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

May 22, 2025

Abstract

We compare human and artificially intelligent (AI) subjects in classification tasks where the optimal decision rule is given by Bayes' Rule. Experimental studies reach mixed conclusions about whether human beliefs and decisions accord with Bayes' Rule. We reanalyze landmark experiments using a new model of decision making and show that decisions can be nearly optimal even when beliefs are not Bayesian. 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 GPT3.5 ignores the prior and is only 75% efficient, whereas GPT 4 and 4o use Bayes' Rule and are 93% and 99% efficient, respectively. Most errors by GPT4/4o are algebraic mistakes in computing the posterior, but GPT4o 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 from helpful feedback from Liu Xiao and Yucheng Liang and seminar participants at University of Texas at Austin and University of 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 the optimal decision rule is given by Bayes' Rule. AI algorithms such as support vector machines or neural networks can be trained to closely approximate optimal Bayesian decision rules given the relative simplicity of this narrow domain classification task. 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 that are trained with large volumes of data to approximate optimal decision rules for specific tasks. Kühl et al. (2022) show that humans possess *general intelligence* whereas most machine learning algorithms are designed to work in narrow domains and will not necessarily make sensible decisions in a huge variety of different (and often unexpected) situations as 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 (LLMs) 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, sounds 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."*

A large literature has studied the rationality of human subjects in a simple binary classification problem used in dozens of previous experimental studies in economics and psychology, which has an optimal solution defined by Bayes' Rule. The consensus is that humans 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 subjects to choose whether 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]

We find this logic compelling, so to give humans the "best shot" we reanalyze 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.

We reanalyze these data using a new structural logit model of decision making that enables us to infer subjects' beliefs using binary choices or 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, about half have small levels of decision noise so their choices frequently coincide with Bayes' Rule, while the other half are subject to large decision noise so their choices frequently differ from Bayes' Rule.

A key contribution of this study is to provide an objective measure of *performance* 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. Since payoffs are based on correctly predicting the selected bingo cage, it is natural to evaluate performance as a ratio of the expected payoffs under the subject's decision rule to optimal expected payoffs under Bayes' Rule. An important insight from the structural 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 noisy Bayesian subjects do not necessarily outperform subjects with biased beliefs.

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. This alllows us to define a simple measure of *efficiency* of a human or GPT subject: the ratio of the expected win probability implied by the subject's decision rule to the optimal win probability under Bayes' Rule. This is a superior 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 (i.e. their choices agree with Bayes' Rule in the same trials), 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 are 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 structural model and experimental design used in the human experiments. We find rapid improvement in the efficiency of GPTs subjects over successive versions of the GPT software. The earliest version, GPT-3.5 (released in 2022), displayed distinctly suboptimal 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]

A key advantage of AI subjects over humans is that AI subjects provide full reasoning behind their answers, broken down into logical steps. As a result, we gain a unique

[3] These experiments were more difficult because subjects were asked to report the probability tha cage A was used to draw the observed sample using a complicated second stage gamble known as the BDM mechansim (described in section 4.3) to incentivize accurate reporting of the posterior.

[4] We also conducted a limited analysis of the most recently released version of ChatGPT o1, and found that its efficiency is close to 100%, with behavior that closely approximates a perfect Bayesian decision maker.

window into their "thought process" and can better isolate where they make their errors. Analyzing the textual responses, we find that the suboptimal performance of GPT-3.5 is due to ignoring prior information, which explains why 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 probability. However, GPT-4o shows remarkable improvement, demonstrating that it understands the applicability of Bayes' Rule as the basis for the decision and make far fewer symbolic algebraic errors in transforming the general formula for Bayes' Rule into specific numerical values for the posterior probabilities.

Why should we care whether GPT is more "Bayesian" than humans? As LLMs evolve and start to outperform highly trained, professional human decision makers in ways that are objectively quantifiable, it has serious implications for "human replacement" even in relatively high skilled occupations. Though the decision problem we study is a simple and somewhat abstract "textbook problem" we believe it is a reasonable approximation to and metaphor for a range of real-world classification and decisions problems that involve significantly higher stakes.

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. Goh et al. (2024) and McDuff et al. (2023)) have demonstrated that LLMs and GPTs can outperform human physicians in the quality and accuracy of their differential diagnoses. However, unlike our problem, there is no objective measure of what the "correct" diagnosis is for most cases. These studies rely on the diagnoses of panels of expert physicians to score the choices made by GPT and less experienced interns and use *ad hoc* scoring rules to evaluate the reasoning that led subjects to choose a particular diagnosis, but without a clear notion of the cost or harm from a misdiagnosis. In comparison, we have an objective way of evaluating loss and classifying "easy" and "hard" problems to identify where subjects make their most costly mistakes.

Our paper contributes to two strands of literature. First, we build on the experimental literature testing human behavioral biases in experimental environment. We show that despite prevalence biases and decision noise humans are surprisingly efficient and nearly optimal decision makers. We show that decision makers with biased beliefs can outperform a noisy Bayesian whose beliefs are unbiased but suffers from high levels of decision noise. We also show that GPT is also susceptible to decision noise including algebraic errors in computing the posterior. Second, we add to the growing literature that compares the behavior of LLMs to humans using experimental methods.[5] Among them, the closest to us are Chen et al. (2023) and Mei et al. (2024), both of which find that GPTs perform similarly to or better than humans. Their conclusions are broadly consistent with our findings, but unlike them we trace the rapid evolution in rationality of GPT over its first three generations, from the non-Bayesian and subhuman performance of GPT3-5 to the superhuman and nearly perfectly Bayesian beliefs and decision making of GPT4o in a span of just 3 years.[6]

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.

[5]The literature find that LLMs excel at textual tasks, for example, divergent thinking tasks (Hubert et al., 2024), ophthalmology exams (Yan et al., 2024; Taloni et al., 2023), essay writing (Herbold et al., 2023), and uniform bar exams (Martınez, 2024). The literature also report that LLMs can fall short of humans, at least in certain aspects, in some non-textual tasks, including complex mathematical proofs and computation (Frieder et al., 2023), analyzing healthcare data (Li et al., 2024), physics coding (Yeadon et al., 2024) and medical board exams (Katz et al., 2024). For additional studies in marketing and finance, see Goli and Singh (2024) and Zhao et al. (2024), and the references there.

[6]Another recent study, Luo et al. (2025), compared the performance of ChatGPT 3.5, 4 and 4o on the 2020/21 Chinese National Medical Licensing Examination (NMLE). They found that "GPT-4o demonstrated significantly higher overall accuracy than GPT-4 and GPT-3.5" and "In the 2020 and 2021 exams, GPT-4o achieved accuracy rates of 84.2% and 88.2%, respectively".

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 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 $\delta(d, \pi, p_A, p_B, D)$ mapping 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

[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 ignore it in the subsequent analysis in this paper.

[8] The BBM 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 4.3 below.

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 probabilistic to the experimenter because the subject's choice depends on additional information or stochastic psychological "decision noise" ν 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 or make use of subjective posterior beliefs about A. In section 3 we introduce a structural 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 perfectly rational Bayesian decision maker.

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

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

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,*

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

An *optimal decision rule* P maximizes the probability of a win, or equivalently it minimizes the probability of a loss. 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*

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

In our comparisons of the performance of human vs AI decision makers, it is convenient to have a single overall scalar summary measure of *decision efficiency* which we define as the ratio of the subject's expected win probability to the optimal expected win probability implied by Bayes' Rule, where we compute expectations over the empirical distributions for the values of the experimental controls in the experiments. For example, we can define an *ex ante* or unconditional expected loss by first taking expectations over the unconditional distribution over the realized values of d given (π,p_A,p_B,D) by

$$\begin{aligned} W_P(\pi,p_A,p_B,D) &= E\{\tilde{W}_P|\pi,p_A,p_B,D\} \\ &= \sum_{d=0}^{D} W_P(d,\pi,p_A,p_B,D)[f(d|p_A,D)\pi + f(d|p_B,D)(1-\pi)]. \end{aligned} \tag{5}$$

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 loss for a subject with decision rule P

in this experiment is given by

$$W_P = E\{\tilde{W}_P\} = \int_\pi \int_{p_A} \int_{p_B} \int_D W_P(\pi, p_A, p_B, D) dH(\pi, p_A, p_B, D). \tag{6}$$

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 structural logit 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 Structural Econometric Model of Subject Responses

This section introduces an econometric model of subject decision making that we refer to as a *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 El-Gamal and Grether (1999). It also includes the optimal Bayesian decision rule as a special case as well. The structural logit 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. A key advantage of the structural logit model is that it enables us to recover estimates of subjective posterior beliefs even when subjects only make binary choices of which cage they believe is more likely to have generated the observed sample d.

We now derive a formula for conditional probability that a subject chooses cage A that depends on the observed information 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 from which the observed sample was drawn. This depends on the public information from the experiment, (d, π, p_A, p_B, D), as well as a private "calculational error" ν that represents algebraic mistakes the 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 probbility as a special case when $\nu = 0$. Specifically, we assume that 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 posterior odds ratio and $\text{LLR}(d, p_A, p_B, D)$ is the log-likelihood ratio given by

$$\begin{aligned} \text{LPR}(\pi) &= \log(\pi/(1-\pi)) \\ \text{LLR}(d, p_A, p_B, D) &= \log(f(d|p_A, D)/f(d|p_B, D)). \end{aligned} \tag{7}$$

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

$$\log\left(\Pi_s(A)/(1-\Pi_s(A))\right) = \beta_0 + \beta_1 \text{LLR}(d, p_A, p_B, D) + \beta_2 \text{LPR}(\pi) + \nu. \tag{8}$$

In our empirical analysis below we assume $\nu \sim N(0, \eta^2)$. Solving equation (8) for $\Pi_s(A|d, \pi, p_A, p_B, D, \nu)$ results in the following logistic specification given by

$$\Pi_s(A|d, \pi, p_A, p_B, D, \nu) = \frac{\exp\{\beta_0 + \beta_1 \text{LLR}(d, p_A, p_B, D) + \beta_2 \text{LPR}(\pi) + \nu\}}{1 + \exp\{\beta_0 + \beta_1 \text{LLR}(d, p_A, p_B, D) + \beta_2 \text{LPR}(\pi) + \nu\}}. \tag{9}$$

Notice that the true Bayesian posterior $\Pi(A|d, \pi, p_A, p_B, D)$ given in equation (1) is a special case of (9) 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* about the posterior probability of cage A, *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 depend 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 that may cause the subject to choose a cage even if it does not have the higher subjective posterior probability. In the presence of these shocks, the subject's decision rule is given by

$$\delta(d, \pi, p_A, p_B, D, \nu, \varepsilon) = \begin{cases} A & \text{if} \quad R\Pi_s(A) + \sigma\varepsilon(A) \geq R\Pi_s(B) + \sigma\varepsilon(B) \\ B & \text{otherwise.} \end{cases} \tag{10}$$

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

$$\begin{aligned} P(A|d, \pi, p_A, p_B, D, \nu) &= \Pr\{\delta(d, \pi, p_A, p_B, D, \nu, \epsilon) = A|d, \pi, p_A, p_B, D, \nu\} \\ &= \frac{1}{1 + \exp\{R[1 - 2\Pi_s(A|d, \pi, p_A, p_B, D, \nu)]/\sigma\}}. \end{aligned} \tag{11}$$

It follows that 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 $P_s(A|d, \pi, p_A, p_B, D, \nu) > 1/2$, and conversely $P(A|d, \pi, p_A, p_B, D, \nu) \to 0$ as $R/\sigma \to \infty$ when $P_s(A|d, \pi, p_A, p_B, D, \nu) < 1/2$.

As we noted, the structural logit model 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 first layer output to determine the probability of choosing cage A, i.e. the CCP. This network is fully determined by a total of 4 "weights" i.e. the three bias/input weights $(\beta_0, \beta_1, \beta_2)$ from the input layer and the single weight R/σ at the output layer.[9] Maximum likelihood of the structural logit model can be interpreted as training the neural network to behave like a human being. The ability of the structural logit 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.[10]

We estimate the parameters using a panel likelihood function since each subject s in the experiment participates in a total of T_s independent trials, so we observe a sequence of binary choices d_{ts} and corresponding experimental control variables (π_{ts}, D_{ts}) for each subject s over trials $t = 1, \ldots, T_s$ assuming (p_A, p_B) remain 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. The likelihood $L(\theta)$ is given by

$$L(\theta) = \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(A|d_{ts}, \pi_{ts}, p_A, p_B, D_{ts}, \theta)^{y_{ts}} [1 - P(A|d_{ts}, \pi_{ts}, p_A, p_B, D_{ts}, \theta)]^{1-y_{ts}}. \quad (12)$$

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 structural logit specification: 1) a 3 parameter *structural probit* that restricts $\sigma = 0$ and normalizes the standard

[9]The formula for the CCP in equation (11) 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 structural model in equation (11) 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".

[10]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.

deviation of the calculational 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 (8) 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 (11) 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

$$\Pi(A|d, \pi, p_A, p_B, D) \leq 1/2 \iff \Pi_s(A|d, \pi, p_A, p_B, D, \theta_\lambda) \leq 1/2. \tag{13}$$

We say that subjective beliefs are *Bayes' compatible* whenever (13) holds.

When there is no decision noise ($\sigma = 0$ and $\eta = 0$), Bayes' compatibility implies that the subject's decision rule is optimal even though their beliefs are not Bayesian. Let L be the *Bayesian classifcation 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 (9) to derive an equivalent *subjective classification threshold* L_s given by the hyperplane $0 = \beta_0 + \beta_1\text{LLR} + \beta_3\text{LPR}$. Thus, an equivalent statement of Bayes' compatibility is that it holds when the *subjective and Bayesian classification thresholds coincide* which holds 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' compatibility, but 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 to suboptimal decisions. The following lemma shows that a subject's posterior beliefs are identified if $\sigma > 0$ and $\eta = 0$ provided there is sufficient "experimental variation".

Lemma L2. Identification of subject beliefs when $\sigma > 0$. *Assume that $\eta = 0$. When $\sigma > 0$, all four parameters of the structural logit model are identified, so the subject's subjective beliefs can be identified from knowledge of their 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 L2 is in Appendix A. Even though Lemma L2 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. Condition (13) 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 4.3 we return to the question of inferring subjective posterior beliefs using the directly elicted 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.[11] Let 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

$$L(\theta, \lambda) = \prod_{s=1}^{S} \sum_{k=1}^{K} \lambda_k L_s(\theta_k), \tag{14}$$

where $L_s(\theta_k)$ is the likelihood function for subject s evaluated at θ_k. We estimated 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.[12]

4 Optimality and revealed beliefs of human subjects

4.1 Reanalysis of California Experiments

We replicated 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.[13] The results are shown graphically in figure 1 which compare 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

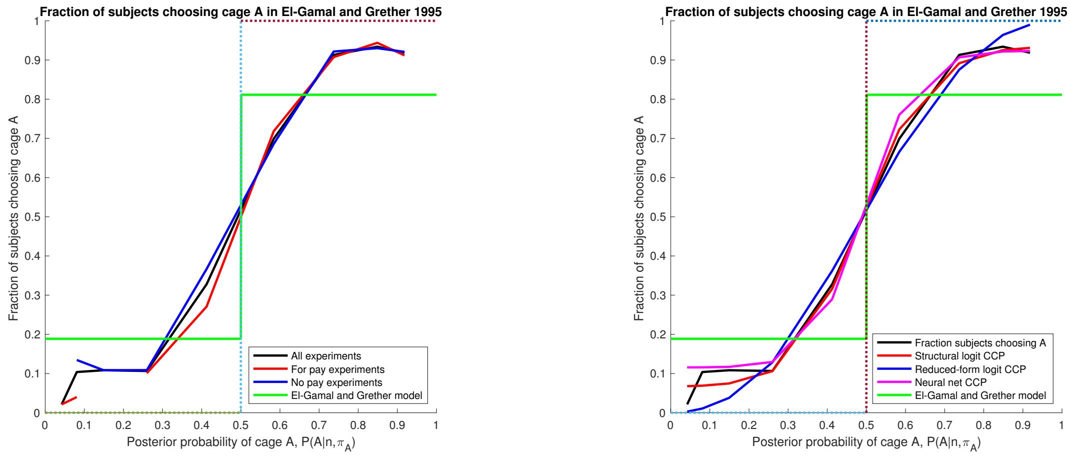
[11] Woutersen and Rust (2025) show that finite mixture models with only a few types can provide highly accurate estimates of the mixed CCP in cases where there are many or even a continuum of types.

[12] 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 (14) 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.

[13] 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.

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.

Figure 1: Comparison of subject behavior and models in the California experiments



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 it has negligible effect on overall behavior.[14] The maximum likelihood predictions from El-Gamal and Grether's model of subject choice are the green curves in both panels. 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.

[14]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.

It is evident from figure 1 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 near the "'hard cases" where the Bayesian 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 1 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 9). 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.[15]

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
Reduced-form logit	3	-1821	3648
Noisy Bayesian	1	-1801	3604
Structural logit	4	-1773	3554
Neural network	5	-1772	3554

[15]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 (11) 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 9, 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: Maximum likelihood estimates of the single type structural logit model

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)

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 display the representativeness heuristic. This differs from their finding using their cutoff rule model that subjects are best described as noisy Bayesians.

Now we show how our conclusions change when we allow for unobserved heterogeneity in subjects' beliefs and behavior. We estimated 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. Rather than presenting the coefficient estimates, we illustrate the predictions and key findings graphically below.

Figure 2: Inferred Posterior beliefs of California subjects

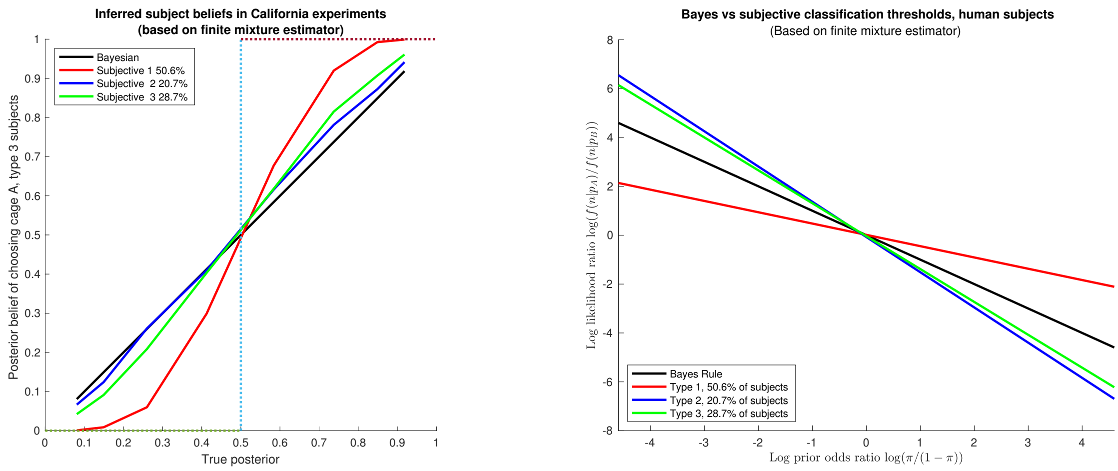
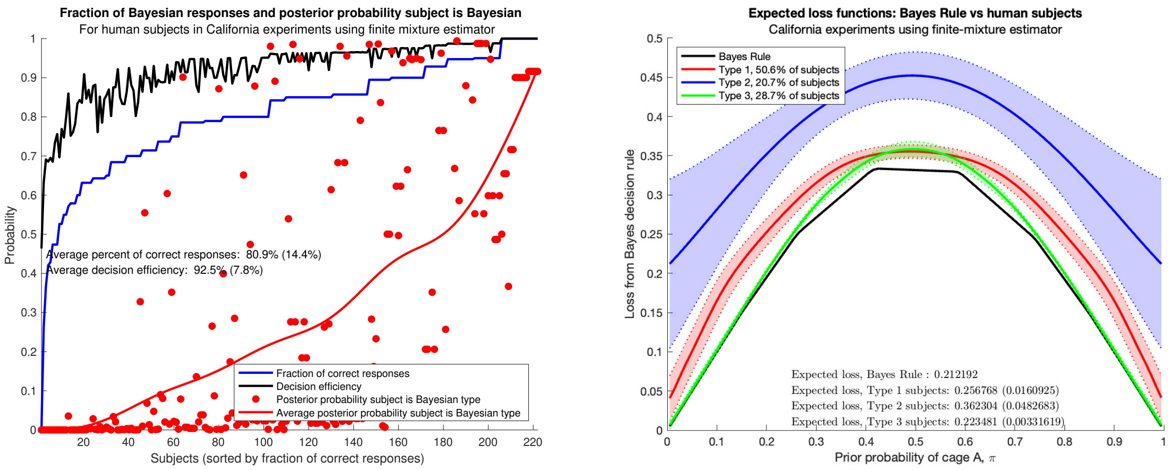


Figure 2 plots the revealed posterior beliefs of the 3 different types of subjects. The left panel plots the estimated subjective posteriors $\Pi_s(A|d, \pi, p_A, p_B, D)$ for against the Bayesian posterior, which is the black 45 degree link. The right panel plots the implied classification thresholds defined in section 3.1. Recall that these are hyperplanes that specify the pairs (LLR, LPR) where the subjective posterior is 1/2, so cage A is more likely for points above these lines and less likely below them. The black line plots the classification hyperlane implied by Bayes' Rule.

The left panel plots the inferred subjective posterior beliefs for each type as a function of the true Bayesian posterior, so the black 45 degree line represents Bayesian beliefs. It is evident that type 2 and 3 subjects have subjective posterior beliefs that are closer to Bayes' Rule than the type 1 subjects (which constitute over 50% of all subjects). The right panel plots the implied classification thresholds. It is less obvious that type 1 is "less Bayesian" from this figure. However we performed Wald tests of the hypothesis of Bayes' compatible beliefs, $\beta_0 = 0$ and $\beta_1 = \beta_2$, and we reject this hypothesis at the 5% level for type 1 subjects but not for type 2 and 3 subjects (P values of .002, .66 and .07, respectively). Type 1 subjects conform to the representativeness heuristic, since the estimated coefficient on LLR is approximately twice as large as the coefficient of LPR. The other two types can be classified as noisy Bayesians since the coefficients on LLR and LPR are not significantly different from each other. However, type 2 subjects are substantially noisier than type 3 subjects: the estimated σ for the former is .75 compared to only .16 for the latter. Thus, we will designate the type 3 subjects as the "most Bayesian" type discovered by the finite mixture method.

Figure 3: Accuracy, efficiency and loss functions for California subjects



The left panel of figure 3 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 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

$$\omega_s = \frac{\sum_{t=1}^{T_s}\left[\Pi(A|d_{ts},\pi_{ts})^{y_{ts}} + [1-\Pi(A|d_{ts},\pi_{ts})]^{(1-y_{ts})}\right]}{\sum_{t=1}^{T_s}\left[\Pi(A|d_{ts},\pi_{ts})^{y^*_{ts}} + [1-\Pi(A|d_{ts},\pi_{ts})]^{(1-y^*_{ts})}\right]}, \tag{15}$$

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 the "most Bayesian type" (i.e. 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

$$\Pi(\tau|y_s, d_s, \pi_s, \hat{\theta}) = \frac{\hat{\lambda}_\tau L(y_s, d_s, \pi_s|\tau, \hat{\theta})}{\sum_{k=1}^{3}\hat{\lambda}_k L(y_s, d_s, \pi_s|k, \hat{\theta})}. \tag{16}$$

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.

The the red line in the left panel plots the local average probability that the subject is the most Bayesian type. We see that this probability is monotonically increasing in the fraction 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 3 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 had the highest efficiency: 97% (0.9%). The average efficiency of all three types of subjects is surprisingly high, 93% even though a minority have beliefs that are well approximated by Bayes' Rule.

Comparing our findings to the main conclusions of El-Gamal and Grether (1995), in their preferred model 63% of subjects were classified as noisy Bayesians and the remaining 37% behaved according to the representativeness heuristic. Our preferred structural model finds fewer noisy Bayesians (49.4%) and more subjects using the representativeness heuristic (50.6%). By introducing the concept of decision efficiency, our analysis reveals the new insight that non-Bayesian subjects can outperform noisy Bayesian subjects when the level of decision noise in the latter is sufficiently high. It follows that decision makers who are "more Bayesian" are not necessarily more efficient. However, Wald tests strongly reject the hypothesis that the subjects are fully efficient decision makers.

4.2 Reanalysis of Wisconsin Experiments

Next we reanalyze data from the "Wisconsin experiments" conducted by El-Gamal and Grether (1999). They recruited 79 student subjects from the University of Wisconsin-Madison and employed a two stage experimental design to test for *context effects* by altering the "California design" used in El-Gamal and Grether (1995) (where $D = 6$ and $p_A = 2/3$ and $p_B = 1/2$) to a new "Wisconsin design" ($D = 7$, $p_A = .4$ and $p_B = .6$). The experiments were conducted on two successive days. On the first day approximately half the subjects began with the 6 ball California design and then switched to the 7 ball Wisconsin design on the second day, whereas for the other half of subjects this order was reversed to see if the ordering of the designs affects subjects' choices.

Table 3 presents the maximum likelihood estimates for a 3 type version of the structural logit model estimated using the 6 ball design for comparability with the results in section 4.1 which also used the 6 ball design.[16]

Table 3: FM estimates of the structural logit model for 6 ball experiments, $LL = -654.25$

Parameter (std error)	Type 1 ($\hat{\lambda}_1 = .32$)	Type 2 ($\hat{\lambda}_2 = .24$)	Type 3 ($\hat{\lambda}_3 = .44$)
σ (noise parameter)	.07 (.08)	.22 (.06)	.32 (.05)
β_0 (bias/intercept)	.14 (.18)	.06 (.11)	-.11 (.08)
β_1 (LLR(n) coefficient)	1.48 (1.91)	1.25 (.36)	2.48 (.67)
β_2 (LPR(π) coefficient)	1.40 (1.81)	2.27 (.70)	1.39 (.36)
P-value for $H_o : \beta_0 = 0, \beta_1 = \beta_2$	.74	.02	.01
P-value for $H_o : \beta_0 = 0, \beta_1 = \beta_2 = 1$	.015	.003	.015

Of the three types, type 1 is the most Bayesian and also the least noisy. The estimated noise parameter σ is less than a third of the value estimated for the other two types. There is no significant overall bias in the beliefs of any of the three types as evidenced by the fact that all three estimates of the bias term β_0 in subjective posterior beliefs in equation (9) is insignificantly different from zero. The next to the last row in the table presents the P values of a Wald hypothesis test that $\beta_0 = 0$ and $\beta_1 = \beta_2$, and we see that there is no evidence against this for type 1 subjects but we reject the hypothesis for type 2 and type

[16]Estimation results using data from the 7 ball design and using the EC rather than the FM algorithm result in similar overall conclusions.

3 subjects. Type 2 subjects have beliefs consistent with representativeness ($\beta_1 > \beta_2$), and type 3 are conservative ($\beta_1 < \beta_2$).

The last row of the table shows that we can strongly reject the hypothesis that any of the three types are noisy Bayesians i.e. who satisfy the hypothesis $H_o : \beta_0 = 0, \beta_1 = 1, \beta_2 = 1$ but allowing decision noise to affect their choices, $\sigma > 0$. However, type 1 subjects satisfy the restriction for Bayes' compatible beliefs $\beta_0 = 0$ and $\beta_1 = \beta_2$ which we showed in section 3.1 implies an optimal decision rule in the absence of decision noise. Thus, we find that type 1 subjects are close to optimal even though their subjective posterior beliefs are distorted relative to Bayesian beliefs.

Figure 4: Fraction of Correct Responses

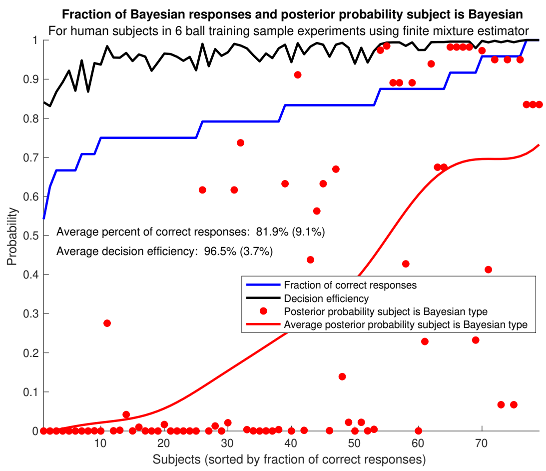
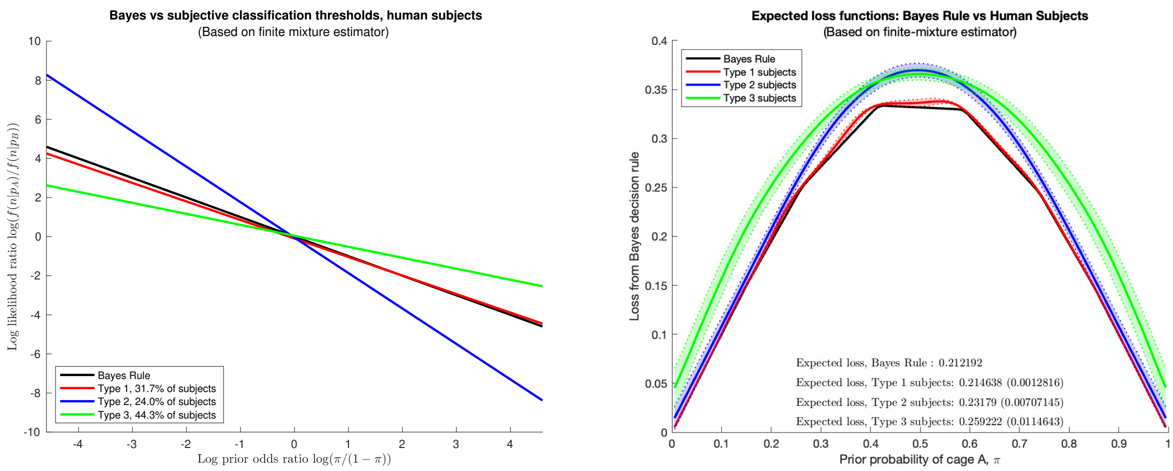


Figure 4 plots the fraction of "correct" responses for each of the 79 subjects, i.e. the fraction of their responses that coincide with the optimal choice of a Bayesian decision maker given the same information (π_t, d_t) in each trial t. We see that the fraction of correct responses ranges from a low of 54% to a high of 100% with an average of over 81% correct responses. The black and red curves help relate this simpler "percent correct" metric of performance to Bayesian behavior. The black curve plots subject-specific decision efficiencies, i.e. the ratio of the sum of each subject's expected win probabilities in all trials to the corresponding total for a perfect Bayesian decision maker. We see that average efficiency is quite high, 96.5%, which exceeds the average of 81.9% correct responses from these subjects. This is due to the fact that most of the mistakes that these subjects made were on the hard cases where the Bayesian posterior probability was close to 1/2, and thus, mistakes for these cases are not very costly for these subjects.

The red curve in figure 4 plots the posterior probability (implied by the FM estimates) that each subject is a type 1 noisy Bayesian and the solid red curve is a local linear regression showing the local average fraction of noisy Bayesians ordered by percent correctly answered. We see a strong positive relation between "correctness" and the probability of being a Type 1 subject, ranging from a low of 0% for subjects who only answered 54% of the trials correctly to a high of 70% for subjects who answered 100% correctly.

Figure 5: Classification Thresholds and Expected Loss Functions, 6 ball experiments



The left panel of figure 5 plots the classification thresholds for each of the three types of subjects along with the prior-specific expected loss functions $L_P(\pi)$, see equation (2). Recall that along these lines the subject believes cages A and B are equally likely, so they each demarcate the decision regions each type of subject would make their classification decisions in the absence of noise (i.e. when $\sigma = 0$). Here it is obvious that the threshold for the Type 1 subjects essentially coincides with the Bayesian classification threshold, but the threshold for the Type 2 subjects is steeper reflecting conservatism ($\beta_2 > \beta_3$), and the threshold for the Type 3 subjects is less steep, reflecting representativeness ($\beta_2 < \beta_3$).

The right hand panel of figure (5) plots the expected conditional loss functions $L_P(\pi)$ for the three types of subjects. The black curve plots the optimal Bayesian loss function $L_{\delta^*}(\pi)$ that necessarily minorizes the loss functions for the human subjects by Lemma L1. We see that the red expected loss curve for the type 1 subjects nearly coincides with the black Bayesian loss function, indicating that the Type 1 subjects are nearly optimal Bayesian decision makers, even though their subjective posterior beliefs are distorted relative to Bayes' Rule (i.e. $\beta_2 > 1$ and $\beta_3 > 1$).

However, the green and blue loss functions are significantly higher than the black and red ones, so it is evident that Type 2 and 3 subjects are using distinctly suboptimal decision rules. The most suboptimal decision makers are the type 3 subjects whose loss function majorizes the other three expected loss functions. Note that the other loss functions are close to zero at values of π equal to 0 or 1, whereas the green curve is significantly positive at those values similar to what we found for the poorest performing types of California subjects. This is a reflection of the high amount of "decision noise" σ for the type 3 subjects that we noted in our discussion of the estimation results in table 3. For example, the model predicts that a type 3 subject will choose cage A about 5% of the time even when π is near 0 or 1.

We also calculated expected win probabilities using the empirical distribution of π for all three types. Type 1 subjects (less noisy Bayesians) have a decision efficiency score of 99.5% (standard error 0.2%), whereas type 2 (Conservative) subjects have a decision efficiency of 96.2% (standard error 0.9%), and type 3 (Representativeness) subjects have an efficiency index of 94.5% (0.1%). While we can strongly reject the hypothesis that any of the three types is a fully optimal decision maker, the overall efficiency of the Wisconsin subjects is remarkably high, 96.5% (0.5%) which is higher than the 93% overall efficiency score for the California subjects.[17]

4.3 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 inefficiency 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

[17] Overall efficiency of the Wisconsin subjects in the 7 ball design was 95.5% (0.6%).

$p_A = 2/3$ 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 $\tilde{p}$. It is not hard to show that the subject's expected payoff from reporting p_r in this second stage BDM lottery is

$$E\{\tilde{G}_R|p_r, d, \pi, p_A, p_B, D\} = R\left[\frac{1-p_r^2}{2} + p_r \Pi_s(A|d,\pi,p_A,p_B,D)\right], \tag{17}$$

where $\Pi_s(A|d,\pi,p_A,p_B,D)$ is the subjective posterior probability for cup A. It is not hard to see that the report p_r that maximizes expected payoff is $p_r^* = \Pi(A|d,\pi,p_A,p_B,D)$.

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 (17), 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.[18]

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

[18] 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".

such a choice (perhaps incentivized by an additional payment for selecting the correct cup) that they would have chosen cup A if $\Pi(A|d,\pi,p_A,p_B,D) > 1/2$ and cup B otherwise.[19] This implies that $\sigma = 0$ in our structural logit model specification in equation (10), 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 prior odds ratio regression specification in equation (8).[20] We also estimated 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).[21] We omit the 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. For the noisy Bayesian subjects, we cannot reject the hypothesis that $(\beta_0^*, \beta_1^*, \beta_2^*) = (0,1,1)$ (i.e. values that result in Bayesian beliefs), though there is significant "calculational noise" as evidenced by the large and significant estimate of $\hat{\eta} = 0.91$ (std error (0.14)). We strongly reject this hypothesis for the type 2 "conservative" subjects and find a small bias against choosing cup A. However, we note that the type 2 subjects are far less noisy than the type 1's with an estimated value of $\hat{\eta} = 0.40$, less than half the value we estimated for type 1 subjects.

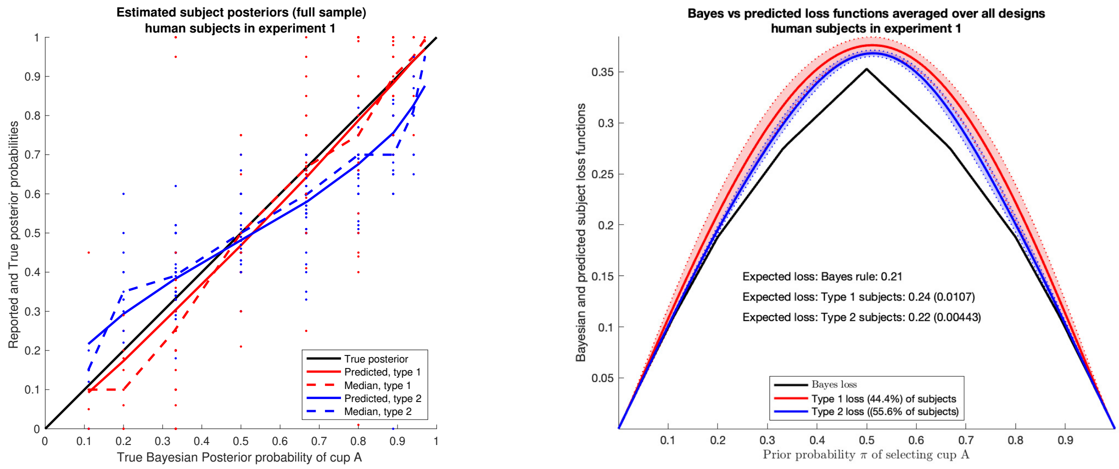
The left hand panel of figure 6 provides a scatterplot of the subject responses plotted as (LPR, LLR) pairs for each subject and trial. For reference the black 45 degree line is the Bayesian posterior probability and the dashed lines are the median values of the subjects' responses. Using the posterior probability in equation (16) we can classify each subject according to whether which posterior probability is more likely: either type 1 (noisy

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

[20]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 estimated 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 verified 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.

[21]The results from the EC and finite mixture models are quite similar, both in the parameters and the estimated fractions of each type.

Figure 6: Predicted vs Actual Median Beliefs and Loss Functions: Holt Smith Experiment 1



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 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' compatabile 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 6 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%.[22] 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 fid for subjects in El-Gamal and Grether's California and Wisconsin experiments.

[22]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.

5 Optimality and revealed beliefs of AI subjects

We will now compare the performance of AI and human subjects by conducting new experiments, *using AI subjects,* that replicate the same design of the experiments conducted by El-Gamal and Grether and Holt and Smith using human subjects. Our new experiments use three versions of of the ChatGPT from OpenAI: 1) GPT-3.5 (introduced in 2022), 2) 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 designed prompts that contained the necessary information for GPT to provide an answer using the realized outcomes of the human experiments (e.g. composition of the cages, selection of the priors, number of draws from the chosen cage, etc). We replicated the Wisconsin experiments, where subjects simply chose the cage they thought was more likely to have been used to draw the observed sample, using the same design as El-Gamal and Grether (1999). We also conducted a second set of experiments using the design of Holt and Smith (2009) where subjects were asked to report their subjective posterior probability that the sample came from cage A, using the same BDM mechanism decribed in section 4.3 to incentivize subjects to report their posterior beliefs as accurately as possible. We then use the same structural logit model introduced in section 3 to provide a parallel analysis of the GPT data and test whether subjects are Bayesian. We also allow for unobserved heterogeneity using both the EC and finite mixture methods just as we did for human subjects. Why would we expect unobserved heterogeneity among GPT subjects? In fact, we intentionally created heterogeneity among the GPTk subjects by choosing different values for the *temperature parameters* of the GPT subjects. This parameter controls the scale of random "noise" in their responses and thus it is akin to heterogeneity in the extreme value σ parameter in equation 11 or the parameter η governing the variance of "calculational errors" ν in the subjective posterior probability $\Pi_s(A|d, \pi, p_A, p_B, D, \nu)$ in equation (8).[23] We show that the EC and finite mixture methods can effectively detect and control for the variability in temperature, which in the case of GPT subjects is actually an observed covariate rather than unobserved heterogeneity.

[23]Lower values of the temperature parameter result in more deterministic and focused responses, where GPT is more likely to choose the most probable next token. A higher temperature increases randomness, allowing for more varied responses. The default temperature for the three versions of GPTs we consider here is 0.7. We varied the temperature randomly across subjects from 0.3 to 1.5.

5.1 Prompt Design for Experiments in ChatGPT

Similar to Chen et al. (2023), gathered data from GPT subjects by submitting inquiries through the public OpenAI application programming interface (API). Using APIs allows us to conduct massive experiments in a timely and cost-effective manner. We formulated prompts by drawing an analogy between GPTs and the human experiments where each GPT subject was matched to a subject in the corresponding human experiment and the same experimental data was provided to the GPT subject. This gives us a unique ability to strongly control outcomes via an exact match between the information that the human and GPT subjects received. Appendix B outlines the algorithm we developed to implement the experiments, which involves looping over different versions of GPTs, temperature settings (subjects), experiments, and trials. We then parsed the responses from ChatGPT and collect its choices of either Cage A or B in the Wisconsin experiment,[24] or the reported subjective probability in the case of the Holt and Smith experiment. Appendix C provides further details about the prompts we used.

One notable difference between the prompt in our study and that of Chen et al. (2023) is that we allow GPTs to report their reasoning process, rather than solely providing the conclusion as done in Chen et al. (2023).[25] We deviate from them for two reasons. First, in the daily use of GPTs, people rarely impose such restrictions. Second, when answering our questions, GPTs typically answer them step by step. This chain of thought may potentially enhance GPT performance, as suggested by Wei et al. (2022). In section 6 we directly analyze the full textual reasoning that lead to the ultimate answer in order to better understand GPT's reasoning process.

5.2 Analysis of GPT subjects in the Wisconsin experiments

We begin by analyzing the binary classification results from the Wisconsin experiments conducted with the GPTs. Following our re-analysis of human subjects, instead of pooling the data from the 6 and 7 ball design we focus on the 6 ball design. We estimated

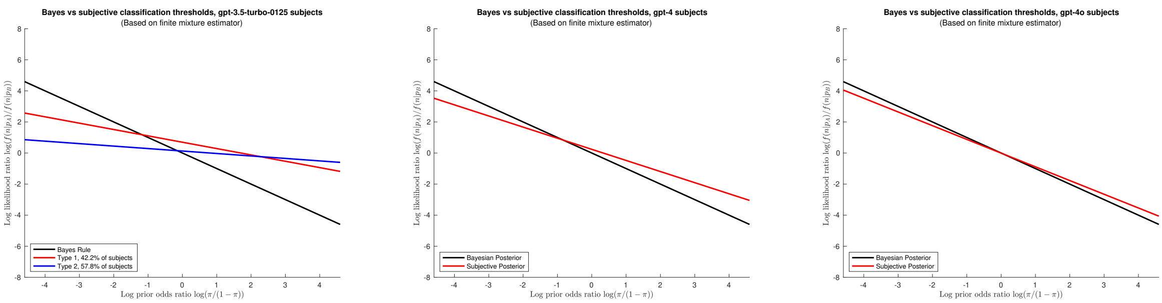
[24]Occasionally, GPTs may stop prematurely before providing an answer regarding the choice between A and B. In such cases, we resubmit the same inquiry until the GPT delivers a classification. We consider this process to be natural, as it mirrors our everyday use of ChatGPT—if it fails to provide a 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.

[25]For example, the prompt for the investment experiment in Chen et al. (2023) ends with "First please *only* tell me the number of points for investing Asset A, then please *only* tell me the number of points for investing Asset B." Including the word *only* forces GPTs to report only the final conclusion without the reasoning process.

a multiple-type structural logit model using the FM method.[26] The estimates detect unobserved heterogeneity among GPT subjects, with different subjective posterior beliefs (determined by the belief parameters β's) and level of idiosyncratic noise (captured by the extreme value scale parameter σ) associated with each type.

The results are presented graphically in figure 7. The FM method identified two types for GPT-3.5 subjects, indicated by the implied classification hyperplanes given by the red and blue lines in the left most panel of the figure. Of the two types we find significant difference in the estimated σ noise parameter, with 1 subjects estimated to have smaller values than type 2 subjects. The difference is due to intentionally injected heterogeneity via the randomly assigned temperature parameters for the GPT 3.5 subjects. The mean temperature for the type 1 subjects is 0.43, nearly half that of the noisier type 2 subjects. Thus, we conclude that higher temperatures correspond to greater noise in subject responses and provides a reassuring confirmation of the effectiveness of the EC and FM algorithms for detecting unobserved heterogeneity. For GPT versions 4 and 4o we only find a single type, and the implied classification hyperplanes for these types are plotted in the middel and right most panels of figure 7, respectively. We see a clear progression toward decision rules that more closely approximate the optimal decision rule implied by Bayes' Rule in the figure.

Figure 7: Estimated classification hyperplanes for GPT Subjects: 6-ball Experiments



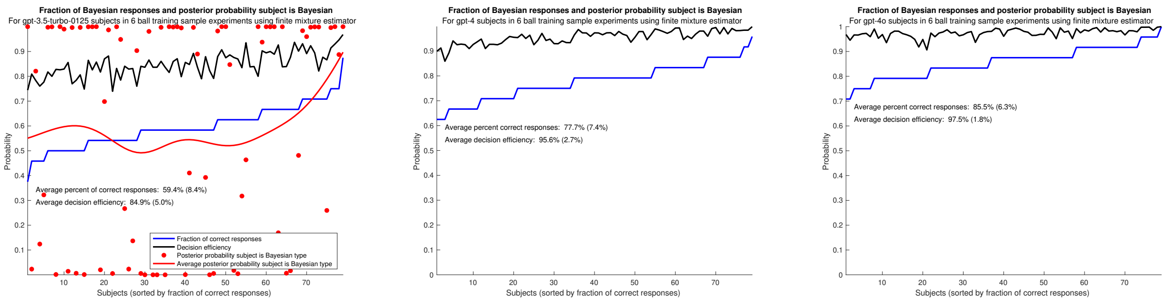
(a) GPT 3.5 (b) GPT-4 (c) GPT-4o

The FM method only identifies a single type of GPT-4 and GPT 4o subjects, respectively. The estimated posterior beliefs for these subjects are almost flat, but they do have

[26] We also estimated the model using the EC algorithm. Although the EC algorithm typically identifies more types than the FM method, we prefer the FM method for two reasons. First, some types identified by the EC algorithm may be redundant due to subtle parameter differences or a small number of subjects in certain types. Second, starting from the estimates of the EC algorithm, the FM method can always consolidate these into fewer types.

a tiny positive slope, leading to the same CCPs as Bayesian decision makers.[27] Their nearly flat posterior beliefs conceal differences in subjective beliefs between GPT-4 and GPT-4o. Figure 7 shows the implied classification hyperplanes, whose slopes equal the negative ratio of coefficients for LLR ($\widehat{\beta_1}$) and LPR ($\widehat{\beta_2}$). The Bayesian classification hyperplanes, which have a slope of -1, are in black. Although the β values are small, the ratios are close to -1, so GPT-4o subjects assign nearly equal weights to LLR and LPR and their subjective classification hyperplane is closer to the Bayesian hyperplane than GPT-4, which places more weight on LPR than on LLR.

Figure 8: Accuarcy and Efficiency of GPT Subjects in Wisconsin 6-ball Experiments



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

Figure 8 illustrates the rapid improvement in the performance of the three generations of GPTs. We calculating accuracy and decision efficiency using the formulas in section 2. The blue step function in Figure 8 shows that accuracy varies from below 40% to 90% among GPT 3.5 subjects, with a very low average of just 59.4%. The accuracy curve moves up for GPT-4 and rises further for GPT-4o, highlighting the improvement in accuracy across successive generations of GPTs. We also see observe a rapid improvement in decision efficiency, given by the black curves in figure 8. When there are multiple types, as in the case of GPT 3.5, we identified the most Bayesian type and used the posterior probability to determine the most likely type of each subject, represented by the red dots. The red Bayesian-type curve plots the average posterior probability of the most Bayesian type, estimated using a local linear regression of these red dots. Similar to what we found for human subjects, there is a fairly monotonic relationship between accuracy, efficiency

[27] We estimate the model from multiple starting values, all converging to the same single-type estimate. To further validate that GPT-4o acts as a single-type "noisy Bayesian" decision maker, with subjective posterior beliefs showing a slight positive slope against true posteriors, we estimate a noisy Bayesian model with β parameters constrained to $[0, -1, -1]$, estimating only σ. The likelihood ratio test does not reject the noisy Bayesian model.

and the probability of being the "most Bayesian" type of GPT3.5 subject.

Table 4 summarizes the overall performance of humans and various GPTs in the Wisconsin 6 ball design experiments, averaging across all estimated types. In terms of both efficiency and accuracy the GPT-3.5 subjects perform significantly worse than humans. Although humans are comparable to GPT-4, they are less efficient than GPT-4o. This underscores the rapid transition in GPT performance from "subhuman" to "superhuman" with just a few upgrades over a relatively short period.

The bottom row in Table 4 reports the number of subject types detected by the FM method. We find that there are fewer types in the later versions of GPT, as their behavior becomes increasingly Bayesian. We find more heterogeneity among human subjects than GPTs, whose behavior is more uniform, despite the intentional introduction of heterogeneity via temperature.

Table 4: Performance of GPTs and humans in the 6-Ball Experiments

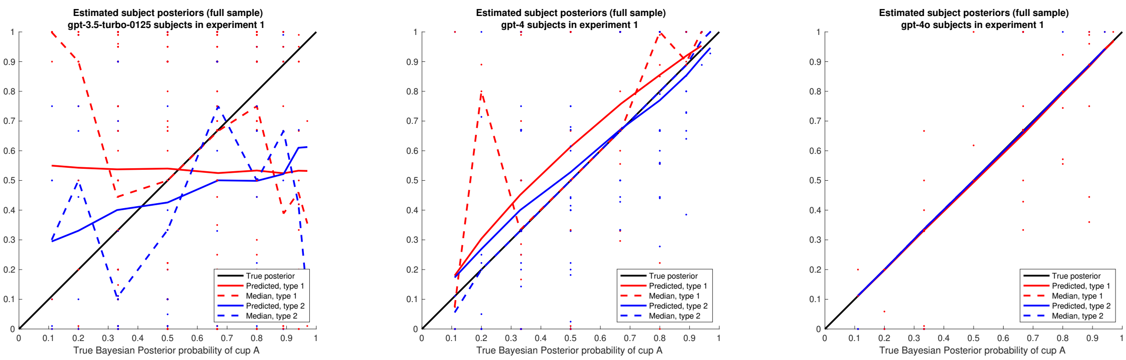
	GPT 3.5	GPT-4	GPT-4o	Humans
Efficiency	84.9	96.0	97.5	96.5
	(0.6%)	(0.3%)	(0.2%)	(0.5%)
Accuracy	59.4	77.7	85.5	81.9
	(8.4%)	(7.4%)	(6.3%)	(9.1%)
No. of Types	2	1	1	3

5.3 Analysis of GPT subjects in the Holt and Smith experiments

Now we turn to an analysis of elicited posteriors using the Holt and Smith design with GPT subjects. These data allow us to directly observe subjective posterior beliefs which are otherwise challenging to identify when subjects report only binary choices. We replicated both the experiments run at the University of Virginia (experiment 1) as well as online (experiment 2). In contrast to our reanalysis of human data, we find that the estimated subjective posterior beliefs are very similar for GPT subjects in the two experiments. This is not surprising, as humans may be more easily distracted in online than in offline experiments, as suggested by the larger variance in the error terms. Such differences do not appear in experiments with GPTs. Therefore, we focus on presenting the results from experiment 1 where the average human decision efficiency was higher, 96% compared to 91% for human subjects in experiment 2.

Figure 9 plots, for the 3 versions of GPT subjects, their reported posterior probabilities against the corresponding Bayesian posterior probabilities, each represented as a dot.[28] The FM algorithm identifies two types of subjects for all three GPTs, denoted by red dots (type 1) and blue dots (type 2), respectively.

Figure 9: True vs Estimated Subjective Posterior Probabilities



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

The dotted lines illustrate the median reported posterior probabilities derived from the data. It is evident that each successive generation of GPT subjects report posterior beliefs that increasingly align with the true Bayesian posterior, as evidenced by their proximity to the 45-degree Bayesian line. GPT-3.5 frequently makes numerical mistakes in calculating its posterior, with both the frequency and severity of these inaccuracies escalating when true posterior probabilities are more extreme, particularly in the ranges below 0.2 and above 0.8, where discrepancies are especially pronounced. For GPT-4, the median subjective posterior probabilities of type 1 subjects closely align with the Bayesian 45-degree line, except that they may slightly overestimate the posterior probabilities when the true posteriors are close to 1. Type 2 subjects also approximate Bayesian decision-making in most trials, except for those with true posterior probabilities near 0, where subjects tend to underestimate probabilities, and near 1, where they tend to overestimate. The median subjective posteriors for both types of subjects in GPT-4o are almost identical to the Bayesian decision makers.

The solid red and blue lines in Figure 9 represent the model-predicted median of subject responses, constructed from the estimated β's. Consistent with the findings from the Wisconsin experiments, we observe a markedly improved model fit as subjects approach Bayesian decision-making. GPT-3.5 exhibits numerous errors in its responses,

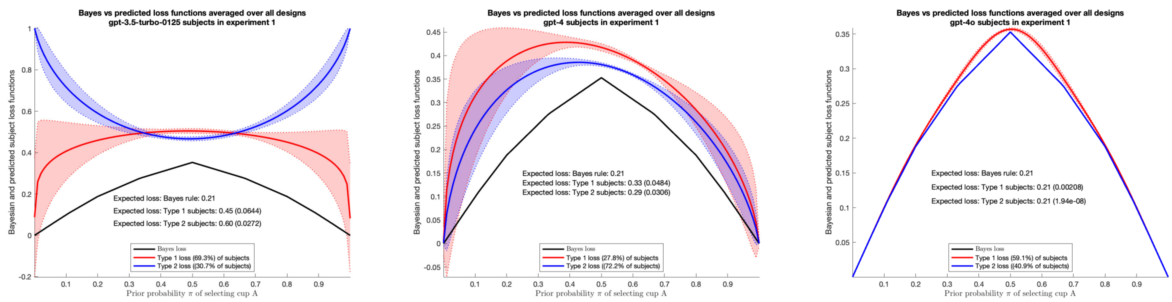
[28]In one trial, a subject from GPT 3.5 reported a posterior probability of 2, which is excluded from Figure 9. However, we do not exclude this observation from our sample.

which makes it challenging for the simple structural logit model to capture the full range of atypical behaviors.

The structural logit model fits much better in GPT-4 and GPT-4o. However, both subject types identified in GPT-4 consistently overestimate posterior probabilities when the true posteriors are below 0.7, while underestimating them otherwise. Both types assign less weight to LPR, however, the type 1 subjects are closer to Bayesian, as evidenced by a smaller difference between the coefficients on LPR and LLR.

The model-predicted and data posterior probabilities both coincide with the Bayesian 45-degree line, illustrating an almost perfect model fit. Both types of GPT-4o subjects put almost equal weight on LPR and LLR and have negligible bias, so their posterior beliefs are close to Bayesian. The two types are different mainly in their estimated noise parameters. The estimated noise parameter $\widehat{\eta}$ reveals another aspect of improvement, namely, the level of calculation noise in their subjective beliefs reduces in each successive generation. For GPT 3.5, the degree of noise is large for both types, with $\widehat{\eta} = 2.8$ for type 1 subjects and $\widehat{\eta} = 2.1$ for type 2 subjects. In contrast, the level of noise is significantly smaller for GPT-4, with $\widehat{\eta} = 0.8$ for type 1 subjects and $\widehat{\eta} = 1.6$ for type 2 subjects. By the time we get to GPT-4o there is almost no estimated noise for type 2 subjects, $\widehat{\eta} = 0.0025$, which means that effectively they are perfect, noiseless Bayesians. In contrast, type 1 subjects have a slightly larger noise of $\widehat{\eta} = 0.4$, making them noisy Bayesian decision makers.

Figure 10: Loss Functions by Types



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

Figure 10 plots the implied loss functions for the three generations of GPT subjects. We see steady improvement in performance with implied loss functions from quite suboptimal behavior for GPT3.5 subjects (with an "inverted loss" function for type 2 subjects that implies highest loss at the easiest cases where the prior is near 0 or 1) to nearly optimal decision rules with a loss function that nearly concides with the optimal Bayesian loss function for GPT4-o. Table 5 compares the performance of human and GPT subjects in Holt and Smith's experiment 1. The table reports our decision efficiency and accuracy scores as well as a new measure of "accuracy", namely, the R^2 of a regression of the reported posterior on a constant and the true posterior. Values closer to 1 for each of these measures can be interpreted as "more Bayesian". We see a steady progression in efficiency and accuracy of the GPT subjects from GPT 3.5 to GPT 4o. The human subjects are superior to GPT 3.5 and 4 in terms of efficiency, but GPT-4o shows superhuman performance, especially in terms of decision efficiency and accuracy as measured by R^2, reflecting the fact that reported posteriors by GPT subjects are significantly closer to the true Bayesian posterior on average than the noisier human subjects.

The FM algorithm finds 2 types of subjects for each version of GPT and human subjects, and in each case the key difference is the level of estimated noise (captured by the standard deviation parameter η representing "calculational errors" in the structural logit model). Except for GPT 3.5, both types of subjects are "noisy Bayesians" but one of the types makes significantly larger calculational errors than the other. Among the GPT-4o subjects, 41% are classified as essentially perfect noiseless Bayesians with 100% decision efficiency. For human subjects, 45% are Bayesians also, but due to the greater level of noise in their response their decision efficiency is lower, 95%. For GPT 4 and GPT 4o this noise correlates directly with the "temperature" parameter we set to generate heterogeneity among the GPT subjects. For example for GPT-4o, the noisier subjects had average temperature of .75 (std error .30), twice the an average temperature of the less noisy subjects identified by the FM algorithm. Thus, as we would expect higher temperatures imply noisier responses by GPT subjects and this noise leads to lower efficiency and accuracy. We can obtain an even better performance for the GPT-4o subjects by reducing the temperature parameter.[29]

[29] We also tested for contextual effects by asking GPT subjects to choose which of two ponds a fisherman might have been fishing on given the observed catch, knowing the prior and the different proportions of two types of fish in each pond. This change had no significant effect on GPT's responses, suggesting that GPT's performance is not due to data mining that discovers an association between Bayes' Rule and bingo cage experiments in statistical texts and the articles we cite in this study. Thus it seems that changes context and experimental design does not "fool" GPT subjects. GPT 4 and 4o appear to recognize that Bayes' Rule applies in a variety of contexts.

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
	(0.1)	(0.1)	(0.02)	(0.1)
R^2	0.7	41.8	88.0	63.5
No. of Types	2	2	2	2

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 comparitively 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 D.

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 reviewed 50 randomly sampled textual responses for each model to verify the GPT grader's performance. Overall, our independent cross-check confirmed the accuracy of GPT o3-mini's grading and classification of errors in the responses from ChatGPT 3.5, 4 and 4o.[30]

[30]In Appendix F, we report the error rates for the same set of 50 samples graded by 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.

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.[31] 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 this information can explain choices that are consistent with conservatism and representativeness. 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, which checks whether the prior probability,[32] the likelihood, and the posterior (or posterior odds ratio) are correctly calculated.[33]

The final category under Panel D examines whether the final decision (Cage A or B) aligns with the previous reasoning, errors in which are reflected by the decision noise term ε in the structural logit model. We instructed the GPT grader to read 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 asked the GPT grader to explicitly predict the outcome based on the reasoning

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

[32]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 E for an example.

[33]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.

just before the final answer and compare whether this prediction is consistent with the subjects' reports.

6.2 Grading Results

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

We applied 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 to each error flag, with 1 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.[34]

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,

[34] 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 = YES), they will also have errors in calculating the prior (error flag 6 = YES) and subsequently the posterior (error flag 8 = YES). 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.

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 "conceptual Bayesian" to "perfect Bayesian" amoung GPT-4 and GPT-4o subjects. In GPT-4o, 78% of responses correctly calculate the posterior, compared to just 28% for GPT-4. 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 comparing the numerator and denominator of the posterior odds. To understand it, we relied on a more detailed analysis of a sample graded by a human grader to confirm this observation.[35] First, we find that GPT-4o is more likely to report the posterior as fractional numbers than GPT-4,[36] probably 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[37], making comparisons for the final decision more challenging.[38] Therefore, although GPT 4o can calculate the posterior more accurately, it is more susceptible to final decision errors in the process of comparing these numbers.

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

[36]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.

[37]See Appendix E 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.

[38]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 when using rounded decimals.

7 Conclusion

Who is more Bayesian? We compared humans to three generations of chatGPT subjects — 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 earliest version, GPT3.5, is not Bayesian because it ignores prior information. In contrast GPT4 and 4o use Bayes Rule but make algebraic errors in evaluating the Bayesian posterior, so they are noisy Bayesians. However the incidence of these errors is dramatically lower in GPT4o so it is 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 Kahneman (1974) and other studies in pyschology and behavioral economics documented "significant deviations from the Bayesian theory of judgment under under uncertainty." We show it is possible to behave like a Bayesian even if beliefs are non-Bayesian. Appealing to statistical decision theory, we introduced an objective measure of what it means to be "more Bayesian", namely *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 show 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. 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% and nearly 99% for the minority of best performing humans. In comparison, GPT3.5's efficiency ranged between 75-85% and GPT4o's efficiency ranged from 97 to 99% across the experiments we analyzed.

We have documented a significant improvement in GPT's performance, from subhuman to superhuman in just three years. GPT now outperforms most humans in a wide variety of other intellectual domains. It 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 is a cause for celebration or panic is beyond the scope of this paper.

References

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.

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

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.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023. URL https://arxiv.org/abs/2301.13867.

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

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

Ali Goli and Amandeep Singh. Frontiers: Can large language models capture human preferences? *Marketing Science*, 43(4):709–722, 2024.

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

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.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. Ai, write an essay for me: A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13(1):1–13, 2023. doi: 10.1038/s41598-023-45644-9. URL https://www.nature.com/articles/s41598-023-45644-9.

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.

Kent F. Hubert, Kim N. Awa, and Darya L. Zabelina. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14:1–10, 2024. doi: 10.1038/s41598-024-53303-w. URL https://www.nature.com/articles/s41598-024-53303-w.

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.

Uriel Katz, Eran Cohen, Eliya Shachar, et al. Gpt versus resident physicians — a benchmark based on official board scores. *NEJM AI*, 1(5):5, 2024. doi: 10.1056/AIdbp2300192. URL https://ai.nejm.org/doi/pdf/10.1056/AIdbp2300192.

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.

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.

Kevin Danis Li, Adrian M. Fernandez, Rachel Schwartz, Natalie Rios, Marvin Nathaniel Carlisle, Gregory M. Amend, Hiren V. Patel, and Benjamin N. Breyer. Comparing gpt-4 and human researchers in qualitative analysis of healthcare data: Qualitative description study. *Journal of Medical Internet Research*, 26:e56500, 2024. doi: 10.2196/56500. URL https://www.jmir.org/2024/1/e56500.

Dingyuan Luo, Mengke Liu2, Runyuan Yu, Yulian Liu, Wenjun Jiang, Qi Fan, Naifeng Kuang, Qiang Gao, Tao Yin, and Zuncheng Zheng. Evaluating the performance of gpt-3.5, gpt-4, and gpt-4o in the chinese national medical licensing examination. *Nature Portfolio*, 15:14119, 2025.

Eric Martınez. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, 2024. doi: 10.1007/s10506-024-09396-9. URL https://link.springer.com/article/10.1007/s10506-024-09396-9.

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.

Andrea Taloni, Massimiliano Borselli, Valentina Scarsi, Costanza Rossi, et al. Comparative performance of humans versus gpt-4.0 and gpt-3.5 in the self-assessment program of american academy of ophthalmology. *Research Square*, 2023. doi: 10.21203/rs.3.rs-3206650/v1. URL https://www.researchsquare.com/article/rs-3206650/v1.

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.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Jing Li, Xianchao Zhu, and Yue Zhang. Benchmarking gpt-4 against human translators: A comprehensive evaluation across languages, domains, and expertise levels. *arXiv preprint arXiv:2411.13775*, 2024. URL https://arxiv.org/abs/2411.13775.

Will Yeadon, Craig P. Testrow, and Alex Peach. A comparison of human, gpt-3.5, and gpt-4 performance in a university-level coding course. *arXiv preprint arXiv:2403.16977*, 2024. URL https://arxiv.org/abs/2403.16977.

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.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*, 2024.

Appendix A Proof of Lemma L2

This appendix provides the proof of identification of the structural logit model when $\sigma > 0$, Lemma L2 of section 3.1. We start by assuming that the agent's decision rule $P(A|\pi, n)$ is identified for all possible values of $\pi \in [0, 1]$ and all possible experiments involving draws from one of two bingo cages where the number of draws from these cages, and hence the possible values of n, can be arbitrarily large. In fact, because we assume that the experimenter can run an arbitrary number of experiments and observe the subject outcomes, the experimenter has the freedom to design experiments with arbitrary values for the log-likelihood ratio LLR and the log-prior odds ratio LPR, so we assume these quantities can take any value between $-\infty$ and $+\infty$ via an appropriately designed set of experiments. Further, since the subject's decision rule depends on (π, n) via the quantities $(\mathrm{LPR}(\pi), \mathrm{LLR}(n))$ and the experimenter has the freedom to design experiments that allow LPR and LLR to take on any values in R^2, we can treat the subject's decision rule as a known function of two continuous arguments, $(\mathrm{LPR}, \mathrm{LLR})$, which constitute "sufficient statistics" for the subject's posterior beliefs, and hence for the subject's decision rule. In short, we can assume that the subject's decision rule is a known conditional probability of the form $P(A|\mathrm{LPR}, \mathrm{LLR})$ mapping R^2 into $[0, 1]$.

Now, let the subject's true decision rule be given by the function

$$P(A|\mathrm{LPR}, \mathrm{LLR}, \sigma^*, \beta^*) = \frac{1}{1 + \exp\{[2\Pi_s(A|\mathrm{LPR}, \mathrm{LLR}, \beta^*) - 1]/\sigma^*\}}, \tag{18}$$

and by assumption, this function is identified, i.e. its value is known for any pair $(\mathrm{LPR}, \mathrm{LLR}) \in R^2$. Now suppose there is some other structural logit that is observationally equivalent to the true one. That is, suppose there is a function $P(A|\mathrm{LPR}, \mathrm{LLR}, \sigma, \beta)$ such that

$$P(A|\mathrm{LPR}, \mathrm{LLR}, \sigma^*, \beta^*) = P(A|\mathrm{LPR}, \mathrm{LLR}, \sigma, \beta), \quad \forall(\mathrm{LPR}, \mathrm{LLR}) \in R^2. \tag{19}$$

We will know that that if this is true, then it must be the case that $\sigma = \sigma^*$, and $\beta = \beta^*$. That is, the parameters of the structural logit model are identified.

We can show this under a *full support condition* which is a version of an "identification at infinity" type of argument. The full support condition implies that the experimenter can conduct sufficient experimentation on subjects that the log-prior ratio LPR takes on any value on the real line. It follows that by taking the limit as $\mathrm{LPR} \to \infty$ we have $\Pi_s(A|\mathrm{LPR}, \mathrm{LLR}) \to 0$. Since equation (19) holds for all $(\mathrm{LPR}, \mathrm{LLR}) \in R^2$ and is a continuous function of these variables, it follows that the equality must hold in the limit so we have

$$\frac{1}{1 + \exp\{1/\sigma^*\}} = \frac{1}{1 + \exp\{1/\sigma\}}, \tag{20}$$

which implies that $\sigma^* = \sigma$, so this parameter is identified. Using this result we can immediately conclude from equations (18) and (19) that

$$\Pi_s(A|\mathrm{LPR}, \mathrm{LLR}, \beta^*) = \Pi_s(A|\mathrm{LPR}, \mathrm{LLR}, \beta). \tag{21}$$

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

$$\beta_0^* + \beta_1^* \mathrm{LPR} + \beta_2^* \mathrm{LLR} = \beta_0 + \beta_1 \mathrm{LPR} + \beta_2 \mathrm{LLR}, \quad \forall(\mathrm{LPR}, \mathrm{LLR}) \in R^2. \tag{22}$$

Let LPR = LLR = 0 (which is possible due to the full support assumption). If follows from equation (22) that $\beta_0^* = \beta_0$. Next, set LLR = 0 and LPR = 1, and it follows that $\beta_1^* = \beta_1$. Finally, set LLR = 1 and LPR = 0, and it follows that $\beta_2^* = \beta_2$. We conclude that the parameters $(\sigma^*, \beta_0^*, \beta_1^*, \beta_2^*)$ of the structural logit model are identified.

Appendix B 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 C 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 D 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 E 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 F Validation of the Automated GPT Grader

The analysis presented in Section 6 utilizes an automated approach, employing a "teacher" GPT grader to evaluate the textual reasoning provided by the "student" GPT models (GPT-3.5, GPT-4, and GPT-4o). To ascertain the reliability of this method, a validation exercise was conducted. The validation methodology involved a randomly selected subset of 150 textual responses (50 for each student model: GPT-3.5, GPT-4, GPT-4o) from the 6-ball Wisconsin experiment dataset. These responses were independently evaluated by three graders: a human co-author expert in the task, the GPT-o1 model, and the primary GPT-o3-mini grader. All graders utilized the finalized grading prompt detailed in Appendix D. Each response was assessed against the 9 binary grading flags described in Section 6.1, which were subsequently grouped into four conceptual panels: Panel A (Data Read-in), Panel B (Bayes Rule Application), Panel C (Posterior Calculation), and Panel D (Final Decision Consistency). Aggregate panel error rates were calculated, where a panel error was recorded if any underlying flag indicating an error within that panel was triggered for a given response. The error rates reported represent the percentage of the 50 responses for each student model flagged with an error in that panel.

Table 7 presents the comparative panel error rates generated by the three graders on the 50-sample validation subset. 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	Panel A	Data read-in errors	0	2	[2]
	Panel B	Bayes Rule application errors	76	76	[76]
	Panel C	Posterior calculation errors	96	100	[98]
	Panel D	Final decision inconsistency errors	6	10	[10]
GPT-4	Panel A	Data read-in errors	0	0	[0]
	Panel B	Bayes Rule application errors	4	4	[4]
	Panel C	Posterior calculation errors	64	74	[82]
	Panel D	Final decision inconsistency errors	0	0	[2]
GPT-4o	Panel A	Data read-in errors	0	0	[0]
	Panel B	Bayes Rule application errors	0	0	[0]
	Panel C	Posterior calculation errors	20	18	[22]
	Panel D	Final decision inconsistency errors	8	12	[14]

Note: 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 (Posterior Calculation), 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 (Final Decision Inconsistency), 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. Importantly, the core qualitative findings reported in Section 6 are robustly identified by all graders. This includes the transition from conceptual errors (Panel B) dominating in GPT-3.5 to calculation errors (Panel C) being more prevalent in GPT-4, followed by significant improvement in calculation accuracy (Panel C) for GPT-4o. This validation exercise confirms that GPT-o3-mini serves as a reliable primary grader for the large-scale textual analysis.