried out at the University of Wisconsin, the subjects participated in two different experimental designs on successive days. Half of the 79 subjects were given the "California design" on the first day, and on the second day these same subjects made choices under a new "Wisconsin design" where (p_A, p_B, D) was set to $(.4, .6, 7)$, and the remaining subjects had the Wisconsin design on day 1 and the California design on day 2. This seemingly minor change in the experimental design has a big effect on the Bayesian classification threshold: under the California design larger values of d are evidence in favor of cage A, whereas under the Wisconsin design, larger values of d are evidence in favor of cage B.

Figure 1 summarizes subjects' choices under the two designs, superimposing the Bayes Rule threshold dividing the outcome space (π, d) into regions where it is optimal to choose cage A (red dots) and cage B (blue dots). The size of the dots is proportional to the number of subject-trial observations and their colors are a convex mixture of the colors red and blue in proportion to the share of subjects choosing cages A and B, respectively. The subjects react to the information (p_A, p_B, D) and change their responses in a manner that is consistent with Bayes Rule. The purple dots near the threshold indicates that "confusion" among the subjects is highest in the regions where we would most expect it to occur, i.e. at points where the true posterior probability of cage A close to 1/2. Though it is possible to specify deeper neural nets that allow (p_A, p_B, D) to enter the model directly as separate inputs, the lack of experimental variation in these variables prevents us from estimating and identifying such a model. The "quasi-Bayesian" behavioral model (15) provides a parsimonious and highly flexible way to allow subjects choices to depend on (p_A, p_B, D), and the estimated model accurately predicts how subjects' choices change in response to the change in experimental design illustrated in figure 1.

We estimated the parameters of the model (15) using a panel likelihood function that accounts for the fact that each subject s in the experiment participates in a total of T_s independent trials. we observe a sequence of binary choices d_{ts} and corresponding experimental control variables (π_{ts}, D_{ts}) for each subject s over trials $t = 1, \ldots, T_s$ but where the probabilities (p_A, p_B) remained fixed across trials. Let y_{ts} be a binary indicator of the choice of subject s in trial t: $y_{ts} = 1$ if the subject chose A and $y_{ts} = 0$ otherwise.

Figure 1: Decision Regions for El-Gamal and Grether's California and Wisconsin Designs



The likelihood $L(\theta)$ is given by

$$L(\theta) = \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(A|d_{ts}, \pi_{ts}, p_A, p_B, D_{ts}, \theta)^{y_{ts}} [1 - P(A|d_{ts}, \pi_{ts}, p_A, p_B, D_{ts}, \theta)]^{1-y_{ts}}. \quad (16)$$

3.1 Identification of Beliefs

The unobserved private shocks ε and ν have similar effects on a subject's choice, it is not clear whether it is possible to separately identify the independent effect of each type of shock using data from experiments where subjects provide only binary responses. For this reason we estimated two different restricted versions of the behavioral model: 1) a 3 parameter *structural probit* that restricts $\sigma = 0$ and normalizes the standard deviation of the errors ν to be unity ($\eta = 1$), 2) the 4 parameter *structural logit* model where we restrict $\eta = 0$ and estimate the scale parameter σ of the extreme value distribution of the preference shocks ε.

In experiments 3 and 4 subjects reported their subjective posterior probabilities but did not make binary choices. So, we assume that subjects would select cage A if and only if their reported subjective posterior probability exceeds 1/2. This is equivalent to restricting the output level bias term to 0 and requiring $\sigma/R = 0$. However, we estimate the standard deviation parameter η of the "calculational errors" in their reported subjective posteriors. It is not hard to see from equation (12) that the belief parameters β and η are parametrically identified given sufficient variation in the experimental controls (π, p_A, p_B, D).

Identification is more challenging if we only observe the binary choices of cage A or B, even under the restriction that $\eta = 0$. First, observe that it is impossible to separately identify the reward R and the error or noise parameter σ since it is obvious from formula (15) that these parameters only appear together as a "signal to noise ratio" R/σ. Thus, we assume that the reward R from making a correct decision is known and normalize the payoff to $R = 1$ and estimate only σ subject to this normalization.

When $\sigma = 0$ there is an additional identification problem reflected by the fact that there is a continuum of non-Bayesian posterior beliefs that are consistent with the optimal Bayesian decision rule $\delta^*(d, \pi, p_A, p_B, D)$. To see this consider a family of beliefs indexed by a single parameter $\lambda > 0$ given by $\theta_\lambda = (\beta_\lambda, \sigma, \eta)$ where $\beta_\lambda = (0, \lambda, \lambda)$ and $\eta = 0$ and $\sigma = 0$. For any λ and any information (d, π, p_A, p_B, D) we have

$$\Pi(A|d, \pi, p_A, p_B, D) \leq 1/2 \iff \Pi_s(A|d, \pi, p_A, p_B, D) \leq 1/2, \tag{17}$$

so the subjective beliefs Bayes-consistent, and thus the implied behavior of a non-Bayesian decision maker with parameter $\beta_\lambda = (0, \lambda, \lambda)$ is observationally equivalent to the behavior of a Bayesian.

When there is no decision noise ($\sigma = 0$ and $\eta = 0$), Bayes' consistency implies that the subject's decision rule is optimal even though their beliefs are not Bayesian. Let L be the *Bayesian classification threshold* i.e., the set of all pairs (LLR, LPR) such that $\Pi(A|\text{LLR}, \text{LPR}) = 1/2$: it is the hyperplane given by $0 = \text{LLR} + \text{LPR}$. We can use (13) to derive an equivalent *subjective classification threshold* L_s given by the hyperplane $0 = \beta_0 + \beta_1\text{LLR} + \beta_2\text{LPR}$. Thus, an equivalent statement of Bayes' consistency is that it holds when the *subjective and Bayesian classification thresholds coincide* i.e. if and only if $\beta_0 = 0$ and $\beta_1 = \beta_2$.

When there is decision noise, the restrictions $\beta_0 = 0$ and $\beta_1 = \beta_2$ no longer imply Bayes' consistency, and we will refer to these subjects as *noisy Bayesians* since there are no biases distorting their beliefs, but they do suffer from random decision noise that causes them to make suboptimal decisions. The following lemma shows that a subject's posterior beliefs are identified for any coefficients if $\sigma > 0$ and $\eta = 0$ provided there is sufficient "experimental variation".

Lemma L3. Identification of subject beliefs when $\sigma > 0$. *Assume that $\eta = 0$. When*

$\sigma > 0$, all four parameters $(\beta_0, \beta_1, \beta_2, \sigma)$ of the structural logit model are identified, so subjective beliefs can be recovered from knowledge of the decision rule $P(A|d, \pi, p_A, p_B, D)$, assuming the latter can be identified from sufficient experimental data on the subject's choices.

The proof of Lemma L3 is in Appendix B. Even though Lemma L3 provides a theoretical justification for the identification of the model when $\sigma > 0$, in practice it can be hard to distinguish the decision rule of a subject with Bayesian posterior beliefs where σ takes on relatively large values (i.e., a "noisy Bayesian") from a decision rule of a non-Bayesian who has a very small value of σ but whose β coefficients are also close to zero.[12] Condition (17) can be viewed as a sufficient condition for the optimality of the decision rule of a non-Bayesian subject and it results in a test for a weaker form of Bayesian rationality: $H_o : \beta_0 = 0$ and $\beta_1 = \beta_2$. If this latter hypothesis is satisfied, then the subject will still be modeled as behaving as a "noisy Bayesian" even though their posterior beliefs are not Bayesian. In Section 4.2 we return to the question of inferring subjective posterior beliefs using the directly elicited beliefs provided by subjects in the experiments by Holt and Smith (2009).

3.2 Accounting for Unobserved Subject Heterogeneity

We control for unobserved heterogeneity among subjects using *random coefficients*. Following Kiefer and Wolfowitz (1956) we posit a *distribution* $\mu(\theta)$ of preference parameters θ in the population and attempt to estimate it. Treating μ as an arbitrary element of the space of all distributions over θ results in an infinite dimensional "parameter space" and the estimation problem can be ill-posed unless some restrictions are imposed. Following Heckman and Singer (1984) we estimate a finite mixture approximation to μ by maximum likelihood using a *sieve* (i.e., an expanding parametric family that increases with sample size S and can eventually approximate any μ when S and the number of mixture components is sufficiently large). Given the relatively short panel dimension in these experiments, the sieve estimator was not able to estimate more than 2 or 3 types.[13] Let

[12] Identification using only binary response data is even more challenging when we allow for noise in beliefs by allowing $\eta > 0$ and attempt to estimate σ and η simultaneously. In section 4 we separately estimated the probit specification ($\sigma = 0$ and $\eta > 0$) and the structural logit specification ($\eta = 0$ and $\sigma > 0$) and find the latter provides a significantly better fit by doing a better job of capturing decision errors subjects make in the "hard cases", i.e. where the Bayesian posterior is close to 0 or 1.

[13] Woutersen and Rust (2025) showed that finite mixture models with only a few types can provide highly accurate estimates of the mixed CCP in cases where there are many or even a continuum of types.

K denote the number of unobserved types to be estimated, and $\theta = (\theta_1, \ldots, \theta_K)$ be the $4K \times 1$ vector of parameters of the "mixed structural logit model" and let $\lambda = (\lambda_1, \ldots, \lambda_K)$ be the corresponding $K \times 1$ vector of population probabilities of each of the K types. Then the mixed logit likelihood function, $L(\theta, \lambda)$ is given by

$$L(\theta, \lambda) = \prod_{s=1}^{S} \sum_{k=1}^{K} \lambda_k L_s(\theta_k), \tag{18}$$

where $L_s(\theta_k)$ is the likelihood function for subject s evaluated at θ_k. We estimate a sequence of models starting with $K = 1$ and increasing the number of types K until a log-likelihood ratio test is unable to reject a model with K types in favor of a model with $K + 1$ types.[14]

4 Optimality and revealed beliefs of human subjects

4.1 Reanalysis of California Experiments

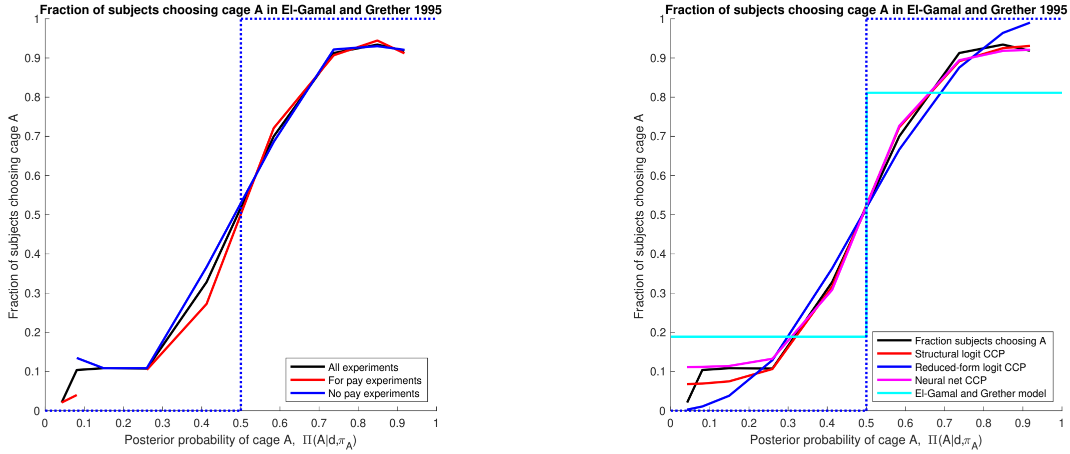
We replicate the results from the California experiments reported in El-Gamal and Grether (1995) using both their original "threshold model" as well our structural logit model.[15] The results are shown graphically in Figure 2 which compares predicted subject choices for several different models and subsamples. In both panels, we plot of the fraction of subjects choosing cage A (y axis) as a function of the Bayesian posterior probability of cage A (x axis). The black lines in both panels are the actual fraction of the 221 subjects who chose cage A in the different trials of the experiment where the values of the "treatment variables" (π, d) are binned so we can plot results on a two-dimensional graph with the Bayesian posterior probabilities, $\Pi(A|\pi, d)$, on the x-axis. The dashed blue line represents the optimal decision rule of a perfect Bayesian decision maker.

The left panel illustrates the effect of the incentive payments on subject behavior (blue curve for the no-pay subjects, red for the subjects who were paid) and it is evident that

[14] We obtain similar results if we choose the value of K with the smallest Akaike Information Criterion (AIC), which equals $2(N_K - L(\hat{\theta}, \hat{\lambda}))$ where N_K is the total number of parameters in the K-type model, is minimized. We also used the *Estimation-Classification* (EC) algorithm of El-Gamal and Grether (1995) to control for heterogeneity. The EC algorithm also maximizes a likelihood function but instead of computing a mixture over types for each subject as in equation (18) the EC algorithm *assigns each subject in the sample their most likely type.* The results from the EC algorithm are similar to those from the finite mixture approach.

[15] We used data for 221 of the 247 subjects reported in El-Gamal and Grether (1995) due to a corrupted data file that made data from 26 subjects from Pomona Community College under the incentivized (i.e., for pay) design unreadable.

Figure 2: Comparison of subject behavior and models in the California experiments



Notes: The left panel displays the fraction of subjects choosing cage A against the posterior probability, $P(A|d,\pi_A)$, from the California experiments reported in El-Gamal and Grether (1995). It includes results from all subject choices, choices in paid experiments only, and choices in non-paid experiments only. The right panel shows the implied conditional choice probabilities (CCPs) derived from estimates of various models: the structural logit model, the reduced-form logit model, the neural net, and the threshold model in El-Gamal and Grether (1995), based on all subject choices in the California experiments. For reference, the CCPs from all subject choices in the right panel are duplicated from the left panel.

it has negligible effect on overall behavior.[16] The maximum likelihood predictions from El-Gamal and Grether's model of subject choice are the cyan colored curve in the right panel. Their model assumes that with probability ε subjects randomly choose cage A or B (with equal probability) and with probability $1-\varepsilon$ they make their choice according to an integer *cutoff rule,* i.e., the subject chooses cage A when $d > c_\pi$ and cage B otherwise, where c_π is one of the 7 integers $\{-1, 0, \ldots, 6\}$ and the subscript π denotes that the cutoffs can depend on the prior π. Note that the optimal Bayesian decision rule takes the form of a cutoff rule: for example when $\pi = 1/3$, the optimal cutoff is $c_{1/3} = 4$, i.e., choose cage A if $d \in \{5, 6\}$ and cage B otherwise. El-Gamal and Grether (1995) found that if they assumed subjects are homogeneous (i.e., all use the same cutoff rule), then Bayes' Rule best describes their behavior in the sense that the cutoffs implied by Bayes' Rule maximized the likelihood function.

It is evident from Figure 2 that their estimated cutoff rule model fails to fit the data well, particularly for the "easy cases", i.e., (π, d) values where the Bayesian posterior probability is near 0 or 1. The model also misses the "'hard cases" where the Bayesian

[16]The average decision efficiency for the 90 subjects in the incentivized trials was 93.5% (std error 0.5%) which is not significantly higher than the 92.3% efficiency of the 132 subjects in the non-incentivized trials (std error 1.5%). We also separately analyzed data from the first and last third of the trials see if there were any substantial "learning by doing" or "experience effects" and these were also negligible.

posterior is near 1/2. This pattern of prediction errors follows from their assumption about subject behavior already discussed, namely that with probability $\sigma = .38$ subjects randomly guess a cage and with probability $1 - \sigma = .62$ the subject follows Bayes' Rule. This implies a discontinuous jump in the predicted probability of selecting cage A right at $\Pi(A|\pi, d) = 1/2$ since at that point the 62% of subjects who are choosing according to Bayes' Rule jump from choosing cage B to choosing cage A.

The right hand panel of Figure 2 plots the predictions from the structural logit model (red curve) as well as several "reduced form" models: 1) a binary logit model with 3 parameters (a constant and two coefficients for π and d), and 2) a 5 parameter two layer neural network that includes an additional bias parameter in the upper output layer (see footnote 10). We can see visually that the structural logit model fits the data significantly better than the El-Gamal and Grether model even though both models have 4 parameters. The El-Gamal and Grether model restricts 3 of the parameters, the cutoffs c_π, to a finite grid of integers, which allows far less flexibility in fitting the data compared to the 4 continuous parameters of the structural logit. The structural logit also outperforms the 3 parameter reduced form logit (which can be regarded as a single layer feedforward neural network), but produces approximately the same predictions as a 5 parameter neural network specification.[17]

Table 1: Log-likelihood values for alternative models of subject choices

Model	Number of Parameters	Log-likelihood	AIC
El-Gamal/Grether discrete cutoff rule	4	-1952	3912
Binary probit	3	-1847	3701
Binary logit	3	-1821	3648
Noisy Bayesian	1	-1801	3604
Structural logit	4	-1773	3554
Neural network	5	-1772	3554

Notes: The table presents the number of parameters, log-likelihood, and the Akaike Information Criterion (AIC) for estimates of various models based on the full sample from the California experiments. We consider the discrete cutoff rule used in El-Gamal and Grether (1995), the structural logit model, the noisy Bayesian model restricting the subjective beliefs in the structural logit model to be Bayesian, the reduced-form logit model, and the neural network.

[17] The structural logit model can be viewed as a restricted 4 parameter version of the 5 parameter neural network where the bias term in the output layer is restricted to be -1/2 times the value of the input weight parameter, which is $1/\sigma$ in the notation of the structural logit model. A likelihood ratio test is unable to reject restriction underlying the structural logit model in equation (15) that subjects choose the cage with the higher payoff, which implies that subjects are equally likely to choose cage A or B when $\Pi_s(A|\pi, d, p_A, p_B, D) = 1/2$. This reflects the identification problem in distinguishing between "biased beliefs" and "biased choices" in Section 3.1.

Table 1 summarizes the fit of the various models of subject behavior. The final column of the table reports the Akaike Information Criterion (AIC) used for model selection and defined as $2(k - LL)$ where k is the number of parameters in the model and LL is the maximized value of the log-likelihood function for that model. Though the 4 parameter El-Gamal and Grether model is not nested as a special case of the 4 parameter structural logit model, using the non-nested likelihood-based specification test of Vuong (1989) we can strongly reject the El-Gamal and Grether model in favor of the structural logit model (P-value 2.5×10^{-4}). The noisy Bayesian model is a restricted 1 parameter version of the structural logit model where we allow σ to be freely estimated and restrict β to impose Bayesian beliefs, i.e., $(\beta_0, \beta_1, \beta_2) = (0, 1, 1)$. A likelihood ratio test strongly rejects the hypothesis that subjects are noisy Bayesians (P-value 9.2×10^{-12}). The structural logit model is a restricted version of the 5 parameter neural network model which allows an extra output layer bias term. Per the comment in footnote 10, the lower layer already accounts for bias via beliefs so an upper layer bias term is superfluous, which explains why a likelihood ratio test fails to reject the structural model (P-value .115) and why it has the same AIC value as the 5 parameter neural network model.

Table 2: Structural logit model

Parameter	σ	η	β_0	β_1	β_2
Estimate	.38	0	.05	2.38	1.86
Standard error	(.02)	(0)	(.05)	(.28)	(.19)

Notes: The table reports the MLE estimates of the single-type structural logit model using subjective choices from the California experiments. We normalize η to be 0 and estimate the remaining belief parameters $(\beta_0, \beta_1, \beta_2)$, and σ .

Table 2 presents the maximum likelihood coefficient estimates for the structural logit model when we assume all subjects are homogeneous. Since the coefficient on $\text{LLR}(d, p_A, p_B, D)$, β_1, is significantly greater than the coefficient on $\text{LPR}(\pi)$, β_2, the estimation results suggest the typical subject in El-Gamal and Grether's California experiments displays the representativeness heuristic. This differs from the conclusion they drew from their integer cutoff rule model that subjects are best described as noisy Bayesians.

Next we show how our conclusions change when we allow for unobserved heterogeneity in subjects' beliefs and behavior. We estimate multiple type models using the finite mixture of types method (hereafter FM) described in Section 3.2. We found that AIC is smaller for a specification with $K = 3$ unobserved types compared to $K = 2$ types or the single type specification presented in Table 2 and likelihood ratio tests strongly reject a model with 1 or 2 types in favor of one with 3 types.

Table 3: FM estimates of the structural logit model for California subjects, $LL = -1655.27$

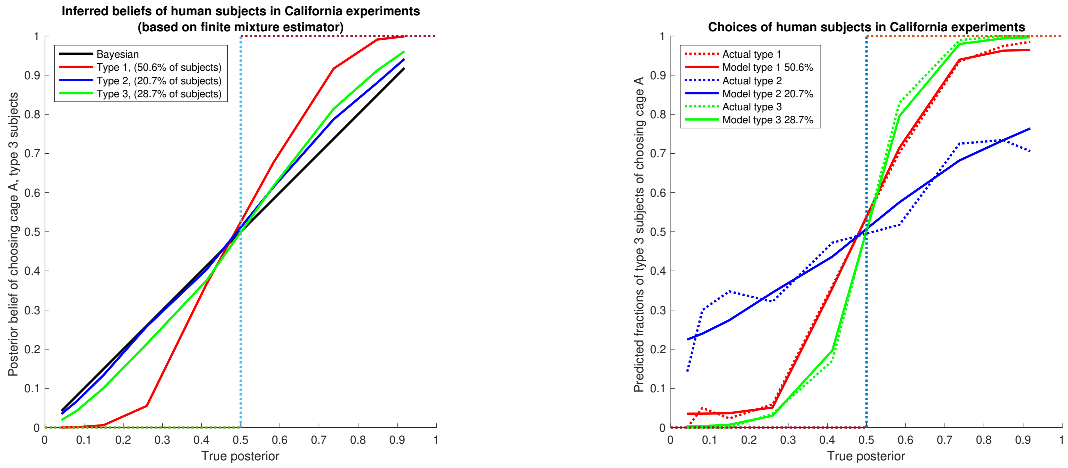
Parameter	Type 1	Type 2	Type 3
σ (noise parameter)	.30	.75	.16
	(.08)	(.37)	(.07)
β_0 (bias/intercept)	.05	-.07	-.05
	(.07)	(.20)	(.05)
β_1 (LLR coefficient)	3.43	.99	1.18
	(.89)	(.77)	(.81)
β_2 (LPR coefficient)	1.58	1.43	1.59
	(.40)	(.53)	(1.01)
λ (population share)	.51	.20	.29
	(.09)	(.10)	(.07)
P-value for $H_o : \beta_0 = 0, \beta_1 = \beta_2$	.002	.59	.16
P-value for $H_o : \beta_0 = 0, \beta_1 = \beta_2 = 1$	.002	.66	.07

Notes: The table presents the maximum likelihood estimates from the FM algorithm for the structural logit model with the constraint $\eta = 0$, using data from the 6-ball design of the California experiments. The standard errors are in parentheses. We stop the FM algorithm when $K = 3$ to avoid overfitting due to limited sample size. In the second to last line, we test whether subjects have Bayesian-consistent beliefs. In the last row, we provide P-value for a test of whether subjects are "noisy Bayesian".

In table 3 we see that 51% of the subjects, the Type 1 subjects, have subjective beliefs consistent with the representativeness heuristic, whereas the remaining subjects can be classified as "noisy Bayesians" who can be further divided into "less noisy" Bayesians (Type 3, 29% of subjects), and "very noisy" Bayesians (Type 2, 20% of subjects). Though the overall fraction of subjects we classify as having approximately "Bayesian" beliefs, 49%, is similar to the 47% that El-Gamal and Grether (1995) classified as Bayesians, we find that a much larger share have beliefs consistent with the representativeness heuristic (51% vs 36%). Though the point estimates of β_1 (LLR) exceed β_2 (LPR) for our type 2 and 3 subjects, we were unable to reject the hypothesis that $\beta_1 = \beta_2$ for both groups. Thus, we did not find any subjects who display "conservatism" (i.e. placing higher weight on the prior), compared to 18% classified as conservatives in El-Gamal and Grether's study.

The left panel of figure 3 compares the estimated or "revealed posterior beliefs" of the 3 different types of subjects to the Bayesian posterior (45 degree line). The right panel compares the implied decision rules (probability of choosing cage A) to the decision rule of a Bayesian decision maker (verical blue dashed line). It is evident that the subjective

Figure 3: Inferred posterior beliefs and predicted vs actual choices of California subjects



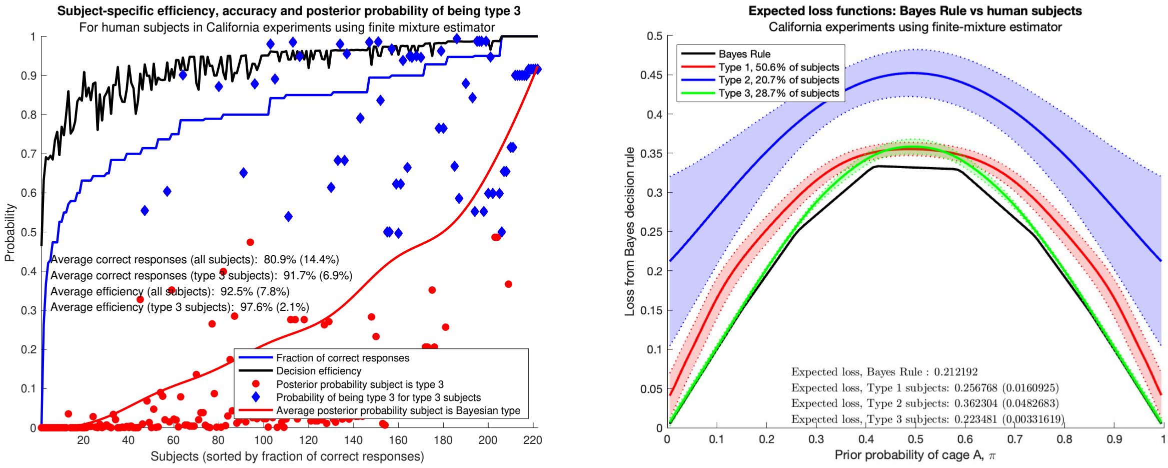
Notes: The left panel shows the implied subjective posteriors, $\Pi_s(A|d, \pi, p_A, p_B, D)$, based on estimates from the three-type structural logit model using subjective choices from the California experiments. These are compared to the true posterior probability, calculated using experimental controls and the realized sample via Bayes' Rule. The model identifies three types of subjects, so we plot each type separately against the true posterior. A 45-degree line representing the true posterior is also included as a reference. The right panel plots predicted vs actual probabilities of choosing cage A for each subject type. The solid lines are the model predictions, the dashed is the fraction of subjects of each type who choose cage A, and the vertical blue dashed line is the choice probability of a perfect Bayesian decision maker, i.e. choose cage A when the true posterior $\Pi(A)$ exceeds 1/2.

posterior beliefs of the noisy Bayesian subjects is close to the Bayesian posterior, whereas the type 1 subjects have distorted beliefs, overestimating the probability of A when the true posterior is above 1/2 and underestimating it otherwise.

The right panel illustrates how belief bias and decision noise contribute to suboptimal decision making, and the point that subjects who are "more Bayesian" are not necessarily more efficient. Recall that the vertical dashed line is the optimal cutoff rule for a perfect Bayesian decision maker, so decision rules that are closer this line have higher win probability and higher overall efficiency. It is evident that green curve plotting the decision rule of the Type 3 least noisy Bayesian subjects is the closest to optimal, whereas the blue line for the Type 2 most noisy Bayesians is the furthest from optimality. In particular, due to the lower estimated decision noise parameter σ, the type 1 subjects who place too much weight on the data (LLR) relative to prior information (LPR) outperform the noisier Type 2 Bayesian subjects, though not the least noisy ones.

The left panel of Figure 4 plots subject-specific accuracy and efficiency scores as well as the posterior probability that the subject is the "Bayesian type" implied by each subject's choices and the estimated structural logit model. The accuracy score is the fraction of

Figure 4: Accuracy, efficiency and loss functions for California subjects



Notes: The left panel displays various performance measures for subjects in the binary decision tasks from the California experiment. The blue line shows the fraction of correct responses (accuracy) for each subject on the x-axis, sorted by their accuracy. The black, wavy line represents the decision efficiency ω_P defined in Section 2 for each subject. The finite-mixture estimates identify three types of subjects: those with representative bias (type 1), noisy Bayesians with larger decision noise (type 2), and noisy Bayesians with smaller decision noise (type 3), considered the "most Bayesian" type. The red dots indicate the posterior probability of each subject being a type 3 (least noisy Bayesian), calculated using equation (20), except we use blue diamonds to identify the type 3 subjects (i.e. those for whom the posterior probability of being a type 3 given their choices, per equation (20) is the highest). A local linear regression estimates the average posterior probability of a subject being type 3, shown by the red curve. The right panel shows the type-specific expected loss associated with its standard errors conditional on the prior π, by integrating out d from the loss function defined in equation 2.

each subject's choices that coincide with the choices of a Bayesian decision maker. The efficiency score is the sum of expected wins in the T_s trials each subject s participated in to the corresponding wins for a perfect Bayesian, i.e., the ratio ω_s given by

$$\omega_s = \frac{\sum_{t=1}^{T_s} \left[\Pi(A|d_{ts}, \pi_{ts})^{y_{ts}} + [1 - \Pi(A|d_{ts}, \pi_{ts})]^{(1-y_{ts})} \right]}{\sum_{t=1}^{T_s} \left[\Pi(A|d_{ts}, \pi_{ts})^{y^*_{ts}} + [1 - \Pi(A|d_{ts}, \pi_{ts})]^{(1-y^*_{ts})} \right]}, \tag{19}$$

where d_{ts} and π_{ts} are the trial outcomes and priors, respectively, and y_{ts} is an indicator for subject s's choice of cage A in trial t, and y^*_{ts} is the choice a perfect Bayesian would make in the same trial. The red dots are the posterior probabilities that each subject is type 3. We use the estimated probabilities of each type as the "prior probability" and the subject-specific likelihood to compute a posterior probability for each type $\tau \in \{1, 2, 3\}$, denoted by $\Pi(\tau|y_s, d_s, \pi_s)$ and given by

$$\Pi(\tau|y_s, d_s, \pi_s, \hat{\theta}) = \frac{\hat{\lambda}_\tau L(y_s, d_s, \pi_s|\tau, \hat{\theta})}{\sum_{k=1}^{3} \hat{\lambda}_k L(y_s, d_s, \pi_s|k, \hat{\theta})}. \tag{20}$$

where y_s is the sequence of choices by subject s in the T_s trials, and d_s and π_s are the corresponding outcomes and priors for these trials, $\hat{\lambda}_k$ is the estimated fraction of type k subjects, and $L(y_s, d_s, \pi_s|k, \hat{\theta})$ is the subject-specific likelihood for subject s at the estimated parameter values $\hat{\theta}$ assuming the subject is type k. We identified the type τ of each subject in the sample as the value for which the corresponding posterior, $\Pi(\tau|y_s, d_s, \pi_s, \hat{\theta})$ is the largest, and we use the blue diamond symbols to highlight the Type 3 subjects in the overall sample.

It is evident from the left hand panel of Figure 4 that the type 3 subjects generally have significantly higher accuracy and efficiency scores, 92 and 98% respectively, than the average for all subjects (81 and 93%, respectively). The red line in the left panel plots the local average probability that the subject is type 3. We see that this probability is monotonically increasing in the fraction of correct responses (accuracy) and is also strongly positively correlated with subject-specific decision efficiency, though variation across subjects in efficiency is not as great as the variation in accuracy. This is a reflection of the observation we made in the introduction that a subject with lower accuracy need not have significantly lower efficiency if their choices deviate from Bayes' Rule mostly for the "hard cases" where the Bayesian posterior is close to 1/2.

The right panel of Figure 4 plots the loss functions for the three types of subjects as a function of the prior π. We see that the implied loss functions are similar, though the estimated standard error bands are larger for the type 2 subjects (blue line, the noisier subset of noisy Bayesians). We calculated expected win probabilities using the empirical distribution of π for the three subject types and compared them to a Bayesian decision maker who has a 70% average probability of choosing the correct cage in these experiments. The implied efficiency scores for the type 1 representativeness subjects is 94.7% (std error 1.4%), whereas the noisier group of the noisy Bayesians have a relatively low efficiency of only 81.6% (4.4%). The less noisy Bayesians have the highest efficiency: 98% (0.9%). The average efficiency of all three types of subjects is surprisingly high, 93% even though only a minority have beliefs that are well approximated by Bayes' Rule.

Thus, we have shown that subjects who are "more Bayesian" in their subjective beliefs are not necessarily better decision makers: the degree of "decision noise" also contributes significantly to suboptimal behavior, especially when it causes mistakes in the "easy cases" where the true posterior is close to 0 or 1.[18]

[18] We also re-analyzed the experimental data collected in El-Gamal and Grether (1999) which involved a

4.2 Reanalysis of Holt and Smith Experiments

The identification problem for subjective beliefs using only binary choice data discussed in Section 3.1 suggests the need for caution in drawing conclusions about the fraction of subjects who have subjective posterior beliefs that are well approximated by Bayes' Rule, though we can be confident our inferences on overall efficiency of human subjects since this measure is based on the CCP which is non-parametrically identified.

In this section we reanalyze experiments reported in Holt and Smith (2009) that directly elicited subjective posterior beliefs. This was done via the *Becker-DeGroot-Marshak* (BDM) mechanism which incentivizes rational subjects to truthfully report their subjective posterior probabilities. They conducted two separate experiments: one at Holt's laboratory at the University of Virginia involving 22 subjects, and a second one done via the Internet involving 30 subjects. In both experiments the design parameters $p_A = 2/3$ and $p_B = 1/3$ were fixed but the number of draws D and the priors varied across multiple trials for each subject. D took values from $\{0, 1, 2, 3, 4\}$ and π varied from $\{1/3, 1/2, 2/3\}$. Here we focus on the reanalysis of the first experiment with 22 subjects and, given space constraints, summarize the key findings from our analysis of their second web-based experiment in a footnote.

The BDM mechanism was implemented as follows: after seeing the prior π and the result of the random drawing d from the selected cage/cup, subjects were asked to report a probability $p_r \in [0, 1]$ that determines their payoff from a second stage lottery. This gamble, denoted by $\tilde{G}_R$, involves drawing a random probability $\tilde{p} \sim U(0, 1)$ and paying the subject a monetary reward of R according to the following rule: if $\tilde{p} < p_r$ the subject receives R if the observed sample was drawn from cup A, otherwise if $\tilde{p} \geq p_r$ the subject receives R with probability $\tilde{p}$. The subject's expected payoff from reporting p_r in this second stage BDM lottery is

$$E\{\tilde{G}_R | p_r, d, \pi, p_A, p_B, D\} = R \left[\frac{1 - p_r^2}{2} + p_r \Pi_s(A | d, \pi, p_A, p_B, D) \right], \tag{21}$$

switch in experimental design to test how subjects change their decision rules in response to fixed experimental design parameters (p_A, p_B, D) as discussed in section 3. The conclusions from our reanalysis of the Wisconsin subjects are broadly similar to the conclusions noted above from our reanalysis of the California subjects. In addition, the behavioral model is able to make accurate out-of-sample predictions of how subjects' responses changed in response to the change in the design parameters (p_A, p_B, D), showing that the behavioral model (15) is "structurally stable" and provides a good yet parsimonious representation of subject behavior. Detailed results of this reanalysis are available on request.

where $\Pi_s(A|d,\pi,p_A,p_B,D)$ is the subjective posterior probability for cup A. It follows that the report p_r that maximizes expected payoff is $p_r^* = \Pi_s(A|d,\pi,p_A,p_B,D)$.

A drawback of the BDM mechanism is that it can be confusing to subjects and potentially harder for them to determine the optimal report p_r than to determine the posterior probability of cup A. Holt (2019) notes that "The use of incentivized elicitation procedures is the norm in research experiments, but there are some problems." one of which is that BDM relies heavily on the presumption of rationality of the human subjects, including the ability to derive the expected payoff function (21), optimize it, and realize that the payoff maximizing report is $p_r^* = \Pi_s(A|d,\pi,p_A,p_B,D)$. If subjects can not do this extra layer of math, the second stage BDM mechanism might actually mislead or confuse them and therefore add extra "decision noise" into experimental outcomes.[19]

Subjects in Holt and Smith's experiments were not asked to make an additional binary choice of which cup they believed the observed sample was more likely to have been drawn from. It seems quite reasonable to assume that if subjects would have been asked to make such a choice (perhaps incentivized by an additional payment for selecting the correct cup), they would have chosen cup A if $\Pi_s(A|d,\pi,p_A,p_B,D) > 1/2$ and cup B otherwise.[20] This implies that $\sigma = 0$ in our structural logit model specification in equation (14), and we use this to generate the implied decision rule. Any inefficiency in subjects' decision making is then due to "calculational noise" ν and bias in subjective posterior beliefs.

We estimated β and the parameter η under the assumption that $\nu \sim N(0,\eta^2)$ by maximum likelihood using the log reported posterior odds ratio regression specification in equation (12).[21] We also estimate multi-type versions of these models using the EC algorithm and finite mixture approaches. We find a strong improvement in the likelihood from going from 1 to 2 types, but we stopped at $K = 2$ types because of the relatively small number of subjects (22 and 24 in experiments 1 and 2, respectively).[22] We omit the

[19] Holt (2019) acknowledges that the "BDM procedures may be difficult for subjects to comprehend." (p. 110). The instructions to subjects in the Holt and Smith (2009) experiments instructed them that their payoff is maximized by truthfully reporting their subjective posterior. To the extent subjects trusted and followed this advice the BDM mechanism may not necessarily confuse subjects or add extra "decision noise".

[20] Or randomly guess if $\Pi_s(A|d,\pi,p_A,p_B,D) = 1/2$.

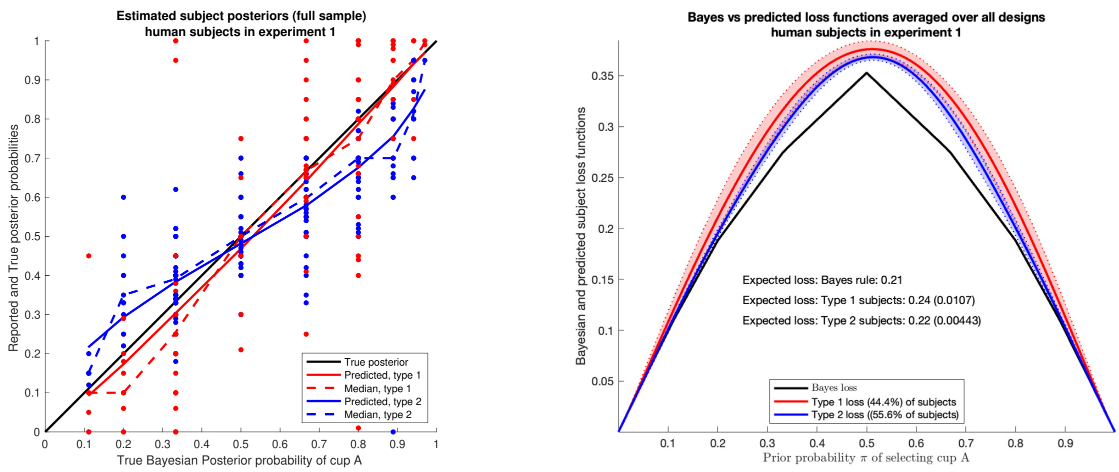
[21] A small fraction of subjects reported posterior probabilities of 0 or 1 for which the log reported posterior odds ratio is undefined. Rather than exclude these observations we estimate a truncated regression specification where we assume that a value of 0 is reported when the subjective posterior is lower than some lower threshold $\underline{p}$ and report a value of 1 when it exceeds an upper threshold $\overline{p}$. It is not hard to show that the maximum likelihood estimates of these additional parameters are the min and max of the subset of reported subjective posterior values that are strictly in the $(0,1)$ interval. We verify that all conclusions are robust to simply excluding the observations with reports of 0 or 1, or recoding them to arbitrary values such as .00001 and .99999.

[22] The results from the EC and finite mixture models are quite similar, both in the parameters and the estimated

actual parameter values and describe the results more informally and graphically below.

For experiment 1 the two types can be described as 1) noisy Bayesians (45% of the subjects) and 2) conservatives, i.e., those who put more weight on LPR than LLR. We classify type 1 subjects as noisy Bayesians since we cannot reject the hypothesis that $(\beta_0^*, \beta_1^*, \beta_2^*) = (0, 1, 1)$ but their reports are subject to significant noise, reflected in the large and significant estimate of $\hat{\eta} = 0.91$ (std error (0.14)). Though we observe significant bias in the reported beliefs of the type 2 subjects, their reports are far less noisy with an estimated value of $\hat{\eta} = 0.40$ that is less than half the value we estimate for the noisy Bayesians.

Figure 5: Predicted vs Actual Median Beliefs and Loss Functions: Holt Smith Experiment 1



Notes: We estimate equation 12 for two types of subjects using the reported prior odds ratios from the Virginia experiments (Holt and Smith, 2009), via maximum likelihood. Subjects are classified into the two types based on the estimated parameters and equation 20. The left panel shows a scatterplot of reported posterior probabilities versus true posterior probabilities, separated by type. We also compare the actual median reported probabilities in the data (dash lines) with the model-predicted probabilities (solid lines) for each type. The right panel presents the expected loss functions for the two types, along with their 95% confidence intervals. For comparison purpose, we also include the expected loss under the Bayesian decision rule.

The left hand panel of Figure 5 provides a scatterplot of the subject responses plotted as (LPR, LLR) pairs for each subject and trial. For reference the black 45 degree line is the Bayesian posterior probability and the dashed lines are the median values of the subjects' responses. Using the posterior probability in equation (20) we can classify each subject according to whether which posterior probability is more likely: either type 1 (noisy Bayesian) or 2 (conservative). The solid red and blue lines are the predicted medians of subject responses from the estimated structural logit model. We see that the model fits the data well and the type 1 subjects have median posterior beliefs that are quite close

fractions of each type.

to the 45 degree line. However, the median beliefs of the type 2 conservative subjects increase less steeply than the Bayesian posterior does, reflecting "underconfidence". While these beliefs are not Bayesian, they are Bayes' compatible according to our definition in Section 3.1, and the type 2 subjects are actually slightly more efficient than the type 1 subjects (97% and 95%, respectively).

This conclusion is verified in the right hand panel of Figure 5 which plots the implied loss functions. The higher level of decision noise (larger value of η) for the type 1 (noisy Bayesian) subjects results in a lower expected win probability compared to the type 2 conservative subjects (blue line). The average efficiency of both types of subjects is relatively high: 96% with a standard error of 0.7%.[23] Nevertheless, a Wald test strongly rejects the hypothesis that the human subjects are fully efficient decision makers. In summary, despite the non-Bayesian beliefs of the type 2 subjects, overall efficiency of these subjects is high and in line with what we find for subjects in El-Gamal and Grether's California and Wisconsin experiments.

5 Optimality and revealed beliefs of AI subjects

To compare the performance of AI and human subjects, we conduct new experiments, *using AI subjects,* with the same design as the human experiments by El-Gamal and Grether and Holt and Smith. We used three versions of the ChatGPT from OpenAI to form our pool of AI subjects: GPT-3.5 (introduced in 2022), GPT-4 (introduced in 2023), and GPT-4o (introduced in 2024) to assess the degree of progress in general-purpose AI capabilities over a short window of time.

We construct our prompts to follow precisely the instructions used in the original human experiments, ensuring that each contains all necessary information for GPT to generate responses based on the realized outcomes of those experiments (e.g., the composition of the cages, prior selection, and the number of draws from the chosen cage). Specifically, we replicate the 6 ball Wisconsin experiments reported by El-Gamal and

[23] We also analyzed data from 30 subjects in Holt and Smith's experiment 2 which was conducted online. The overall conclusions are similar to those from our reanalysis of experiment 1, except that the EC algorithm no longer finds any noisy Bayesians: 62% of subjects put excessive weight on the prior and the remaining 38% put too much weight on the data. The level of calculational noise for these subjects, η, is also significantly higher. The higher degree of noise in subjects' reports implies significantly higher loss and thus lower efficiency. Average decision efficiency for all subjects in all trials in experiment 2 was 91% (0.8%), lower than the 93% efficiency of the subjects in El-Gamal and Grether's 6 ball California experiments and lower than the 96% efficiency of human subjects in Holt and Smith's experiment 1.

Grether (1999) and the design of Holt and Smith (2009) where subjects were asked to report their subjective posterior probability.

We employ the structural logit model introduced in Section 3 to estimate GPT subjects' beliefs and decision noise. The importance of recovering these beliefs follows from Lemma L2, which shows that a decision rule can be optimal for all payoff functions only if the underlying belief is Bayesian. Because LLM end users are highly heterogeneous and have diverse and typically unknown payoff structures, identifying the model's beliefs is essential for assessing the efficiency of its decisions. We then use these estimates to quantify the decision efficiency of LLMs.

As with the human subjects, we also allow for *unobserved* heterogeneity among the GPT subjects. Why might we expect such heterogeneity? In fact, we intentionally introduce it by varying the *temperature parameter*, which governs the degree of randomness ("decision noise") in GPT's responses. The temperature parameter plays a role analogous to the extreme-value scale parameter σ in equation (15), or to the variance parameter η governing the "calculational errors" ν in the subjective posterior probability $\Pi_s(A|d, \pi, p_A, p_B, D, \nu)$ in equation (12).[24] The EC and finite mixture methods capture the variability in temperature, even though we do not include it explicitly as an observable in the estimation. This allows us to examine how decision noise influences GPT performance and to assess the effectiveness of the EC and FM estimators in accounting for such variability.

5.1 Prompt Design for Experiments in ChatGPT

We collect data from GPT subjects by submitting inquiries through the public OpenAI application programming interface (API), following the approach of Chen et al. (2023). Each GPT subject is matched to a corresponding participant in the human experiment and receives the identical experimental information. This design allows us to precisely control the inputs, ensuring an exact match between the information provided to human and GPT subjects. Appendix C details the algorithm used to implement the experiments, which iterates over different GPT versions, temperature settings (subjects), experimental

[24]Lower temperature values yield more deterministic and focused outputs, making GPT more likely to select the most probable next token. Higher temperatures increase stochasticity, producing more variable responses. The default temperature for the three GPT versions considered here is 0.7. In our experiments, we randomly varied the temperature across subjects from 0.01 to 1.2.

designs, and trials. We then parse the responses from ChatGPT and collected its choices of either Cage A or B in the Wisconsin experiment,[25] or the reported subjective probability in the case of the Holt and Smith experiment. Appendix D provides an example of the prompts we used.[26]

We allow GPT to display its full reasoning process rather than restricting it to a final answer, as this reflects how GPT is typically used in practice.[27] An additional advantage of allowing GPT to provide its full reasoning is that it enables us to examine its underlying "thought process," including its understanding of probability theory and its recognition of when Bayes' Rule applies. We analyze GPT's textual responses in Section 6.

5.2 Analysis of GPT subjects in the Wisconsin experiments

We begin by presenting the binary classification results from the 6-ball Wisconsin experiment using GPT subjects.[28] Figure 6 plots the estimated β's by displaying the corresponding classification hyperplanes for each subject type. The FM algorithm identifies two types among GPT-3.5 subjects, but only a single type for GPT-4 and GPT-4o. Panel (a) shows that the estimated hyperplanes for both GPT-3.5 types are substantially flatter than the Bayesian hyperplane, indicating that both types exhibit representativeness. This is also confirmed by our textual analysis in Section 6. Panels (b) and (c) show that the beliefs of GPT-4 and GPT-4o are both very close to Bayesian.

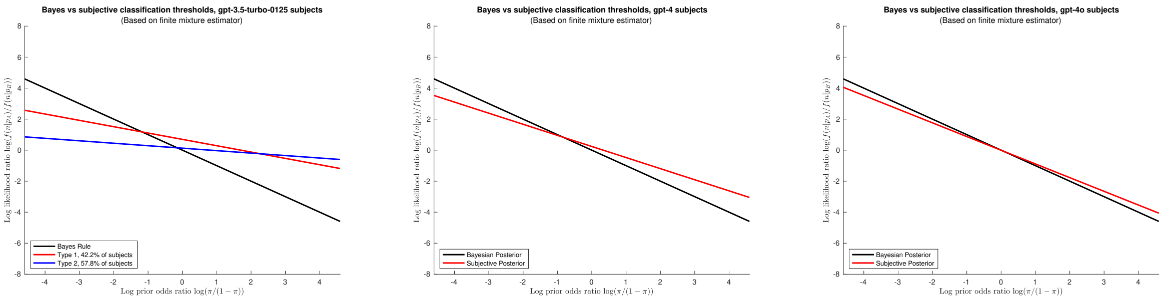
[25] Occasionally, GPTs may stop prematurely before providing an answer regarding the choice between A and B. In such cases, we resubmit the same inquiry until the GPT delivers a classification. We consider this process to be natural, as it mirrors our everyday use of ChatGPT—if it fails to provide a satisfactory answer due to an unexpected stop, we simply ask again. In our implementation, it takes a maximum of 5 iterations to resolve any missing answers in our experiment.

[26] Although we replicate the same questions with identical wording as those used in the human experiments reported in El-Gamal and Grether (1995), El-Gamal and Grether (1999) and Holt and Smith (2009), concerns remain that the performance of LLMs may be sensitive to prompt variations. To address this, we conduct a parallel experiment using a "pond and fish" context, which differs substantially from the traditional balls-and-urns framing. We find that our main conclusions are robust under this alternative setting. In addition, we explore other prompt formats, including versions that allow for chain-of-thought reasoning (Wei et al., 2022) and others that suppress intermediate steps. Across these variations, the central results remain stable.

[27] We also conduct experiments that suppress the reasoning process, following Chen et al. (2023), by adding the instruction "Please only tell ..." to the prompts. Although shutting down the reasoning process may deteriorate performance, our main results are robust to this change in the prompt.

[28] Similar results are obtained from the GPT replication of the 7-ball Wisconsin experiment.

Figure 6: Estimated classification hyperplanes for GPT Subjects: 6-ball Experiments



(a) GPT-3.5 (b) GPT-4 (c) GPT-4o

Notes: The three panels plot the implied classification hyperplanes for each type identified from the estimates of the multi-type structural logit model using finite mixture methods. In panel (a), the estimates are based on the 6-ball design from the Wisconsin experiments using GPT-3.5, in panel (b) using GPT-4, and in panel (c) using GPT-4o. The black lines represent the Bayesian classification hyperplane, as a reference.

All versions of GPT exhibit substantial decision noise, although GPT-4o subjects display slightly less noise than GPT-4 and considerably less noise than GPT-3.5. Consequently, GPT-4 and GPT-4o subjects are better described as "noisy Bayesians" rather than as "perfect Bayesians". For GPT-3.5, type 1 subjects are considerably less noisy than type 2 subjects. The mean temperature for type 1 is 0.43, roughly half that of the noisier type 2 group. Since higher temperatures correspond to greater randomness in GPT responses, the association between temperature and estimated noise provides evidence that the FM algorithm effectively detects unobserved heterogeneity.

Table 4 compares the performance of humans and GPT subjects. We report both decision efficiency and accuracy, averaging across all estimated types weighted by their population shares. Humans are remarkably efficient, achieving 96.5% of the payoff of a perfectly Bayesian decision maker, even though their choices coincide with the Bayesian decision rule in only 81.9% of cases. Relative to the GPTs, humans outperform GPT-3.5 and GPT-4 on average, although their performance falls short of GPT-4o, which attains an average efficiency of 97.5%. Despite this, the most efficient human subjects (type 1) achieve a decision efficiency of 99.5%, exceeding the average performance of GPT-4o.

Comparing different versions of GPTs, we observe a convergence from two distinct types, each exhibiting substantial representative biases, to a single type that behaves as a noisy Bayesian. Both decision efficiency and accuracy highlight the rapid shift in GPT performance from "subhuman" to "superhuman" levels with only a few model upgrades over a relatively short period.

Table 4: Comparison of Humans and GPTs in the 6-Ball Wisconsin Experiment

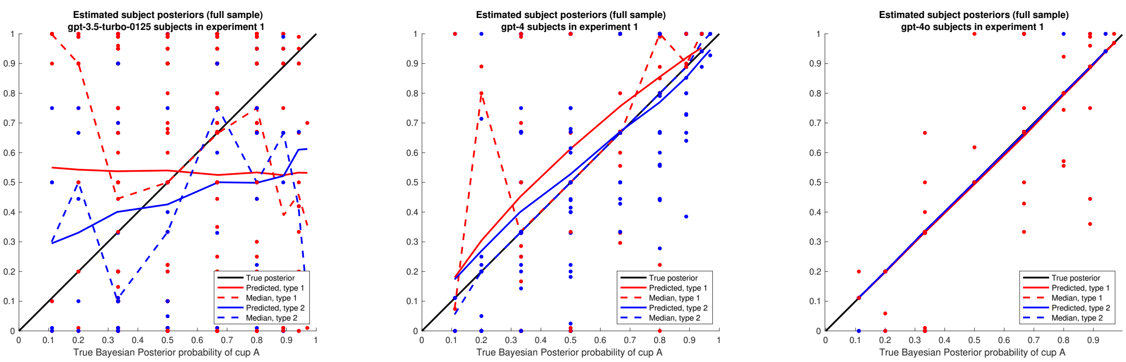
	GPT-3.5	GPT-4	GPT-4o	Humans
Efficiency	84.9	96.0	97.5	96.5
	(0.6%)	(0.3%)	(0.2%)	(0.5%)
Accuracy	59.4	77.7	85.5	81.9
	(1.0%)	(0.8%)	(0.7%)	(1.0%)
No. of Types	2	1	1	3

Notes: The table summarizes the overall performance of AI and human subjects in the 6-ball design of the Wisconsin experiments. It presents the average accuracy and decision efficiency for all subject types, weighted by their population shares. Standard errors are shown in parentheses. Efficiency is derived from the estimated multi-type structural logit model using the finite-mixture method, along with the population shares. The number of types detected by the finite mixture method is reported at the bottom of the table.

5.3 Analysis of GPT subjects in the Holt and Smith experiments

In this section, we analyze the elicited posteriors from the replication of the Holt and Smith experiment 1 using GPT subjects.[29] The FM algorithm identifies two types for all versions of GPTs.[30] Same as Figure 5 from the re-analysis of human experiments, we compare posteriors reported by GPT subjects and estimated subjective posteriors for each type in Figure 7.

Figure 7: True vs Estimated Subjective Posterior Probabilities



(a) GPT-3.5 (b) GPT-4 (c) GPT-4o

Notes: This figure is analogous to Figure 5, but for GPT subjects, with one panel for each version of GPTs. Using the reported prior odds from each version of GPT subjects in the replication of Virginia experiments, we estimate Equation (12) for up to two types via maximum likelihood. Subjects are then classified into the two types based on the estimated parameters and Equation (20). Each panel shows a scatterplot of reported versus true posterior probabilities, separated by type. We also compare the actual median reported probabilities in the data (dashed lines) with the model-predicted probabilities (solid lines) for each type. The black 45-degree lines represent the true Bayesian posterior probability for reference.

[29] We also replicate and analyze data from the online experiment 2 reported in Holt and Smith (2009), applying it to GPT subjects. The results are similar to those of experiment 1.

[30] Same as in the re-analysis of human experiments, we limit the FM algorithm up to two types to avoid overfitting. Likelihood ratio tests reject the single-type model for all three versions of GPTs, leading us to conclude with two types for each version.

Panel (a) shows large biases for GPT-3.5 near the extremes: the median type 1 subject reports a posterior of 1 when the true posterior is 0.1, and the median type 2 subject reports a posterior of 0.1 when the true posterior is 1. Because these are "easy cases," such extreme errors lead to substantial efficiency losses. Human subjects are not fully optimal, but their reported posteriors generally increase with the true Bayesian posterior. In contrast, GPT-3.5 often fails even to produce a monotonic mapping, a consequence of its failure to employ Bayes Rule to calculate the probability it reports.[31]

We observe a substantial improvement in GPT-4, as its reported subjective posteriors align much more closely with the Bayesian reference line. The posteriors of type 2 subjects nearly coincide with those of Bayesian decision makers. Type 1 subjects, however, remain less consistent with Bayesian predictions, showing a noticeable bias toward Cage A at true posteriors of 0.2 and 0.8. This bias disappears in GPT-4o, where the subjective posteriors of both types almost coincide with the true Bayesian posteriors.

The estimates of the noise parameter η reveal another aspect of improvement: the level of calculation noise in their subjective beliefs decreases with each successive generation. The estimate $\widehat{\eta}$ decreases from 2.8 and 2.1 for the two types of GPT-3.5 subjects to 0.4 and 0.0025 for the two types of GPT-4o subjects.

Table 5 summarizes the performance of human and GPT subjects in Holt and Smith's experiment 1. In addition to decision efficiency and accuracy, the table reports the R^2 from regressing the reported posterior on a constant and the true posterior. Higher values represent behavior that is closer to Bayesian. Consistent with the Wisconsin experiments, we observe a sharp performance jump from GPT-3.5 to GPT-4 across all metrics. Humans achieve roughly 3 percent point higher decision efficiency than GPT-4, and GPT-4o exceeds GPT-4 by about 3.4 percentage points. The improvement across successive GPT versions is more pronounced in the Holt–Smith experiment than in the Wisconsin experiment.

[31]We also find that the average reported posterior probabilities for GPT-3.5 are not increasing in either the LPR or the LLR. This non-monotonicity, which indicates limited rationality in its decisions, can reduce the structural logit model's ability to fit the data, since the model implicitly imposes monotonicity in both LPR and LLR. We could extend the model to allow non-monotonic patterns by including higher-order terms of the LLR and LPR. However, given the limited number of observations, such flexible specifications may lead to overfitting. For this reason, we retain the simple structural logit model.

Table 5: Performance of humans and GPT in the Holt-Smith Experiments

	GPT-3.5	GPT-4	GPT-4o	Humans
Efficiency	75.0	93.0	99.4	96.0
	(1.0%)	(0.4%)	(0.3%)	(0.7%)
Accuracy	58.1	84.0	98.2	87.4
	(2.3%)	(1.7%)	(0.6%)	(1.5%)
R^2	0.7%	41.8%	88.0%	63.5%
No. of Types	2	2	2	2

Notes: The table summarizes the overall performance of human subjects in the first day of experiments at the University of Virginia, as well as various GPT subjects replicating those experiments. It presents the average accuracy and decision efficiency for all subject types, weighted by their population shares. Standard errors are shown in parentheses. Efficiency is derived from the estimated multi-type structural logit model using the finite-mixture method, along with the population shares. The second to last row presents the R^2 from a simple regression of the reported posterior on a constant and the true posteriors. The last row reports the number of types detected by the finite mixture method.

6 Analysis of Errors from AI Subject Response Text

Most econometric models treat humans as black boxes and apart from some research from neuroscience, we know little about how humans process information. However, using the textual responses of GPT, we have the unique advantage of observing the reasoning of GPT subjects, opening the door to analyze where GPTs make mistakes. There are two challenges in such an analysis: the errors made by GPTs can be highly diverse, and the textual responses are not well-structured.

We overcome the first challenge by exploiting the simple structure of the binary decision problem, which allows us to classify errors into nine binary error flags under four broad categories. To obtain a distribution of GPT errors across categories, we then develop a GPT grader to efficiently process large-scale unstructured responses and determine the value of error flags within each category. Key inputs for the GPT grader include a reference answer using Bayes' rule, detailed grading rubrics, original experiment prompts and responses. We present our grading prompt in Appendix E.

We use a more advanced version, GPT-o3-mini, to grade three less advanced models considered, ensuring the grader has superior performance in the binary classification tasks and general intelligence. We also manually review 50 randomly sampled textual responses from each model to verify the GPT grader's performance. Overall, our independent cross-check confirms the accuracy of GPT-o3-mini's grading and classification of errors in the responses from ChatGPT 3.5, 4 and 4o.[32]

6.1 Error Taxonomy

We focus on the textual responses from the 6-ball Wisconsin experiment. Table 6 presents the four broad categories and the error flags under each category.[33] The first type of error under Panel A examines whether GPTs understand the context and correctly interpret experimental parameters, including cage composition (i.e., the number of N balls in each cage), sample size (D), and the observed number of N balls (d).

The second category of errors concern whether GPT subjects are "conceptually" Bayesian by checking whether they take the prior information into consideration and if they use the sample information as inputs to their decision. Failure to use prior or sample information can explain choices that are consistent with representativeness or conservatism. We emphasize the *consideration* of both prior and likelihood in the second category, leaving the examination of *numerically correct* posterior calculation to the third category under Panel C. We break down the calculation of posteriors into three components: the prior probability,[34] the likelihood, and the posterior (or posterior odds ratio).[35]

The final category in Panel D evaluates whether the final decision (Cage A or B) aligns with the preceding reasoning. Such errors are captured by the decision noise term ε in the structural logit model. Allowing for these errors is important because chain-of-thought output is not always faithful and may not be consistent with the final answer reported by GPTs (Turpin et al., 2023; Lanham et al., 2023). We instruct the GPT grader to read

[32] In Appendix G, we report the error rates for the same set of 50 samples graded by both GPT-o3-mini and a human grader. Additionally, we include the grading results using the most advanced version of GPT-o1, which, while offering slightly superior performance, is significantly more costly.

[33] To demonstrate the error flags, we provide excerpts from textual responses as example answers classified under each error flag in Appendix F.

[34] To calculate the prior probability, subjects must understand the process of rolling a 10-sided die and divide the specified range of results for Cage A by 10. While this is straightforward for humans, we occasionally find that GPT subjects use an incorrect numerator or denominator when calculating the prior probability for Cage A. See Appendix F for an example.

[35] A few details are worth discussing regarding the grading rubric for Panel C. First, we explicitly instructed the GPT grader to allow for the omission of binomial coefficients when calculating the likelihood, as they will cancel out and do not affect the posterior calculation. Second, we find that some GPT subjects, instead of calculating the posterior probability $\Pi(A|d, \pi, p_A, p_B, D)$ as in equation (1), make decisions by calculating and comparing the product of prior and likelihood, $\pi \times f(d|p_A, D)$ and $(1 - \pi) \times f(d|p_B, D)$. This approach is consistent with Bayes' rule, and in such cases, the subject passes error flag 8. Third, similar to the identification challenge discussed in 3.1, it is sufficient for subjects to make correct decisions if the posterior is on the same side of $1/2$ as the Bayesian posterior. However, since GPT subjects usually report the posterior probabilities, we apply a stricter grading rubric by marking the subject as making an error in flag 8 if their calculated posterior (or posterior odds) was incorrect regardless of whether it leads to the same decision as the true posterior.

the overall reasoning of the textual responses and then predict the expected outcome based on the reasoning flow. Almost all GPT-4 and 4o subjects answer the problem by calculating the posterior or posterior odds. In such cases, we define error flag 9, final decision contradicting the previous reasoning, as choosing cage A when the posterior of Cage A is below $\frac{1}{2}$, or when $\pi \times f(d|p_A, D)$ is smaller than $(1-\pi) \times f(d|p_B, D)$. In other cases, we ask the GPT grader to explicitly predict the outcome based on the reasoning just before the final answer and compare whether this prediction is consistent with the subjects' reports.

6.2 Grading Results

We apply the grading prompt to evaluate 500 randomly selected text responses for each GPT model, regardless of whether the final decision aligns with Bayes' rule. The output of the grading algorithm assigns a value zero or one to each error flag, with one indicating that the student's response contains a mistake. Table 6 presents the error rates for nine types of errors. We also report the fraction of responses inconsistent with the Bayes' rule at the bottom of the Table.[36]

[36]The sum of error rates across all categories doesn't necessarily match the fraction of incorrect responses for two reasons. First, errors are not mutually exclusive. If a subject ignores the prior (error flag 4 = 1), they will also have errors in calculating the prior (error flag 6 = 1) and subsequently the posterior (error flag 8 = 1). Second, since the final decision's consistency with Bayes' rule only requires comparing the posterior to $\frac{1}{2}$, an error in calculating the posterior does not necessarily lead to a mistake in the final decision.

Table 6: Error Distribution from GPT Responses

	Error Flag (Yes/No)	GPT-3.5 (%)	GPT-4 (%)	GPT-4o (%)
Panel A. Data read-in errors				
1	Error reading the compositions of the two cages	0.4	0.4	0.0
2	Error reading the number of balls drawn from the two cages	0.0	0.0	0.0
3	Error reading the outcome of the draws	0.0	0.0	0.0
Panel B. Errors in the application of Bayes' Rule				
4	Ignoring the prior	82.4	1.4	0.6
5	Ignoring the likelihood	0.2	0.0	0.0
Panel C. Errors in computing the posterior probability				
6	Error calculating prior probability	83.4	2.2	0.6
7	Error calculating the likelihood	70.0	67.8	13.6
8	Error calculating the posterior (or posterior odds)	98.4	72.0	21.8
Panel D. Errors in the final decision				
9	Final decision contradicting the previous reasoning	4.4	3.0	4.6
Fraction of responses inconsistent with Bayes' rule		35.2	14.2	9.2

Notes: The table presents nine error flags grouped into four broad categories, one in each panel. We use GPT-o3-mini to analyze the textual responses for 500 randomly selected responses for each version of GPT. The third to fifth columns report the fraction of sample responses with an error flag of 1, as judged by the GPT-o3-mini grader, for GPT-3.5, GPT-4, and GPT-4o, respectively. The error flags are not mutually exclusive, so they do not necessarily sum to 1. The bottom of the table reports the fraction of responses in the randomly selected sample that are consistent with Bayes' Rule.

Table 6 reinforces our conclusion about the rapid improvement from GPT-3.5 to GPT-4 and GPT-4o, as GPT-4 and GPT-4o make significantly fewer errors in almost all types. This not only reflects higher accuracy, but also indicates fewer mistakes in calculating posterior probabilities or other steps even when final decisions align with Bayes' rule. Panel A shows that GPT subjects rarely make mistakes in reading experimental parameters, suggesting a good understanding of the experimental setup.

Panel B shows the remarkable transformation from non-Bayesian to conceptual Bayesian reasoning from GPT-3.5 to GPT-4. More than 80% of GPT-3.5 decisions rely only on likelihood, lacking a Bayesian rationality even conceptually. This aligns with our findings in Section 5. Our manual review of the text responses shows that they often use representativeness heuristics, making decisions by matching observed patterns in the sample with the composition of the two cages, or simply comparing the likelihood of each cage being the sample's source. Interestingly, GPT-3.5 almost never ignores information from the sample, suggesting it is less prone to conservative bias. GPT-4 and GPT-4o almost always demonstrate at least "conceptual Bayesian" reasoning, either by explicitly writing down Bayes' formula or informally considering both prior and likelihood information.

Panel C highlights the transition from being "conceptually Bayesian" to being more fully "Bayesian" between GPT-4 and GPT-4o subjects. While both groups utilize prior probabilities and sample information, only 28% of GPT-4 subjects accurately calculate the posterior, compared to 78% of GPT-4o subjects. GPT-3.5 struggles to calculate both the likelihood and prior probability, often due to overlooking prior information.

Panel D shows that all GPT models experience decision noise. Surprisingly, although GPT-4o generally outperforms GPT-4, its final decisions are more likely to be inconsistent with the calculated posterior, whether comparing it to $\frac{1}{2}$ or the posterior odds. To understand it further, we examined a sample of 50 scored by a human grader who also confirmed this finding.[37] First, we find that GPT-4o is more likely to report the posterior as fractional numbers than GPT-4,[38] due to its ability to more accurately calculate the posterior, as shown in Panel C. However, the fractional numbers reported by GPT-4o, though precise, are more complex with more digits[39], making comparisons for the final decision more challenging.[40] Thus while GPT-4o calculates the posterior more accurately, it is more susceptible to final decision errors in the comparison process of these numbers.

7 Conclusion

Who is more Bayesian? We compared humans to the first three generations of ChatGPT — 3.5, 4, and 4o — in an exact match of trials from a series of experiments conducted by El-Gamal and Grether (1995), El-Gamal and Grether (1999) and Holt and Smith (2009). The first version, GPT-3.5, is not Bayesian because it ignores prior information. GPT-4 and GPT-4o use Bayes' Rule but make algebraic errors in computing the Bayesian posterior, so they are noisy Bayesians. But the level of noise is dramatically lower for GPT-4o and its reported beliefs are nearly perfectly Bayesian.

What about humans? The conventional wisdom is that most humans are not Bayesian, because as Gennaioli and Shleifer (2010) noted, the influential work of Tversky and Kah-

[37]The human grader reported an error rate of 14% for GPT-4o and 2% for GPT-4 regarding the decision inconsistency error flag. See Appendix G.

[38]Out of the 50 samples, 86% are reported as fractional numbers in GPT-4o, with the remainder as rounded decimals. This percentage decreases to 66% in GPT-4.

[39]See Appendix F for an example where the posterior for Cage A is calculated and reported as $\frac{2612736}{4200459}$, which equals 0.62 as a decimal. The GPT-4o model should have chosen Cage A, but instead, it chose Cage B.

[40]Out of 17 responses where GPT-4o reported posteriors as fractional numbers, 7 resulted in mistakes during the final decision. In contrast, none of the 7 cases with fractional posteriors from GPT-4 had such issues. We note that only 1 decision was inconsistent with the calculated posterior out of 43 cases for GPT-4 and 33 cases for GPT-4o using rounded decimals.

neman (1974) and other studies in psychology and behavioral economics documented "significant deviations from the Bayesian theory of judgment under uncertainty." Our reanalysis of elicited human posterior beliefs from experiments by Holt and Smith (2009) shows that at least in this specific experiment, human beliefs do correspond rather closely to Bayes' Rule. However we do find substantial random noise and subject-specific heterogeneity in reported beliefs, with a majority of subjects (55%) putting excessive weight on the prior. As a result, we conclude that GPT-4o is more Bayesian than humans.

Why should we care if GPT-4o is more Bayesian than humans? Beyond pure intellectual interest, the practical reason why we care is the supposition that decision makers who are more Bayesian are also quantifiably *better* decision makers, in the sense of obtaining higher expected payoffs/welfare than decision makers who are "less Bayesian". Indeed, a theoretical literature has arisen that provides an objective way of ranking the performance of different decision makers with potentially distorted/noisy subjective beliefs "Given a true signal distribution, we deem one bias more harmful than another if it yields lower objective expected payoffs in all decision problems." Frick et al. (2024), (p. 1612). They show how subjective beliefs can be partially ordered to provide "a welfare-founded approach to quantify and compare the severity of many well-documented learning biases." (p. 1613). Specializing this framework to the binary classification experiment analyzed in this paper, we show that only way a decision maker can make fully optimal decisions in *all possible* decision problems is for their subjective posterior beliefs to coincide with Bayes' Rule.

However it is challenging and may not even be possible to rank the optimal expected payoffs for decision makers with different subjective beliefs. Therefore we focus on empirically ranking decision rules used by different subjects for the specific symmetric payoff function used in these binary classification experiments, namely a payoff of 1 if the subject correctly chooses the cage used to draw the sample and 0 otherwise. Appealing to statistical decision theory, we introduced an objective measure of *decision efficiency,* the ratio of the subject's probability of being correct to the optimal probability of being correct under Bayes' Rule.

Using a model of decision making with potentially non-Bayesian beliefs, we showed how belief biases and decision noise contribute to inefficiency, and how subjects with distorted beliefs can be more efficient than subjects with Bayesian beliefs but high levels decision noise. We also showed theoretically and empirically that it is possible to behave like a Bayesian even if beliefs are non-Bayesian, i.e. the existence of subjects with non-Bayesian beliefs whose behavior can be fully or nearly fully efficient. Efficiency depends on where mistakes are made: mistakes on "easy cases" where the true posterior is close to 0 or 1 reduce efficiency much more than mistakes on the "hard cases" where the posterior is close to 1/2. Despite their biased beliefs and decision noise, human subjects make most of their mistakes on the hard cases so their overall efficiency is surprisingly high, typically over 95% to nearly 99% for the minority of best performing humans. In comparison, GPT-3.5's efficiency ranges between 75-85% and GPT-4o's efficiency ranges from 97 to 99% across the experiments we analyzed.

A shortcoming of the simple "balls drawn from urns" experimental design of Grether (1978) is that it eliminates extraneous information and inherent ambiguity that are features of most real world decision problems, reducing the external validity of the findings. As we noted in the Introduction, this was intentional, as Grether hypothesized that the rejection of rational Bayesian behavior in experiments by psychologists such as Tversky and Kahneman (1974) might be due to the possibility that ambiguity and extraneous information could amplify potential for framing effects and stereotyping to distort/bias subjects' choices.

Given that the human brain evolved to deal with a world full of ambiguity and extraneous information, we would expect that these are the environments where humans should outperform AI, especially in situations where there is payoff-relevant sensory and contextual information that humans use but is difficult to collect and digitize for use by AI algorithms. However the range of information that LLMs can process does include audio/visual data, as well as digital data such as X-rays and mammograms that AI may have a comparative advantage in processing. The studies we cited in the introduction that compare the performance of human and AI decision makers in medical diagnosis and decision making generally find that AI significantly outperforms humans. It remains unclear whether AI to is superior to humans in its ability to separate payoff-relevant from extraneous information, and whether AI is superior in how it frames and uses the information to solve an ambiguous decision problem. One might expect that an even more effective approach would be to combine the intuition and judgment of a human expert with AI's ability to rapidly process large volumes of data, but several studies of medical diagnosis that we cited in the introduction find that AI alone outperforms AI-assisted human decision makers.

The biggest challenge to comparing the performance of human and AI decision making in complex real-world applications such as medical diagnosis and treatment decisions is the lack of an objective criterion of the "correctness" or "optimality" of decisions. This is due to the fundamental ambiguity of most decision problems where one does not know the correct prior, likelihood, or payoff function and there is no omniscient outside observer or oracle that knows the truth and can determine what the optimal or "correct" decision should be. Without an objective metric of performance, there is a potential for "AI hype" to create misleading conclusions about the superiority of AI over humans (or vice versa).

One way forward is to further develop and empirically implement and test theories of decision making and belief updating under ambiguity, such as Gilboa and Schmeidler (1989), Gilboa and Schmeidler (1993). However, progress has been slow in developing an objective framework to compare the optimality of different decision rules that we can agree on. Etner et al. (2010) notes "This literature still produces hot debates" and Cohen et al. (2000) notes that "the question of updating 'ambiguous' beliefs was raised in economic theory and artificial intelligence. Several contradictory answers were put forward, which were axiomatically justified in the theoretical works of the authors. However, both the axioms and the derived rules can only be judged through introspection."

Without adequate guidance from theory and experimental studies to define an objective measure of performance it is understandable that most of the existing comparisons of human and AI decision making in real-world environments rely on simple measures of decision accuracy that define a "correct" decision to be the one made by a majority of experts. But how do we determine who the "experts" are, and do we require them to be human? If we acknowledge that AI can outperform human experts in domains where outcomes are easy to observe (for example in the ability to win in chess), the need for an objective definition of "performance" to compare human and AI decision makers in cases where the outcomes/consequences of decisions are less clear cut is obvious.

We have documented a significant improvement in GPT's performance, from sub-human to super-human in just three years, using an objective measure of performance, decision efficiency. There is widespread concern that AI will soon be able to outperform most humans in a wide variety of much more complicated and challenging intellectual tasks, and it is already threatening the jobs of of high-skilled workers including computer programmers. According to Google's Gemini "In the first seven months of 2025, over 10,000 job cuts were directly attributed to the adoption of generative AI by private employers." While this certainly represents a great leap forward in the field of AI and is further evidence that we are at the verge of artificial general intelligence, whether this should be a cause for celebration or panic is beyond the scope of this paper.

References

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. *National Bureau of Economic Research Working Paper Series*, 31422, 2024.

Cuimin Ba, J. Ainslinn Bohren, and Alex Imas. Over- and underreaction to information: Belief updating with cognitive constraints. *Working paper, University of Pittsburgh*, 2025.

G. Becker, M. H. DeGroot, and J. Marshak. Measuring utility by a single-response method. *Behavioral Science*, 9:226–232, 1964.

Daniel J. Benjamin. Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 69–186. Elsevier, 2019.

Andrew Caplin, David Deming, Shijie Li, Daniel Martin, Paul Marx, Ben Weidmann, and K. J. Ye. The abcs of who benefits from working with ai: Ability, beliefs, and calibration. *Management Science*, 2025.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt. Technical report, Tsinghua University, 2023.

M. Cohen, I. Gilboa, J. Y. Jaffray, and D. Schmeidler. An experimental study of updating ambiguous beliefs. *Risk, Decision and Policy*, 5:123–133, 2000.

Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58:1–73, 1996.

Mahmoud A. El-Gamal and David M. Grether. Are people bayesian? uncovering behavorial strategies. *Journal of the American Statistical Association*, 90-432:1137–1145, 1995.

Mahmoud A. El-Gamal and David M. Grether. Changing decision rules: Uncovering behavioral strategies using estimation/classification (ec). In *Beliefs, interactions, and preferences in decision making*, pages 3081–3143. Dordrecht; Boston: Kluwer Academic, 1999. URL https://www.worldcat.org/title/beliefs-interactions-and-preferences-in-decision-making/oclc/1012461235.

Benjamin Enke and Thomas Graeber. Cognitive uncertainty. *Quarterly Journal of Economics*, 138:2021–2066, 2023.

Johanna Etner, Meglena Jeleva, and Jean-Marc Tallon. Decision theory under ambiguity. *Journal of Economic Surveys*, 26:234–270, 2010.

Sara Fish, Julia Shephard, Minkai Li, Ran I. Shorrer, and Yannai A. Gonczarowski. EconEvals: Benchmarks and litmus tests for llm agents in unknown environments. *arXiv preprint arXiv:2502.xxxxx*, 2025.

Mira Frick, Ryota Iijima, and Yuhta Ishii. Welfare comparisons for biased learning. *American Economic Review*, 114:1612–1649, 2024.

Nicola Gennaioli and Andrei Shleifer. What comes to mind. *Quarterly Journal of Economics*, 125-4:1399–1432, 2010.

Itzak Gilboa and David Schmeidler. Maximin expected utility with a non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

Itzak Gilboa and David Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59:33–49, 1993.

Ethan Goh et al. Large language model influence on diagnostic reasoning a randomized clinical trial. *JAMA Open*, 7-10:1–12, 2024.

David M. Grether. Recent psychological studies of behavior under uncertainty. *American Economic Review*, 68-2:70–74, 1978.

David M. Grether. Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics*, 95:537–557, 1980.

J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52-2:271–320, 1984.

Charles A Holt. *Markets, Games, and Strategic Behavior: An Introduction to Experimental Economics*. Princeton University Press, Princeton, New Jersey, 2019.

Charles A. Holt and Angela M. Smith. An update on bayesian updating. *Journal of Economic Behavior and Organization*, 69:125–134, 2009.

J. Wesley Hutchinson and Robert J. Meyer. Dynamic decision making: Optimal policies and actual behavior in sequential choice problems. *Marketing Letters*, 5-4:369–382, 1994.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27-4:887–906, 1956.

Matthew Kovach, Daniel Martin, and Gerelt Tserenjigmid. Learning from an unknown dgp: Experimental evidence on belief updating with ai recommendations. Slides available at https://www.dropbox.com/scl/fi/707jcawh0evvscnvlazor/General_Information_Experiment_LMU.pdf?rlkey=zta3mmqkxandbw9vtbu74kudu&e=1&dl=0, 2025.

Niklas Kühl, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, 76:78–92, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial intelligence index report 2024. *Stanford Institute for Human-Centered Artificial Intelligence*, 2024. URL https://aiindex.stanford.edu/report/2024.

Daniel McDuff et al. Towards accurate differential diagnosis with large language models. Technical report, Google Research and Google DeepMind, 2023.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, pages 105–141. Academic Press, 1974.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Nikhil Raman, Thomas Lundy, Simon Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. Steer: Assessing the economic rationality of large language models. *arXiv preprint arXiv:2402.09552*, 2024.

Nikhil Raman, Thomas Lundy, Taha Amin, Jesse Perla, and Kevin Leyton-Brown. STEER-ME: Assessing the microeconomic reasoning of large language models. *arXiv preprint arXiv:2502.13119*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Amos Tversky and Daniel Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185-4157:1124–1131, 1974.

Quang Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 52-2:307–333, 1989.

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

Tiemen Woutersen and John Rust. Why so few types? Technical report, Georgetown University, 2025.

Xintian Yang et al. Multiple large language models versus experienced physicians in diagnosing challenging cases with gastrointestinal symptoms. *NPJ Digital Medicine*, 8: 1–12, 2025.

Heera Yoen and Jung Min Chang. Artificial intelligence improves detection of supplemental screening ultrasound-detected breast cancers in mammography. *Journal of Breast Cancer*, 26-5:504–513, 2023.

Appendix A Proof of Lemma L2

Proof. The optimal decision rule of an objective Bayesian decision maker (i.e. one who has correct unbiased beliefs given by the Bayesian posterior Π in (1) section 2), can be written as a threshold rule,

$$\delta^*(u,\Pi)(d,\pi,p_A,p_B,D) = \begin{cases} A & \text{if } E\{u|A\}(d,\pi,p_A,p_B,D) \geq E\{u|B\}(d,\pi,p_A,p_B,D) \\ B & \text{otherwise} \end{cases} \quad (22)$$

where $E\{u|a\}(d,\pi,p_A,p_B,D)$ is the expected payoff from taking action $a \in \{A,B\}$ in state (d,π,p_A,p_B,D) given by

$$E\{u|a\}(d,\pi,p_A,p_B,D) = u(A,a)\Pi(A|d,\pi,p_A,p_B,D) + u(B,a)[1-\Pi(A|d,\pi,p_A,p_B,D)]. \quad (23)$$

It is not hard to see that the threshold rule (22) defining the states for which the choice of cage A is optimal is equivalent to the following cutoff rule for the posterior probability Π

$$\Pi(A|d,\pi,p_A,p_B,D) \geq \frac{u(B,B)-u(B,A)}{u(A,A)-u(A,B)+u(B,B)-u(B,A)} \equiv \overline{p}. \quad (24)$$

Similarly, a decision rule $\delta^*(u,\Pi_s)$ that is optimal with respect to *subjective beliefs* Π_s can also be expressed as a cutoff rule of the form (24), but with Π_s in place of Π. Since $\delta^*(u,\Pi)$ defined in terms of the true (objective) Bayesian posterior Π is an optimal decision rule, whereas a decision rule $\delta^*(u,\Pi_s)$ defined in terms of a potentially biased subjective posterior belief Π_s cannot have a higher expected payoff when we evaluate this expectation with respect to the true probability measure governing outcomes of the experiment (which is given by Π). It follows that the expected payoff from the optimal decision rule $\delta^*(u,\Pi)$ must exceed the expected payoff from the decision rule $\delta^*(u,\Pi_s)$, since the latter is only optimal with respect to the subjective posterior beliefs Π_s but not necessarily with respect to the objective posterior beliefs Π that generate the actual outcomes in the experiment.

Now suppose the expected payoffs (with respect to the true Bayesian posterior Π) for the two decision rules $\delta^*(u,\Pi)$ and $\delta^*(u,\Pi_s)$ are equal for all payoffs u and all information states (d,π,p_A,p_B,D). We now show that this implies that $\Pi_s(A|d,\pi,p_A,p_B,D) = \Pi(A|d,\pi,p_A,p_B,D)$ for all states (d,π,p_A,p_B,D). Suppose, by way of contradiction, that this equality does not hold for some state (d,π,p_A,p_B,D). For example, assume that $\Pi_s(A|d,\pi,p_A,p_B,D) > \Pi(A|d,\pi,p_A,p_B,D)$. Choose any utility function u such that the inequality (24) actually holds with equality at (d,π,p_A,p_B,D) with respect to the true Bayesian beliefs Π. Call the threshold value of this probability $\overline{p}$ as shown in (24). Thus at $\Pi(A|d,\pi,p_A,p_B,D) = \overline{p}$ for this particular payoff function u, the Bayesian decision maker is indifferent between choosing A and B at state (d,π,p_A,p_B,D). By assumption, the subjective posterior $\Pi_s(A|d,\pi,p_A,p_B,D)$ is a continuous function of π, so it follows that if $\Pi_s(A|d,\pi,p_A,p_B,D) > \Pi(A|d,\pi,p_A,p_B,D) = \overline{p}$ there is a $\pi' < \pi$ such $\Pi_s(A|d,\pi',p_A,p_B,D) > \overline{p}$. Since the same cutoff rule (24) determines the optimal decision rule $\delta^*(u,\Pi_s)$ of a subjective Bayesian decision maker with subjective posterior beliefs Π_s, it follows that $\delta^*(u,\Pi_s)(d,\pi',p_A,p_B,D) = A$ and thus $E\{u|\delta^*(u,\Pi_s)\}(d,\pi',p_A,p_B,D) = E\{u|A\}(d,\pi',p_A,p_B,D)$. However, it is easy to see that the Bayesian posterior Π is a continuous, monotonically increasing function of the prior π, so it follows that for any $\pi' < \pi$ the optimal decision of a Bayesian decision maker is B, i.e. $\delta^*(u,\Pi)(d,\pi',p_A,p_B,D) =$

B. It follows that $E\{u|\delta^*(u,\Pi)\}(d,\pi',p_A,p_B,D) > E\{u|A\}(d,\pi',p_A,p_B,D)$. But if $E\{u|\delta^*(u,\Pi)(p,\pi,p_A,p_B,D) = E\{u|\delta^*(u,\Pi_s)\}(d,\pi,p_A,p_B,D)$ for all possible u and all states (d,π,p_A,p_B,D), then we have a contradiction, since we have

$$\begin{aligned} E\{u|A\}(d,\pi',p_A,p_B,D) &= E\{u|\delta^*(u,\Pi_s)\}(d,\pi',p_A,p_B,D) \\ &= E\{u|\delta^*(u,\Pi)\}(d,\pi',p_A,p_B,D) \\ &> E\{u|A\}(d,\pi',p_A,p_B,D). \end{aligned} \tag{25}$$

□

Appendix B Proof of Lemma L3

This appendix provides the proof of identification of the structural logit model when $\sigma > 0$, Lemma L3 of section 3.1. We assume that the subject's decision rule $P(A|\pi, d, p_A, p_B, D)$ is identified. Since the information (π, d, p_A, p_B, D) enters P via LLR and LPR, they can take any value between $-\infty$ and $+\infty$ via an appropriately designed experiment. So hereafter we write P as $P(A|\text{LLR}, \text{LPR})$ which is also identified, i.e., known for all LPR and LLR in R^2. Now, suppose the subject's true decision rule is defined by the parameters (σ^*, β^*). Recalling equation (15) we have

$$P(A|\text{LLR}, \text{LPR}, \sigma^*, \beta^*) = \frac{1}{1 + \exp\{[2\Pi_s(A|\text{LLR}, \text{LPR}, \beta^*) - 1]/\sigma^*\}}. \tag{26}$$

Now suppose there is some other parameter (σ, β) that is observationally equivalent to (σ^*, β^*), so we have

$$P(A|\text{LLR}, \text{LPR}, \sigma^*, \beta^*) = P(A|\text{LLR}, \text{LPR}, \sigma, \beta), \quad \forall(\text{LLR}, \text{LPR}) \in R^2. \tag{27}$$

We will show that equation (27) holds if and only if $\sigma = \sigma^*$, and $\beta = \beta^*$. That is, the parameters of the structural logit model are identified.

Consider the formula for Π_s in equation (12). Suppose the coefficient β_2^* of LPR is positive. Then if we let $\text{LPR} \to -\infty$ we have $\Pi_s(A|\text{LPR}, \text{LLR}) \to 0$. Since equation (27) holds for all $(\text{LPR}, \text{LLR}) \in R^2$ and is a continuous function of these variables, it follows that the equality must hold in the limit so we have

$$\frac{1}{1 + \exp\{1/\sigma^*\}} = \frac{1}{1 + \exp\{1/\sigma\}}, \tag{28}$$

which implies that $\sigma^* = \sigma$, so this parameter is identified.[41] Using this result equations (26) and (27) imply the following equality

$$\Pi_s(A|\text{LPR}, \text{LLR}, \beta^*) = \Pi_s(A|\text{LPR}, \text{LLR}, \beta). \tag{29}$$

Since $\Pi_s(A|\text{LPR}, \text{LLR}, \beta) = 1/(1 + \exp\{\beta_0 + \beta_1\text{LPR} + \beta_2\text{LLR}\}$, it follows that we have

$$\beta_0^* + \beta_1^*\text{LPR} + \beta_2^*\text{LLR} = \beta_0 + \beta_1\text{LPR} + \beta_2\text{LLR}, \quad \forall(\text{LPR}, \text{LLR}) \in R^2. \tag{30}$$

Let $\text{LPR} = \text{LLR} = 0$. It follows from equation (30) that $\beta_0^* = \beta_0$. Next, set $\text{LLR} = 0$ and $\text{LPR} = 1$, and it follows that $\beta_1^* = \beta_1$. Finally, set $\text{LLR} = 1$ and $\text{LPR} = 0$, and it follows that $\beta_2^* = \beta_2$. We conclude that there is only a single solution to equation (27), i.e., the parameters $(\sigma^*, \beta_0^*, \beta_1^*, \beta_2^*)$ of the structural logit model are identified.

[41] If $\beta_2^* < 0$, then we repeat the same argument except we let $\text{LPR} \to \infty$.

Appendix C Algorithm for Data Collection from LLMs

Algorithm 1 Data Collection and Processing

```
Initialize data collection specifications based on the experiment specified by run_name,
  generate textual prompts and model settings, and save to disk
'send' prompts and model settings to OpenAI via API
while any request is not completed do
    for every request do
        'retrieve' the request's status
        if the request has been completed and not retrieved then
            Retrieve the responses and save to disk
        else if the request has failed then
            'resend_failed' request(s)
        else if
             then continue
        end if
    end for
    'finalize' the responses, which includes:
        Align responses with the original prompts
        Parse the responses into the final answers like Cage A or B or a numerical value
        Append metadata and informational columns
        Save different formats of the collected data to disk
        Check for invalid responses that cannot be parsed into a final answer
    if any response is invalid then
        'resend_invalid' prompts and model settings to OpenAI via API
    end if
end while
```

Appendix D Prompts used to collect data from LLMs

In this section, we provide our prompt to replicate using LLMs the two experiments at the University of Wisconsin-Madison and the two experiments reported in Holt and Smith (2009).

There are no `developer` / `system` messages, only one `user` message for each chat completion request for each trial. The exact numerical values will reflect the specifications of the actual trial, the followings are examples.

C.1 Wisconsin

For the experiment that allows for `reasoning`, the example `user` message is:

```
You are participating in a decision-making experiment, where you
   can earn money based on the number of correct decisions you
   make.

There are two identical bingo cages, Cage A and Cage B, each
   containing 6 balls. Cage A contains 4 balls labeled "N" and 2
   balls labeled "G", while Cage B contains 3 balls labeled "N"
   and 3 balls labeled "G".

A 10-sided die is used to determine which of the two cages will
   be used to generate draws. If a random roll of the die shows 1
    through 3, I will use Cage A; if it shows 4 through 10, I
   will use Cage B. You will not know the outcome of the roll of
   the die or which cage I use.

Once a cage is chosen at random based on the roll of the die, it
   is used to generate draws with replacement.

I have drawn a total of 6 balls with replacement. The result is 3
    "N" balls and 3 "G" balls.
After observing this outcome, which cage do you think generated
   the observations? Your decision is correct if the balls were
   drawn from that cage.

YOU ARE WELCOME TO ALSO DESCRIBE YOUR REASONING, BROKEN INTO
   SEPARATE STEPS, TO EXPLAIN HOW YOU ARRIVED AT YOUR FINAL
   ANSWER.
Please state your answer in the following format at the end.
"Final answer: Cage A." or "Final answer: Cage B.".
```

For the experiment that prohibits `reasoning`, the last section of the `user` message is substituted with:

```
PLEASE JUST REPORT YOU FINAL ANSWER AND DO NOT PROVIDE ANY
   REASONING AS TO HOW YOU ARRIVED AT YOUR FINAL ANSWER.
Please state your answer in the following format.
"Final answer: Cage A." or "Final answer: Cage B.".
```

C.2 Holt and Smith

For the experiment that allows for `reasoning`, the example `user` message is:

```
This is an experiment in the economics of decision making.
   Various agencies have provided funds for the experiment. Your
   earnings will depend partly on your decisions and partly on
   chance. If you are careful and make good decisions, you may
   earn a considerable amount of money, which will be paid to you
   , privately, in cash, at the end of the experiment. In
   addition to the money that you earn during the experiment, you
    will also receive $6. This payment is to compensate you for
   showing up today.

This experiment involves two stages. In stage 1 we will show you
   some information including the result of a drawing of 1 ball
   from one of two possible cages, each containing different
   numbers of light and dark balls. Then at the start of stage 2
   you will report a number P between 0 and 1. After your report,
    we will draw a random number U that is equally likely to be
   any number between 0 and 1. Your payoff from this experiment
   will either be $1000 or $0 depending on your report P and the
   random number U.

Let's describe the two stages in more detail now. In stage 1 we
   will show you 1 ball that are drawn at random from one of two
   possible urns labelled A and B.

Urn A contains 2 light balls and 1 dark ball.
Urn B contains 1 light ball and 2 dark balls.

We select the urn, A or B, from which we draw the sample of 1
   ball by the outcome of throwing a 6 sided die.
We do not show you the outcome of this throw of the die but we do
    tell you the rule we use to select urn A or B.

If the outcome of the die throw is 1 to 3 we select urn A.
If the outcome of the die throw is 4 to 6, we use urn B to draw
   the random sample of 1 ball.

Once you see the outcome of the sample of 1 ball, stage 1 is over
    and stage 2 begins.

At the start of stage 2 we ask you to report a number P between 0
    and 1. Your payoff from this experiment depend on another
   random number, which we call U, which we draw after you report
    the number P. We draw the random number U in a way that every
    possible number between 0 and 1 has an equal chance of being
   selected.
```

```
Here is how you will be paid from participating in this
   experiment. There are two possible cases:

Case 1. If the number U is less than or equal to P then you will
   receive $1000 if the sample of 1 ball we showed you in stage 1
    was from urn A and $0 otherwise.
Case 2. If the number U is between the number P you report and 1,
    you will receive $1000 with probability equal to the realized
    value of U, but with probability 1-U you will get $0.

OK, this is the setup. Let's now start begin this experiment,
   starting with stage 1.

We have tossed the die (the outcome we don't show to you) and
   selected one of these urns according to the rule given above (
   i.e., urn A if the die throw was 1 to 3, and urn B otherwise).
    We have drawn 1 ball from the selected urn and the outcome is
    D, i.e., Dark.

Now, we are at stage 2 where we are asking you, given the
   information from stage 1 to report a number P between 0 and 1
   that in conjunction with the random number U will determine if
    you get either $1000 or $0 according to the rule given in
   cases 1 and 2 above.

Please report a number P between 0 and 1 that maximizes your
   probability of winning $1000 in this experiment.

YOU ARE WELCOME TO ALSO DESCRIBE YOUR REASONING, BROKEN INTO
   SEPARATE STEPS, TO EXPLAIN HOW YOU ARRIVED AT YOUR FINAL
   ANSWER P.
Please state your answer in the following format at the end.
Final answer: [your P value here].
```

For the experiment that prohibits `reasoning`, the last section of the `user` message is substituted with:

```
PLEASE JUST REPORT P AND DO NOT PROVIDE ANY REASONING AS TO HOW
   YOU ARRIVED AT THE VALUE P.
Please state your answer in the following format.
Final answer: [your P value here].
```

Appendix E Grading Prompts

We provide a grading prompt that evaluates GPT-generated textual responses and identifies errors. First, we introduce the grading task and the role of the grader to GPT-01. Next, we present the experiment design prompt along with the original responses from GPT students. The key input for our grading prompt is a meticulously crafted step-by-step solution to the binary choice problem, which includes calculating the prior probability, determining the likelihood, applying Bayes' rule to find the posterior probability, and making a final decision based on the posterior. This solution is crucial as a benchmark for the GPT grader to understand the binary decision task and identify errors in the student responses. We then provide detailed grading rubrics that allow the GPT grader to classify different error types. Finally, the prompt specifies the format for outputting error flags. We develop Python code to loop over different student responses, parameters of trials, and the corresponding benchmark solutions.

General Instruction about the Role of the Grader

```
You are an expert and detailed grader for this task:
```

Introduce the Experiment Prompt

```
You are participating in a decision-making experiment, where you
   can earn money based on the number of correct decisions you
   make.

There are two identical bingo cages, Cage A and Cage B, each
   containing 10 balls. Cage A contains 4 balls labeled "N" and 6
    balls labeled "G", while Cage B contains 6 balls labeled "N"
   and 4 balls labeled "G".

A 10-sided die is used to determine which of the two cages will
   be used to generate draws. If a random roll of the die shows 1
    through 6, I will use Cage A; if it shows 7 through 10, I
   will use Cage B. You will not know the outcome of the roll of
   the die or which cage I use.

Once a cage is chosen at random based on the roll of the die, it
   is used to generate draws with replacement.

I have drawn a total of 7 balls with replacement. The result is 3
    "N" balls and 4 "G" balls.
After observing this outcome, which cage do you think generated
   the observations? Your decision is correct if the balls were
   drawn from that cage.

YOU ARE WELCOME TO ALSO DESCRIBE YOUR REASONING, BROKEN INTO
   SEPARATE STEPS, TO EXPLAIN HOW YOU ARRIVED AT YOUR FINAL
   ANSWER.
Please state your answer in the following format at the end.
"Final answer: Cage A." or "Final answer: Cage B.".
```

Student textual responses

```
Based on the outcome of 3 "N" balls and 4 "G" balls, here is my
   reasoning broken down into steps:

Step 1: Calculate the probability of getting 3 "N" balls and 4 "G
   " balls from each cage.

- Probability of getting 3 "N" balls and 4 "G" balls from Cage A:
    (4/10)^3 * (6/10)^4 = 0.00324
- Probability of getting 3 "N" balls and 4 "G" balls from Cage B:
    (6/10)^3 * (4/10)^4 = 0.00307

Step 2: Compare the probabilities of the two cages.

- The probability of getting the observed outcome is slightly
   higher for Cage A than Cage B.

Step 3: Consider the initial probabilities of each cage being
   chosen.

- Cage A has a 6/10 chance of being chosen, while Cage B has a
   4/10 chance.

Step 4: Combine the probabilities of the observed outcome and the
    initial probabilities of the cages being chosen.

- Probability of getting the observed outcome from Cage A:
   0.00324 * 0.6 = 0.00194
- Probability of getting the observed outcome from Cage B:
   0.00307 * 0.4 = 0.00123

Step 5: Compare the combined probabilities.

- The combined probability of getting the observed outcome is
   higher for Cage A than Cage B.

Final answer: Cage A.
```

Step-by-step Solution Benchmark

```
Step 1: Prior from die roll.
  p(A) = 3/5 = 0.6000;  p(B) = 2/5 = 0.4000
  Prior ratio p(A)/p(B) = 3/2 = 1.500000
Step 2: Cage compositions.
  Cage A: p(NA) = 2/5 = 0.4000, p(GA) = 3/5 = 0.6000
  Cage B: p(NB) = 3/5 = 0.6000, p(GB) = 2/5 = 0.4000
Step 3: Observed outcome & binomial likelihood.
  Observed: 3 'N', 4 'G' (total 7).
  L(A) = 4536/15625 = 0.290304, L(B) = 3024/15625 ≈ 0.193536
  Likelihood ratio L(A)/L(B) = 3/2 = 1.500000
```

```
Step 4: Posterior components & probabilities (fraction & decimal)
   .
  p(A)*L(A) = 13608/78125 ≈ 0.174182
  p(B)*L(B) = 6048/78125 ≈ 0.077414
  Post(A)   = 9/13 = 0.692308
  Post(B)   = 4/13 = 0.307692
  Posterior ratio Post(A)/Post(B) = 9/4 = 2.250000
Step 5: Decision.
  Final answer: Cage A.
```

Grading Rubrics

```
Evaluate the student's answer based on the following criteria:

Part I. Did They Make a Mistake When Reading the Data?

Instructions: For (1), (2), and (3), we want to see if the
   student understands the basic experimental setup and correctly
    incorporates the trial parameters into their reasoning. Focus
    on whether they read the relevant information accurately, not
    on how they use it later. For example, if a student correctly
    identifies the number of N balls in cages A and B but makes a
    comparison error later, you should still answer YES if the
   criterion is reading the number of N balls correctly.

(1) Cage Composition: Do they explicitly mention or implicitly
   acknowledge the number of N and G balls in each cage? Answer
   Yes or No.

(2) Draw Count: Do they explicitly mention or implicitly
   acknowledge the total number of balls in the sample? Answer
   Yes or No.

(3) Observed Data: Do they explicitly mention or implicitly
   acknowledge the number of N draws from the sample? Answer Yes
   or No.

Part II. Are they conceptually Bayesian?

Instructions: A Bayesian decision maker should consider both
   prior information (the announced probability of using a cage)
   and posterior information (the likelihood that the sample was
   drawn from a cage). Criteria (4) and (5) assess whether the
   student incorporates both prior and posterior information in
   their reasoning. We are not looking for explicit numerical
   calculations, but both types of information should be part of
   their reasoning process.

(4) Ignoring Prior: Do they make a decision using only the
   likelihood or observed data, ignoring the prior conceptually?
```

```
   Answer Yes or No.

(5) Ignoring Likelihood: Do they make a decision using only the
   announced probability of using cage A, ignoring the sample
   information conceptually? Answer Yes or No.

Part III. Can they correctly calculate the Bayesian posterior
   probability?

Instructions: To answer correctly, the student should apply Bayes
   ' rule and calculate the posterior probability accurately.
   This involves three steps:

Prior Probability: Calculate the prior probability that the
   sample is drawn from cage A and B, based on the announced
   probability in the experiment.
Likelihood: Calculate the likelihood that a sample is drawn from
   each cage, using the number of N draws in the sample and the
   cage composition.
Posterior Probability: Either calculate the posterior probability
    or compute the product of likelihood and prior for each cage.
Note: Values are equal if they round to the same number at two
   decimal places. For example, 0.333 and 0.33 should be treated
   as the same. If the student doesn't attempt the calculation or
    leaves it incomplete, answer No. If interrupted, also answer
   No.

(6) Prior Computed: Do they calculate the prior probability for
   each cage correctly? Answer Yes or No.

(7) Likelihood Computed: Do they calculate the likelihood
   correctly? Answer Yes or No. Note, that omitting the binomial
   coefficient is acceptable, as it is a constant for both cages.

(8) Posterior Computed: Do they apply Bayes' rule and calculate
   the posterior probability correctly? Alternatively, answer Yes
    if they correctly compute and compare the product of
   likelihood and prior probability. Answer Yes or No.

Part IV. Do they make a final decision that is consistent with
   their reasoning process? Instructions: The student should
   reach a conclusion based on their reasoning process, and the
   final answer should align with that conclusion. You should
   examine the student's reasoning and predict what they should
   report (e.g., cage A or cage B), then compare it to their
   actual report. Provide YES if they are consistent.

Here are two examples of inconsistency.
```

```
A student calculates the posterior probability of cage A as 2/3.
   Since 2/3 is greater than 1/2, they should report cage A but
   instead report cage B.
A student finds the product of likelihood and prior to be 2/3 for
    cage A and 1/3 for cage B. Since 2/3 is greater than 1/3,
   they should choose cage A but report cage B.
(9) Inconsistency: Based on the student's reasoning process and
   decision rule, determine what the final decision should be. Is
    the decision they actually make consistent with their
   reasoning? Answer Yes or No.
Note, that we are looking for cases where the student has reduced
    the answer to decimals or simple fractions, and decides
   against what they should have decided.
We do want to penalize fraction-to-decimal conversion errors if
   they clearly lead the student to the wrong final choice.
If the student never attempts a numeric calculation or fails to
   provide a final decision at all, also answer No.
```

Output Instructions

```
Provide detailed reasoning first and then at the end output a
   JSON. Your answers must use: Yes = 1, No = 0.

    {{
      "1_correct_cage_composition": 1/0,
      "2_correct_draw_count_and_with_replacement": 1/0,
      "3_correct_use_of_observed_data": 1/0,
      "4_conceptual_ignore_prior_likelihood_only": 1/0,
      "5_conceptual_ignore_likelihood_prior_only": 1/0,
      "6_compute_prior_exact": 1/0,
      "7_compute_likelihood_exact": 1/0,
      "8_compute_posterior_exact": 1/0,
      "9_decision_consistency": 1/0
    }}
```

Appendix F Examples of Textual Responses under Error Flags

In this section, we excerpt from representative examples of *students*' original textual responses, classified by the GPT grader under each category, along with the *grader*'s responses explaining the reasoning behind its classifications. We don't present examples for Error Flags 2, 3 and 5 because they are rarely, if any, detected.

1. Cage Composition

Student

Based on the given information, here is my reasoning broken into separate steps:
Step 1. Since there are more N balls in Cage A than in Cage B, and the outcome after 7 draws is 4 N balls and 3 G balls, it's more likely that the draws came from Cage A because there are more N balls available in that cage.
Step 2. The probability of getting a N ball from Cage A is $4/10 = 0.4$, and the probability of getting a G ball from Cage A is $6/10 = 0.6$.
Step 3. The probability of getting a N ball from Cage B is $6/10 = 0.6$, and the probability of getting a G ball from Cage B is $4/10 = 0.4$.
Step 4. Considering the outcomes of 4 N balls and 3 G balls after 7 draws, it is more probable that the draws came from Cage A because the probability of drawing a N ball from Cage A (0.4) is higher than drawing a N ball from Cage B (0.6).
Step 5. Therefore, my final answer is: Final answer: Cage A.

Grader

The student states there are more N balls in Cage A than in Cage B. This is incorrect because Cage A actually has 4 N balls and Cage B has 6 N balls. Thus, they misread the cage composition.

4. Ignoring Prior

Student

Step 1: Calculate the probability of getting 3 "N" balls and 3 "G" balls from each cage.
- For Cage A: The probability of drawing an "N" ball from Cage A is $4/6 = 2/3$. The probability of drawing a "G" ball from Cage A is $2/6 = 1/3$. The probability of getting 3 "N" balls and 3 "G" balls from Cage A is $(2/3)^3 * (1/3)^3 = 8/729$.
- For Cage B: The probability of drawing an "N" ball from Cage B is $3/6 = 1/2$. The probability of drawing a "G" ball from Cage B is also $1/2$. The probability of getting 3 "N" balls and 3 "G" balls from Cage B is $(1/2)^3 * (1/2)^3 = 1/64$.
Step 2: Compare the probabilities of the outcomes from each cage.
- The probability of getting 3 "N" balls and 3 "G" balls is higher for Cage A (8/729) compared to Cage B (1/64).
Step 3: Make a decision based on the higher probability outcome.
Final answer: Cage A.

Grader

The student never mentioned or used the die roll prior probabilities (3/10 for Cage A and 7/10 for Cage B) and made the decision solely based on the likelihood calculation. Thus they did ignore prior information.

6. Prior Computed

Student

Step 1: Calculate the probability of drawing 6 ”N” balls from each cage: For Cage A: $(4/6)^6 = 0.012$, For Cage B: $(3/6)^6 = 0.015$
Step 2: Compare the probabilities: - The probability of drawing 6 ”N” balls from Cage B is slightly higher than from Cage A.
Step 3: Consider the initial composition of the cages: - Cage A has more ”N” balls compared to Cage B, which could make it more likely to draw 6 ”N” balls from that cage.
Final answer: Cage A.

Grader

The student does not calculate or even consider the prior probabilities (the announced cage probabilities).

7. Likelihood Computed

Student

Step 1: Calculate the probability of getting 2 ”N” balls and 5 ”G” balls from each cage.
For Cage A: P(2 ”N” balls and 5 ”G” balls from Cage A) = P(2 ”N” balls from Cage A) * P(5 ”G” balls from Cage A) = $(4/10)^2 * (6/10)^5 \approx 0.0288$
For Cage B: P(2 ”N” balls and 5 ”G” balls from Cage B) = P(2 ”N” balls from Cage B) * P(5 ”G” balls from Cage B) = $(6/10)^2 * (4/10)^5 \approx 0.013824$
Step 2: Compare the probabilities calculated above.
Since the probability of getting 2 ”N” balls and 5 ”G” balls from Cage A is higher than from Cage B, it is more likely that the observations were generated from Cage A.
Final answer: Cage A.

Grader

Step 3: Observed outcome and binomial likelihood. Observed: 2 ’N’, 5 ’G’ (total 7). $L(A) = 20412/78125 \approx 0.261274$ $L(B) = 6048/78125 \approx 0.077414$ Likelihood ratio $L(A)/L(B) = 27/8 = 3.375000$

The student attempts to compute the likelihoods, but they did not include the proper terms (or at least did not compute using the correct values as the proper binomial coefficient is omitted, and the numerical values do not match the correct likelihood calculation). Moreover, their computed numbers do not lead to the correct likelihood ratio.

8. Posterior Computed

Grader

To determine which cage is more likely to have generated the observed outcome, we can use Bayes' theorem. The steps for reasoning through this problem are as follows:
Step 1: Determine Prior Probabilities
- **Probability of using Cage A**: Since the die shows 1 through 4 for Cage A, $P(A) = \frac{4}{10} = 0.4$.
- **Probability of using Cage B**: Since the die shows 5 through 10 for Cage B, $P(B) = \frac{6}{10} = 0.6$.

Step 2: Determine Likelihoods - **Likelihood of drawing 5 "N" and 1 "G" from Cage A**: - Probability of drawing "N" from Cage A: $\frac{4}{6} = \frac{2}{3}$.
- Probability of drawing "G" from Cage A: $\frac{2}{6} = \frac{1}{3}$.
- Probability of the observed outcome from Cage A:

$$\left(\frac{2}{3}\right)^5 \times \left(\frac{1}{3}\right)^1 = \frac{32}{243}.$$

- **Likelihood of drawing 5 "N" and 1 "G" from Cage B**: - Probability of drawing "N" from Cage B: $\frac{3}{6} = \frac{1}{2}$.
- Probability of drawing "G" from Cage B: $\frac{3}{6} = \frac{1}{2}$.
- Probability of the observed outcome from Cage B:

$$\left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^1 = \frac{1}{64}.$$

Step 3: Apply Bayes' Theorem - We want to compare the posterior probabilities $P(A \mid \text{observed outcome})$ and $P(B \mid \text{observed outcome})$.

Using Bayes' theorem:

$$P(A \mid \text{observed outcome}) \propto P(\text{observed outcome} \mid A) \times P(A) = \frac{32}{243} \times 0.4$$

$$P(B \mid \text{observed outcome}) \propto P(\text{observed outcome} \mid B) \times P(B) = \frac{1}{64} \times 0.6$$

Calculate these products: - For Cage A: $\frac{32}{243} \times 0.4 = \frac{12.8}{243}$. - For Cage B: $\frac{1}{64} \times 0.6 = \frac{0.6}{64}$.

Step 4: Compare Posterior Probabilities To compare the likelihoods, it is easier to compare the ratios: - The ratio of the likelihood for Cage A to Cage B is:

$$\frac{\frac{12.8}{243}}{\frac{0.6}{64}} = \frac{12.8 \times 64}{243 \times 0.6} = \frac{819.2}{145.8} \approx 5.62$$

Since the ratio is greater than 1, Cage A is more likely to have produced the observed outcome.

Conclusion Given that the posterior probability is higher for Cage A, it is more likely that Cage A generated the observations.
Final answer: Cage A.

Grader

The student applies Bayes' Rule by multiplying the (miscomputed) likelihood by the correct prior. Although they combine the two pieces of information, the incorrect likelihood for Cage A (and thus the resulting ratio of products) leads to a numerical error in the posterior computation. Therefore, the computation is not exact.

9. Decision Inconsistency

Student

To determine which cage is more likely to have generated the result of 3 "N" balls and 3 "G" balls, let's analyze the problem step by step.
Step 1: Determine the Probability of Choosing Each Cage - The probability of choosing Cage A is determined by the roll of the die showing 1 through 7. So, $P(A) = \frac{7}{10} = 0.7$. - The probability of choosing Cage B is determined by the roll of the die showing 8 through 10. So, $P(B) = \frac{3}{10} = 0.3$.
Step 2: Calculate the Probability of the Observed Result for Each Cage
Cage A: - Probability of drawing "N" from Cage A $= \frac{4}{6} = \frac{2}{3}$. - Probability of drawing "G" from Cage A $= \frac{2}{6} = \frac{1}{3}$. - To get exactly 3 "N" and 3 "G" in 6 draws, we use the binomial probability formula:

$$P(3N, 3G|A) = \binom{6}{3} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^3$$

$$= 20 \times \left(\frac{8}{27}\right) \times \left(\frac{1}{27}\right)$$

$$= 20 \times \frac{8}{729}$$

$$= \frac{160}{729}$$

Cage B: - Probability of drawing "N" from Cage B $= \frac{3}{6} = \frac{1}{2}$. - Probability of drawing "G" from Cage B = $\frac{3}{6} = \frac{1}{2}$. - To get exactly 3 "N" and 3 "G" in 6 draws:

$$P(3N, 3G|B) = \binom{6}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^3$$

$$= 20 \times \left(\frac{1}{8}\right) \times \left(\frac{1}{8}\right)$$

$$= 20 \times \frac{1}{64}$$

$$= \frac{20}{64} = \frac{5}{16}$$

Step 3: Use Bayes' Theorem to Update Beliefs

We need to calculate $P(A|\text{3N, 3G})$ and $P(B|\text{3N, 3G})$. According to Bayes' Theorem:

$$P(A|\text{3N, 3G}) = \frac{P(\text{3N, 3G}|A) \times P(A)}{P(\text{3N, 3G})}$$

$$P(B|\text{3N, 3G}) = \frac{P(\text{3N, 3G}|B) \times P(B)}{P(\text{3N, 3G})}$$

The denominator $P(\text{3N, 3G})$ is the same for both, so we can compare the numerators directly:

$$P(A|\text{3N, 3G}) \propto \frac{160}{729} \times 0.7$$

$$P(B|\text{3N, 3G}) \propto \frac{5}{16} \times 0.3$$

Calculating these: - $P(A|\text{3N, 3G}) \propto \frac{160}{729} \times 0.7 = \frac{112}{729}$ - $P(B|\text{3N, 3G}) \propto \frac{5}{16} \times 0.3 = \frac{1.5}{16} = \frac{1.5 \times 729}{16 \times 729} = \frac{1.5 \times 729}{11664}$

Comparing these values, $P(B|\text{3N, 3G})$ is greater.

Conclusion

The observed result is more likely with Cage B than with Cage A, given the probabilities calculated. Therefore, the draws are more likely to have come from Cage B.

Final answer: Cage B.

Grader

Step 1: Prior from die roll.

p(A) = 7/10 = 0.7000; p(B) = 3/10 = 0.3000

Prior ratio p(A)/p(B) = 7/3 ≈ 2.333333

Step 2: Cage compositions.

Cage A: p(N—A) = 2/3≈ 0.6667, p(G—A) = 1/3 ≈ 0.3333

Cage B: p(N—B) = 1/2 = 0.5000, p(G—B) = 1/2 = 0.5000

Step 3: Observed outcome and binomial likelihood. Observed: 3 'N', 3 'G' (total 6).

L(A) = 160/729 ≈ 0.219479, L(B) = 5/16 = 0.312500

Likelihood ratio L(A)/L(B) = 512/729 ≈ 0.702332
Step 4: Posterior components and probabilities (fraction and decimal).
p(A)*L(A) = 112/729 ≈ 0.153635, p(B)*L(B) = 3/32 ≈ 0.093750
Post(A) = 3584/5771 ≈ 0.621036, Post(B) = 2187/5771 ≈ 0.378964
Posterior ratio Post(A)/Post(B) = 3584/2187 ≈ 1.638775
Step 5: Decision.
Final answer: Cage A.

Because the correctly computed product for Cage A (≈ 0.1537) is higher than that for Cage B (≈ 0.09375), the decision should have been Cage A. The student's final answer Cage B is inconsistent with the calculations.

Appendix G Validation of the Automated GPT Grader

The analysis in Section 6 employs an automated approach, using a "teacher" GPT grader to evaluate the textual reasoning provided by the "student" GPT models (GPT 3.5, GPT-4, and GPT-4o). To validate the grading ability of the GPT grader, we compare its results with those of a human grader. Both GPT o3-mini and the human grader were tasked with grading a randomly selected subset of 150 textual responses—50 from each student GPT model—from the 6-ball Wisconsin experiment. Additionally, we conduct the same exercise using the more recent but more expensive GPT o1 for comparison with the less expensive GPT o3-mini.
Table 7 presents the error rate for each category (the share of observations with at least one of the error flags in the category is one). The results demonstrate a high degree of concordance between the primary automated grader (GPT o3-mini), the advanced GPT-o1 grader, and the human expert, particularly for Panels A and B, where error rates are nearly identical. This confirms the automated graders' proficiency in identifying fundamental comprehension and conceptual errors related to Bayes' rule application.

Table 7: Comparison of Panel Error Rates (%) Across Graders (N=50 per student model)

Student Model	Panel	Error Category	GPT-o3-mini Grader	GPT-o1 Grader	Human Grader
GPT-3.5	A	Data read-in errors	0	2	2
	B	Bayes' Rule application errors	76	76	76
	C	Posterior calculation errors	96	100	98
	D	Final decision inconsistency errors	6	10	10
GPT-4	A	Data read-in errors	0	0	0
	B	Bayes' Rule application errors	4	4	4
	C	Posterior calculation errors	64	74	82
	D	Final decision inconsistency errors	0	0	2
GPT-4o	A	Data read-in errors	0	0	0
	B	Bayes' Rule application errors	0	0	0
	C	Posterior calculation errors	20	18	22
	D	Final decision inconsistency errors	8	12	14

Notes: Values represent the percentage of the 50 responses flagged with at least one error within the specified panel by the respective graders.

Minor discrepancies arise in Panels C and D, which involve assessing complex numerical calculations and logical consistency. For Panel C, particularly with GPT-4 responses, GPT o1 aligns more closely with the human grader (74% error rate) than GPT o3-mini does (64% error rate, vs. 82% for human). This suggests GPT o1 has a superior, albeit still imperfect, ability to evaluate intricate numerical steps. Similarly, for Panel D, GPT o1 again tracks human judgment more closely, especially for GPT-4o (12% vs. 14% for human, compared to 8% for o3-mini). These differences likely stem from the challenges LLMs face in precisely evaluating complex fraction comparisons and conversions, a task where GPT o1 demonstrates marginally better performance.
Despite these minor variations in evaluating complex numerical reasoning, the overall

agreement across graders is substantial and the main findings in Section 6 are robust to all graders. This validation exercise confirms that GPT-o3-mini serves as a reliable primary grader for the large-scale textual analysis.