Strategic trade-offs between choice biases and accuracy for decisions based on asymmetric, rare evidence

Tahra L Eissa[1,*], Joshua I Gold[2,‡], Krešimir Josić[3,4,‡], and Zachary P Kilpatrick[1,5,‡]

[1]Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO, USA
[2]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Mathematics, University of Houston, Houston, TX, USA
[4]Department of Biology and Biochemistry, University of Houston, Houston, TX, USA
[5]Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA
[‡]equal contributions
[*]corresponding author: tahra.eissa@colorado.edu

Contributions: TLE, JIG, KJ, and ZPK all contributed to project development, experimental design, and writing the manuscript. JIG ran the human subject task. TLE performed all data and model analyses.

Abstract. Decisions based on rare events are challenging because the evidence provided by each event can be difficult to interpret. We examined an insidious form of this challenge, in which the evidence is rare and asymmetric (promoting biases), but the priors are symmetric (discouraging biases). Both a Bayesian model and most human subjects performing a simple decision task exhibited choice biases that reflected a tendency to rationally interpret a paucity of rare events as evidence for the alternative likely to produce fewer events, even when the two alternatives were equally likely. These biases varied across individual subjects, reflecting a distinction between: 1) imperfect Bayesian-based strategies that produced generally higher accuracy but when mis-tuned had degraded accuracy and excessive biases; and 2) heuristic strategies that tended to produce slightly lower accuracy and more consistent, near-normative biases. Our results show how humans balance biases and accuracy when making decisions based on rare events.

Pronounced decision biases can result from a wide variety of factors and have different effects on decision outcomes [1–3]. Some biases are the result of incomplete or improper knowledge of the structure of the world and tend to reduce accuracy compared to unbiased decision-makers [4]. Other biases account for asymmetric priors or rewards and thus improve performance [5]. Another source of decision bias is an asymmetry in the quality and availability of evidence associated with the alternatives [6]. Despite the general nature of this problem, how evidence asymmetries affect human decision biases and accuracy, and how these effects relate to normative theory, are not well understood.

Consider a doctor deciding whether or not to give a patient a drug based on the frequency of occurrence of a particular symptom (e.g., a seizure or bout of arrhythmia) relative to some baseline value (e.g., no symptoms). The absence of symptoms within a given time interval provides evidence that the drug is not needed. In contrast, the occurrence of the symptom in that same interval provides evidence that the drug is needed. A rational doctor will weight both types of evidence, collected over some time interval (e.g., days), to make a decision. However, if the symptoms are sufficiently rare and the decision is time sensitive, even a sick patient may exhibit no symptoms over the allotted observation time. As a result, even if half of the patients in a group are expected to need the drug *a priori*, the optimal decision based on the likely absence of symptoms will be biased towards not administering the drug. Such behavior represents a decision bias arising from the asymmetric and rare evidence available.

An additional challenge for rare-event decisions is that people often either underestimate [7–9] or overestimate [10, 11] the probability of occurrence of rare events. These kinds of misestimates have been framed in terms of a "description-experience gap". According to this theory, when the task structure and associated event probabilities are described explicitly, people tend to overestimate rare-event probabilities [12]. These overestimates are particularly prevalent with events representing negative outcomes [13], which is consistent with Prospect Theory [10]. In contrast, when the task structure is learned from experience, people tend to underestimate rare event probabilities [7, 14, 15]. These underestimates are thought to arise from sampling biases, and are prominent with small samples. Limited evidence is misleading, because the paucity of rare events may not be representative of the long-term objective probability [13, 16–18]. However, this delineation comes with caveats: subjects respond to feedback by adopting more objective probability estimation [19],

modifying estimates based on a regression to mean values [18, 20], displaying recency effects [21, 22], and exhibiting a diversity of other biases [23]. Yet, we lack an account on how these misestimations are integrated into a subject's specific decision strategy.

Our goal was to understand decision biases in rare-event problems in terms of their relationship to normative theory, their large individual variability, and the specific decision strategies that underlie them. We show that humans tend to exhibit choice biases during rare-event based decision-making that are similar to those produced by normative (Bayesian) models. Individual biases vary, but this variability can be attributed to differences in strategies and their effective use. More generally, we show that this individual variability yields insights into the relative advantages and disadvantages of two broad classes of decision strategies whose general roles in human decision-making are often considered and debated: 1) Bayesian-based strategies, which are capable of the highest accuracies but rely on estimates of task parameters that we show can be faulty, leading to systematically reduced accuracy and excessive biases for some subjects; and 2) heuristic strategies, which tended to yield slightly lower accuracies than the Bayesian-based strategies but no systematic mis-tuning that could lead to excessive biases.

Results

We used a task based on a classic inference problem that required each subject to infer which of two *a priori* equally likely jars filled with red and blue balls was the source of a sample of balls drawn with replacement (Fig. 1**a**). On each trial, the sample of 2, 5, or 10 balls was shown all at once, with the contents of each jar visible at all times, and subjects were told that each jar was equally likely to be the source (Fig. S1). Across different blocks, the proportions of red and blue balls in each jar were varied, thereby altering the ideal evidence weight of each observation. Under "symmetric" conditions, the ratios of the two ball colors in the two jars were reciprocal. In contrast, under "asymmetric" conditions, the ratios were non-reciprocal. Specifically, one ball color (either blue or red) accounted for fewer than half the balls in both jars and thus was "rare". This rare color was selected randomly for each subject but was then fixed throughout the experiment. We asked how ideal Bayesian observers and human observers compare in their use of symmetric and asymmetric information to infer which jar is the source of an observed sample.

Ideal Bayesian Observer

We first derived the strategy of an ideal Bayesian observer who knows the task structure. We defined the ratios of balls in the jars as h_+ and h_-, where h refers to the proportion of rare-colored balls in the given jar. The h_+ jar (termed the 'high' jar) included more balls of the rare color, whereas the h_- jar (termed the 'low' jar) included fewer balls of the rare color, so that $0 < h_- < h_+$. When the proportions were symmetric, $h_+ = 1 - h_-$. When the proportions were asymmetric, $0 < h_- < h_+ < 0.5$. Because the two jars were always visible, we assumed the fraction of rare balls, h_+ and h_-, in the low and high jars is known to the ideal observer.

The ideal observer sees a sample of draws, $\xi_{1:n}$, where $\xi_i = 1 (\xi_i = -1)$ if a rare (common) ball is drawn, and computes as evidence the log odds ratio (*belief*), $z_n = \log \frac{P(h_+|\xi_{1:n})}{P(h_-|\xi_{1:n})}$, between the probabilities that the sample of draws came from either jar. Balls were seen all at once, but the recursive belief-update equation shows the effect of each observation,

$$\begin{aligned} z_n &= \log \frac{P(\xi_n|h_+)}{P(\xi_n|h_-)} + z_{n-1} \\ &= \Psi(\xi_n) + z_{n-1}, \end{aligned}$$

where the belief increment after the nth observation is

$$\Psi(\xi_n) = \log \begin{cases} h_+/h_-, & \xi_n = +1 \\ (1-h_+)/(1-h_-), & \xi_n = -1. \end{cases}$$

The most likely choice given n observations is determined by the sign of z_n ($z_n > 0 \mapsto$ choose the high jar; $z_n < 0 \mapsto$ choose the low jar).

Asymmetric evidence leads to asymmetric belief updates, such that rare observations provide stronger evidence. On symmetric trials, the ratio of ball colors is reciprocal, and thus the magnitude of the belief increments is the same for all observations, $|\Psi(+1)| = \log \frac{h_+}{1-h_+} = -\log \frac{1-h_+}{h_+} = |\Psi(-1)|$. In contrast, on asymmetric trials ($h_- \neq 1 - h_+$), the magnitude of the belief updates are not equal. Thus, for rare asymmetric evidence, $0 < h_- < h_+ < 0.5$, observing a rare ball leads to a larger belief update than observing a common ball ($|\Psi(+1)| > |\Psi(-1)|$; Fig. 1**b**, green trace). By comparison, an asymmetric prior would shift the observer's initial belief (Fig. 1**b**, brown trace).

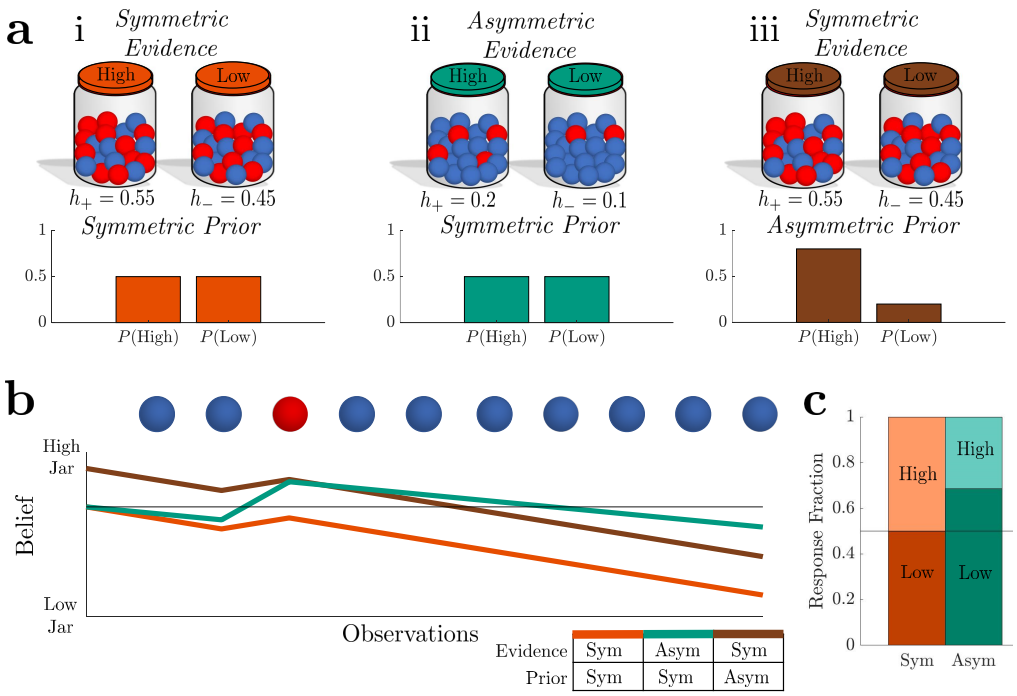

Figure 1: **Different sources of decision biases.** **a.** Schematic of the Jar-Discrimination Task. Balls were drawn with replacement from one of two equally probable jars with different ratios of red to blue balls. Here $h_\pm$ denotes the probability that a red ball is drawn from the "high" probability (h_+) and "low" probability (h_-) jar. We considered conditions with: (i) symmetric evidence ($h_- = 1-h_+$) and symmetric priors, or (ii) asymmetric evidence ($h_- \neq 1-h_+$) and symmetric priors. For comparison, we also show (iii) symmetric evidence and asymmetric priors ($P(High) \neq P(Low)$). **b.** Example belief state of a Bayesian observer as a function of observations (balls drawn) using the task structures shown in **a**. In the asymmetric-evidence environment, rare (here red) balls are weighed more heavily, causing larger increments in belief. With asymmetric priors, but symmetric evidence, belief updates are of equal magnitude, though the initial belief is biased. **c.** Fraction of low-jar to high-jar responses for an ideal observer across a block of simulated trials (that display 2,5, or 10 ball draws) in the environments with a symmetric prior and symmetric evidence (orange) versus asymmetric evidence (green) shown in **a**.

These differences in belief increments can lead an ideal observer to respond differently to the same sample when observed in symmetric-evidence or asymmetric-evidence environments. If the observer knows the two jars are equally likely to be the source of the observed sample *a priori*, we might expect the observer to choose both jars in equal proportions over the course of many trials. If the ball proportions are symmetric, the ideal observer's choices match the prior. However, if the ball proportions are asymmetric, the ideal observer chooses the jar containing fewer rare balls (fraction h_-, low jar) more often than its counterpart (Fig. 1**c**). The reason for this response bias (here, bias will be used to describe both artificial and human observers, regardless of the bias' benefits) is clearest when samples are small and often contain no rare balls at all. Thus, an ideal observer presented with a sample containing no rare balls will always choose the low jar based on the evidence, producing a response bias even though the prior is symmetric.

Human Responses are Similar to the Ideal Observer on Average

To determine whether human observers display the response biases exhibited by an ideal observer, we recruited 201 subjects using the crowdsourcing platform Amazon Mechanical Turk (MTurk) to perform a pre-registered Jar-Discrimination task (Fig. 1**a**i and ii). Each subject performed $T = 42$ trials ($T = 60$ in the control (CT) block) under each of four conditions that varied in difficulty and evidence asymmetry: hard asymmetric (HA), hard symmetric (HS), easy asymmetric (EA), and easy symmetric (ES) (Fig. 2**a**). The task structure and parameters were informed by simulations and pilot studies (see Methods).

In accord with the ideal observer, subjects tended, on average, to exhibit choice biases in the asymmetric, but not symmetric, conditions (Fig. 2**b**). In general, these biases tended to be largest when the number of balls observed was small (2 or 5 balls; Fig. 2**c**). The population trends of the human subjects were similar to those of the ideal observer under the same conditions. Specifically, within a given block, we identified all trials in which at least one rare ball was observed (rare-ball trials) and found an inverse relationship between the percentage of these rare-ball trials and the low-jar response frequency. Thus both human and ideal observers tended to show more/less of a preference for the low

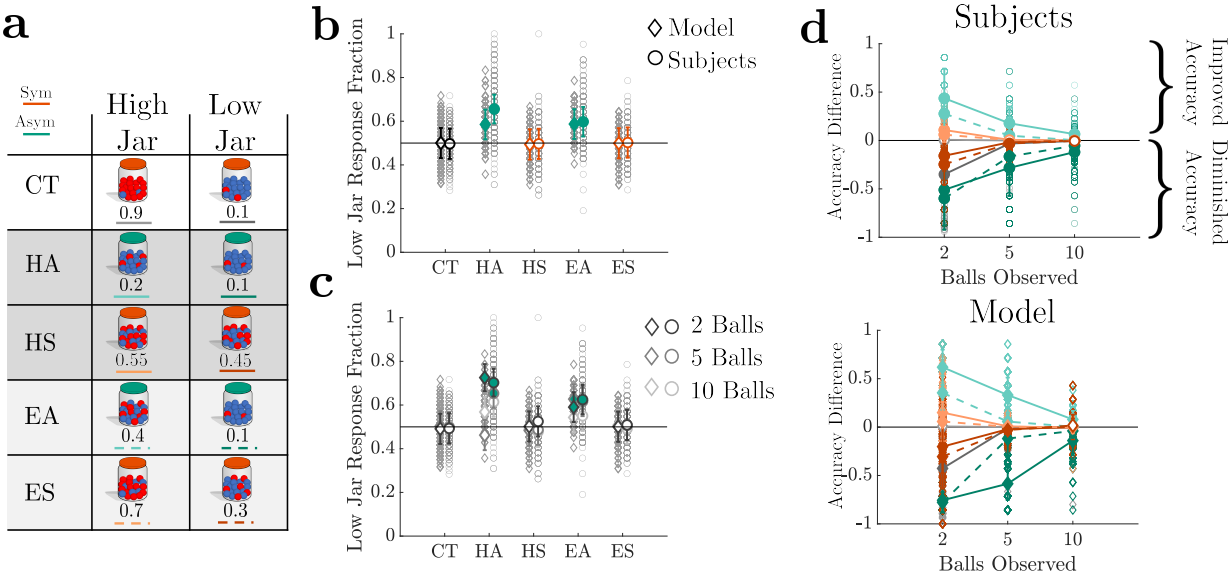

Figure 2: **Ideal observers and human participants show similar behaviors, on average.** **a.** Schematic showing the different variants of the task and parameters used with human subjects ($N = 201$): CT = Control; HA = Hard Asymmetric; HS = Hard Symmetric; EA = Easy Asymmetric; ES = Easy Symmetric. Colors and line type shown under each jar correspond with plots in **d**. **b.** Low jar-response fractions for each subject (grey circles) and sample-matched ideal observer responses (grey diamonds) for each block. Population bootstrapped means (1000 iterations) and 95% confidence intervals shown in corresponding symmetry color. Filled markers denote a significant population shift away from 0.5 ($p < 0.05$). **c.** Low-jar response fraction as in **b** for each trial length, separating trials in which 2, 5, or 10 balls were observed. Filled markers denote a significant population shift away from 0.5 ($p < 0.05$). **d.** Difference in performance on trials with at least one rare ball compared to performance from all trials for the given condition for subject (top) and ideal observer (bottom) responses, plotted as a function of the number of balls observed (trial length) for each condition. Differences in accuracy that corresponded with response asymmetries tended to be larger on asymmetric blocks with fewer observations in both human and ideal observers. Individual circles (diamonds) represent individual subjects (sample-matched models). Summary statistics as in **b**. Filled markers denote a significant population shift away from 0 ($p < 0.05$).

jar when rare-ball trials were less/more common (Fig. S2). These biases did not depend on the specific sequence of the sampled balls (Fig. S3). Moreover, accuracy (i.e., fraction correct) of both ideal observers and experimental subjects tended to be higher for the high jar (light curves in Fig. 2**d**) and lower for the low jar (dark curves in Fig. 2**d**) on rare-ball trials compared to all trials. These tendencies were strongest for short trials and asymmetric blocks (green versus orange curves in Fig. 2**d**), and the population summaries were qualitatively similar between the human subjects and sample-matched ideal observers.

When do Human Subjects' Responses Differ from the Ideal Observer?

Despite qualitative similarities in the performance of humans and ideal observers on the asymmetric tasks, they showed quantitative differences in the magnitude of response asymmetries and their dependence on trial length. These differences were highest in the hard-asymmetric task (Fig. 2**b**,**c**). To better understand these differences, we related them to non-normative serial biases that are common in human decision-making [24–26]. Specifically, we measured the frequencies of four possible two-trial response sequences [26]: 1) high-jar repetition; 2) low-jar repetition; 3) high-low jar alternation; and 4) low-high jar alternation. Because the correct jar was chosen independently on each trial, we could identify each subject's stereotypical response patterns by comparing the relative frequencies of two-trial response sequences during each block (Fig. 3**a**). This procedure allowed us to distinguish response biases that reflected a preference for a jar (i.e., low versus high) from preferences for a response sequence (i.e., alternations versus repetitions) for each block and subject. For example, a subject who strongly favors high-low alternations but does not favor low-high alternations shows a bias in favor of the low jar. We could then compare the subjects' serial biases to those of an ideal observer presented with the same samples of observations.

Many subjects exhibited consistent low-jar preferences in asymmetric blocks but no clear preferences in symmetric blocks (measured as distance from the origin along the negative diagonal in Fig. 3**b**). The low-jar preference on asymmetric trials persisted for some subjects even after subtracting the sequence frequencies of the sample-matched ideal observer, implying that their bias exceeded that predicted by the optimal observer (Fig. 3**c**). Some subjects showed an

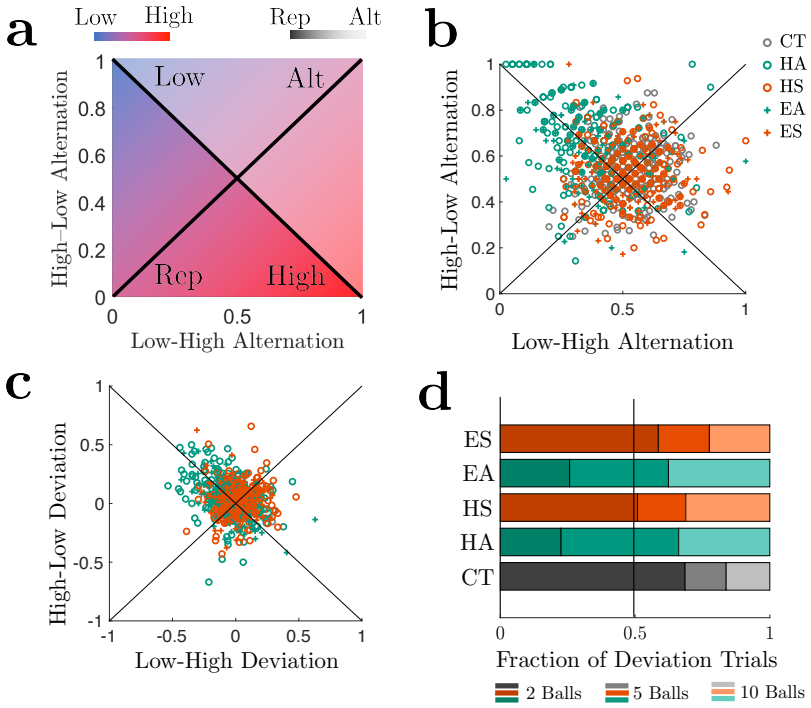

Figure 3: **Human subject deviations from the ideal observer. a.** Schematic depicting the preferences that can be inferred by comparing the response-alternation probabilities (low-high or high-low). The increasing diagonal corresponds to response sequence probabilities (repetition to alternation). The decreasing diagonal corresponds to jar preferences (low jar to high jar). The intersection point corresponds to unbiased responses. **b.** Human subjects' alternation plots. A majority of individuals displayed a low-jar preference (top-left quadrant: High-Low alternation > 0.5 and Low-High alternation < 0.5) in asymmetric blocks (70% of subjects in HA, 54% in EA) but not in symmetric blocks (20% in CT, 16% in HS, 19% in ES). Only asymmetric blocks showed multiple subjects with statistically significant low-jar bias ($p < 0.05$) as described by bootstrapping and multiple comparison correction (45% of subjects in HA, 21% in EA, 0% in CT, 0.5% in HS, and 0.5% in ES). Few subjects showed a repetition preference (bottom-left quadrant) in any block (10–15% of the subjects in a given block). **c.** Alternation statistics relative to the matched ideal observer. There was an excess low-jar preference for some subjects that was stronger and more common in asymmetric blocks (green points in the top-left quadrant; 48% in HA, 32% in EA compared to 17% in CT, 23% in HS, and 21% in ES). The percentage of subjects with a statistically significant enhanced low-jar preference ($p < 0.05$) as described by bootstrapping and multiple comparison correction was 23% in HA, and 8% in EA but was minimal in symmetric blocks (0% in CT, 2% in HS, 0.5% in ES) **d.** Fraction of trials where subjects' responses did not match the ideal observer (deviations) for each trial length. Symmetric blocks all show > 0.5 of deviations occur on two-ball trials.

additional repetition bias, but such biases were not common or consistent across blocks (Fig. 3).

To understand what factors drove the excessive, non-normative preference for the low jar for some subjects, we analyzed performance from all individual trials on which the subject gave a response that differed from the sample-matched ideal observer (deviation trials). For all blocks, the deviations occurred throughout a block (Fig. S4) and thus did not appear to be a learned compensatory mechanism that might occur near the end of the block to rebalance asymmetric choice distributions to match symmetric prior probabilities. Moreover, the deviations were not associated with particular response sequences (Fig. S5). For symmetric blocks, these deviation trials were most common for short (2 ball) trials, implying a straightforward sample size effect (Fig. 3**d**). However, trial length did not account for these deviations in asymmetric blocks, suggesting that deviations in this block are not simply the result of noise in the ideal observer's strategy but some alternative cause. We therefore examined if and how specific decision strategies gave rise to these deviations.

Formal Model Comparisons

To better understand the diversity of human behavioral patterns, we fit choice data from each subject and each block to several parameterized Bayesian-based models and several heuristic models (Fig. 4).

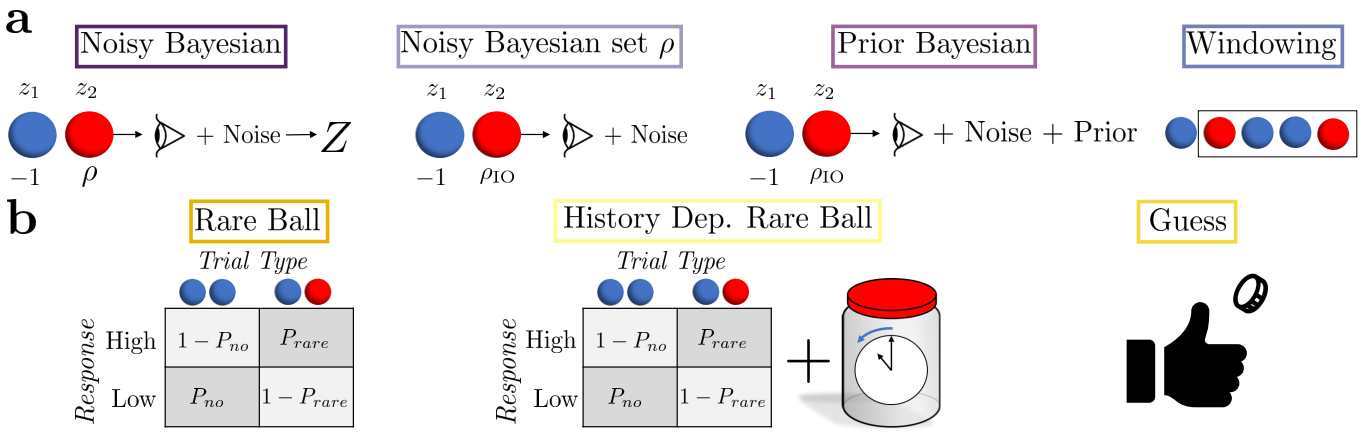

Figure 4: **Decision-making models. a.** Bayesian models. The Noisy Bayesian model includes noise and a variable rare ball weight ρ. The Noisy Bayesian set ρ model sets ρ to the ideal observer's rare-ball weight (ρ_{IO}). The Prior Bayesian model includes a jar bias (prior) but ρ matching the ideal observer. The Windowing model applies the Noisy Bayesian model to a selected window of data beginning at the rightmost ball in the sample (the final ball, assuming subjects view the sample moving from left to right). **b.** Heuristic models. The Rare-Ball model has a preference for either response based on the presence or lack of a rare balls. The History-Dependent Rare-Ball model extends the Rare Ball model to include the previous response. The Guess model guesses the high jar with some probability.

Our base Bayesian model, the Noisy Bayesian, was defined by computing a noisy version of the log likelihood,

$$z_n = \left\{ \begin{array}{ll} \rho, & \xi_n = +1 \\ -1, & \xi_n = -1 \end{array} \right\} + z_{n-1} + w_n,$$

where w_n is a normally distributed random variable with zero mean and variance a^2, and ρ is a free parameter representing the belief update in response to observing a rare ball, $\xi_n = 1$, ("Noisy Bayesian", Fig. 4**a**). When $\rho > 1$, the model weights a rare-ball observation more strongly than an observation of a common ball. Setting

$$\rho \equiv \rho_{\text{IO}} = \frac{\log h_+/h_-}{\log(1-h_+)/(1-h_-)}$$

recovers a rescaled noisy version of the ideal observer ("Noisy Bayesian set ρ").

We considered two other extensions of the Noisy Bayesian model (Fig. 4**a**), including: 1) a Noisy Bayesian model with a parameterized prior that allows for uneven likelihoods of the low and high jars, and the free parameter, $\rho \equiv \rho_{\text{IO}}$, set as with the noisy version of the ideal observer ("Prior Bayesian"); and 2) a windowing model ("Windowing") in which observations are used as in the Noisy Bayesian model, but the number of observations is a free parameter, fixed across trials. The number of observations could be smaller or larger than the number of balls in a single-trial sample and thus could include observations from previous trials. We assumed that the window began at the rightmost ball in the sample (the final ball, assuming subjects viewed the sample from left to right).

We also considered several heuristic strategies that are not based on likelihoods computed from samples (Fig. 4**b**). The first was a rare-ball model ("Rare Ball"), in which the probability of choosing the high jar is determined by the presence or absence of rare balls in the sample. An observation of one or more rare balls in a sample $(\max(\xi_{1:N}) = 1)$ leads the observer to choose the high jar with probability P_{rare} and the low jar with probability $1 - P_{\text{rare}}$. In contrast, on trials with no rare ball $(\max(\xi_{1:N}) = -1)$, the observer chooses the low jar with probability P_{no} and the high jar with probability $1 - P_{\text{no}}$. The second ("History Dependent Rare Ball") was an extension of the Rare-Ball strategy that accounts for the response on the previous trial (high or low jar) in the observer's probabilities of response. The third was a simple guessing strategy ("Guess") in which the observer selects the high jar with some probability, regardless of the observations (Fig. 4**b**).

We used Bayes factors to select the model that best matched each subject on each block (after confirming model identifiability; see Figs. S6, S9, and Methods). For all blocks, a majority of the subjects' behavior was most consistent with one of the Bayesian models (Figs. 5**a** and **b**), although individuals exhibited more varied approaches in asymmetric blocks. Subjects' responses did not typically match the same model across blocks, but approximately half (49%) of subjects were consistently well matched to models in the Bayesian class (Fig. S10).

The majority of subjects who had a strong preference for the low jar in asymmetric blocks were best described by Bayesian models, specifically the Noisy Bayesian or Prior Bayesian models (Fig. 5**c**). From the fits to these models, we identified two factors that could account for the subjects' excessive preferences (beyond that of the true ideal observer)

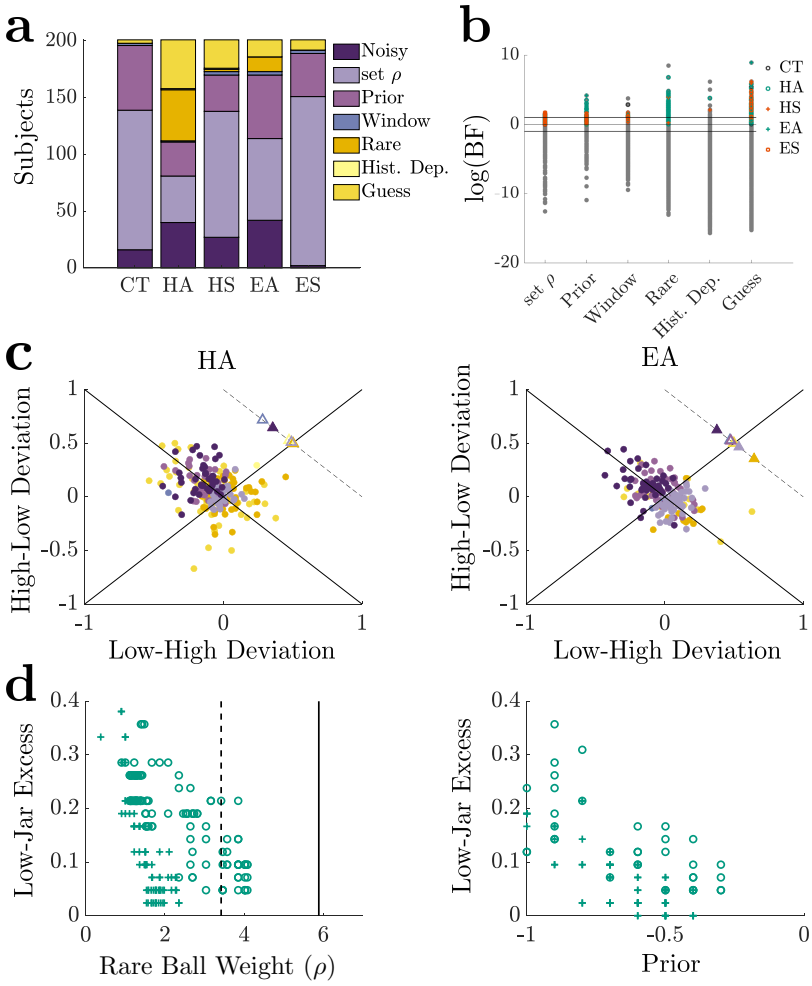

Figure 5: **Subjects used a variety of Bayesian-based and heuristic strategies. a.** Subjects categorized by the model that best matches their responses. Colors and labels as in Fig. 4. A majority of subjects in all blocks were matched to Bayesian models (98% in CT, 55% in HA, 86% in HS, 86% in EA, and 95% in ES), but with a relatively high percentage of heuristic strategies in the HA condition. **b.** Log Bayes factors for each subject-block, computed between each alternative model and the Noisy Bayesian model. log(BF)> 0 indicates the alternative model is more likely, with log(BF)> 1 or < -1 (black lines) considered to be strong evidence in favor of the preferred model [27]. Colored (grey) dots indicate the associated model is (is not) the most likely model. **c.** Asymmetric-block alternation plots as in Fig. 3**c**, color coded by the model that best matches a subject's responses. Medians for each model are shown on the dashed negative diagonal to indicate excess jar bias. Filled (open) triangles were significantly (not significantly) different from zero based on a Wilcoxon rank-sum test with $p < 0.05$. A majority of subjects with excess low-jar bias (top-left quadrant) were matched to the Noisy Bayesian or Prior Bayesian models (60% in HA, 89% in EA). **d.** Left: Fraction of excess low-jar responses versus the maximum likelihood estimate (MLE) of the rare ball weight, ρ, for subjects best matched to the Noisy Bayesian model in the asymmetric blocks (marker legend as in **b**). Black lines indicate the ideal observer's corresponding rare ball weight (HA: solid line, EA: dashed line). Right: Low-jar response excess versus the MLE of the response bias (Prior) for subjects best matched to the Prior Bayesian model in the asymmetric block (marker legend as in **b**). Negative values correspond with a bias in favor of the low jar.

for the low jar. First, subjects well matched to the Noisy Bayesian model tended to underweight the evidence from a rare-ball observation, ρ, compared to the ideal observer, with a stronger low-jar preference corresponding to a lower ρ weight (Fig. 5**d**; left). Second, subjects well matched to the Prior Bayesian model tended to *a priori* favor the low jar (Fig. 5**d**; Right).

The magnitude of these excessive low-jar biases (relative to the ideal observer) exhibited by subjects using Bayesian-based strategies also covaried with overall accuracy across subjects. Subjects who exhibited larger biases also had lower accuracy, compared to the bias and accuracy of the ideal observer presented with the same observations (Fig. 6). This relationship was not evident in any of the symmetric conditions, or for subjects who used heuristic strategies, and whose responses tended to be less accurate but also less biased compared to subjects using Bayesian-based strategies (Fig. 6).

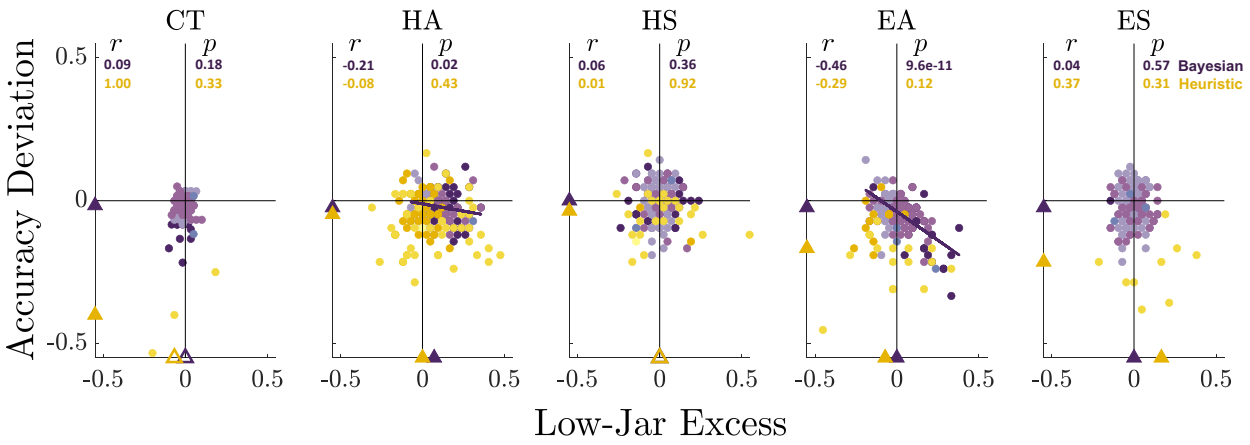

Figure 6: **Relationships between bias and accuracy for Bayesian-based, but not heuristic, strategies.** Accuracy deviation (negative values indicate worse accuracy) and low-jar excess (positive values indicate more low-jar responses) relative to the true ideal observer for all subject-blocks, color coded by each subject's best-fit model (colors as in Fig. 4). Median values for all subjects best-fit by Bayesian-based (purple triangles) or heuristic (yellow triangles) models are shown on each axis. Filled triangles indicate a significant difference between groups ($p < 0.05$) via Wilcoxon rank-sum test. Correlation coefficients (r) and p-values (p) are reported for each group and block. Regression lines are shown for group-blocks with significant correlations ($p < 0.05$).

Strategy-specific trade-offs between information use, accuracy, and biases

Our model-based analysis showed that the subjects' choices exhibited biases and accuracy limitations that were related to their particular decision strategies. In some cases, we could explain these trends in terms of fitted model parameters; e.g., some subjects underweighted evidence obtained from observing a rare ball, whereas others assumed unequal prior probabilities for the jars. To better understand these trends without relying strongly on specific models, we used a complementary, model-free analysis that relied on the complexity of the strategies each subject used for each task condition. As detailed below, strategy complexity helped to not only distinguish between heuristic and Bayesian-based strategies, but also how different versions of those strategy types led to different patterns of biases and accuracy limitations.

Specifically, we computed the mutual information (MI) between the observations and responses across trials from each participant in each block:

$$I_B^i(|\xi|, n; r) = \sum_{n\in\{2,5,10\}} \sum_{|\xi|\in\{0,\ldots,n\}} \sum_{r=h_\pm} P_B^i(|\xi|, n, r) \log_2 \frac{P_B^i(|\xi|, n, r)}{P_B^i(|\xi|, n) P_B^i(r)},$$

where i is a subject or model, B is the block, n is the size of a sample, $|\xi|$ is the number of rare balls in the sample, and r is the corresponding vector of responses in a block. Following principles of rate-distortion theory [28], we compared these values to the subject's accuracy (i.e., the fraction of correct responses within a block) and the maximum possible expected accuracy in a block with infinite trials (asymptotic bound) based on a fixed level of mutual information (MI) between the observations and the response (the accuracy "bound") (Fig. 7**a**). This approach allowed us to identify two important aspects of a subject's strategy: 1) its complexity, denoted by the amount of shared (mutual) information between responses and observations (ball sample in the current trial) I; and 2) how effectively that level of complexity was used to make decisions, represented by the proximity of a subject's accuracy to the upper bound. Importantly, a complex strategy does not guarantee high accuracy. However, decisions based on little information tend to be less accurate than decisions based on more information, when information is used effectively (i.e., the bound on the accuracy increases with MI in Fig. 7**a**).

In general, subjects who used more complex strategies (in terms of their use of information from the current trial) also tended to be more accurate for all five blocks, such that subjects using the most-complex strategies closely approached the optimal accuracy bound (Fig. 7**a**). This result is interesting because it implies that there was not an inherent limitation to how well relevant information could be used to maximize performance accuracy (which might occur, for example, if the only way to use that extra information effectively would be to adopt an impossibly convoluted strategy). In contrast, subjects whose responses were consistent with strategies of lower complexity tended to not only be less accurate and further from the bound, but also exhibit more excessive (relative to the ideal observer) low-jar biases, particularly for the asymmetric blocks. These trends of lower accuracy and higher biases for less-complex strategies in asymmetric blocks were apparent both within strategy types (Bayesian-based versus heuristic) and across strategy types, with generally lower complexity, lower accuracy, and lower biases for subjects best matched to heuristic versus Bayesian-based

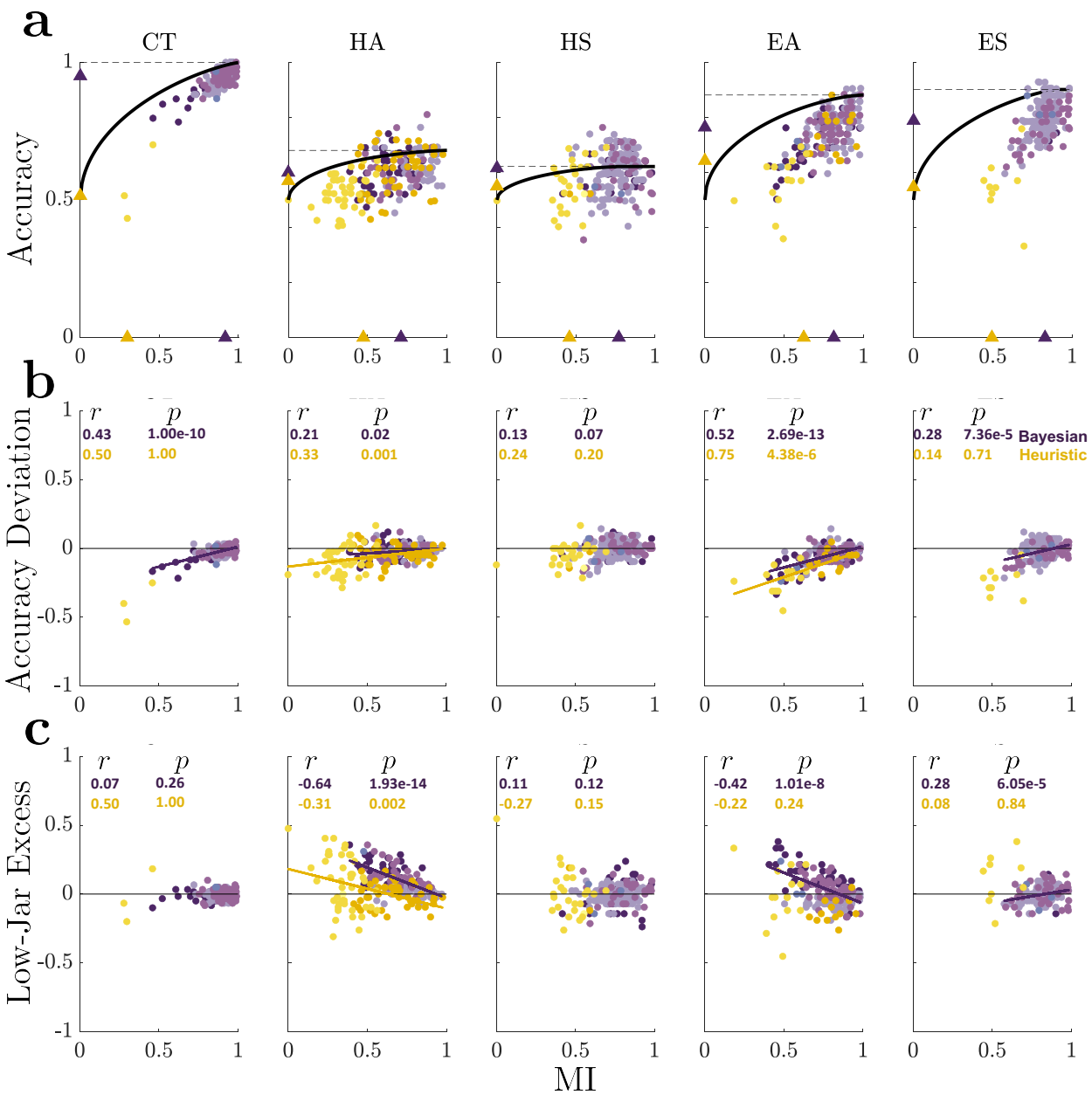

Figure 7: **Human strategies reflected systematic differences in the effective use of task information. a.** Accuracy versus mutual information (MI, computed as bootstrapped means from 1000 iterations per subject) for each block (columns, as labeled). Dots represent data from individual subjects, color coded by the model that best matched a subject's responses (colors as in Fig. 4). The bound is the maximum accuracy attainable in the limit of many trials by a normative observer under the constraint of a fixed amount of MI. Subjects with higher MI values were closer to the optimal bound. Note that points exceeding the bound resulted from the limited number of trials for each subject. The dashed horizontal lines indicate the bounds on accuracies for maximum MI values. Median values for Bayesian and heuristic groups are indicated on each axis. In each case, filled triangles denote statistically significant differences between classes ($p < 0.05$) based on a Wilcoxon rank sum test. **b.** Relationships between MI and accuracy deviation relative to the ideal observer. Correlation coefficients, r, and p-values for each group are shown in each panel. Regression lines are shown when $p < 0.05$. **c.** Relationships between MI and low-jar excess preference relative to the ideal observer, plotted as in **b**.

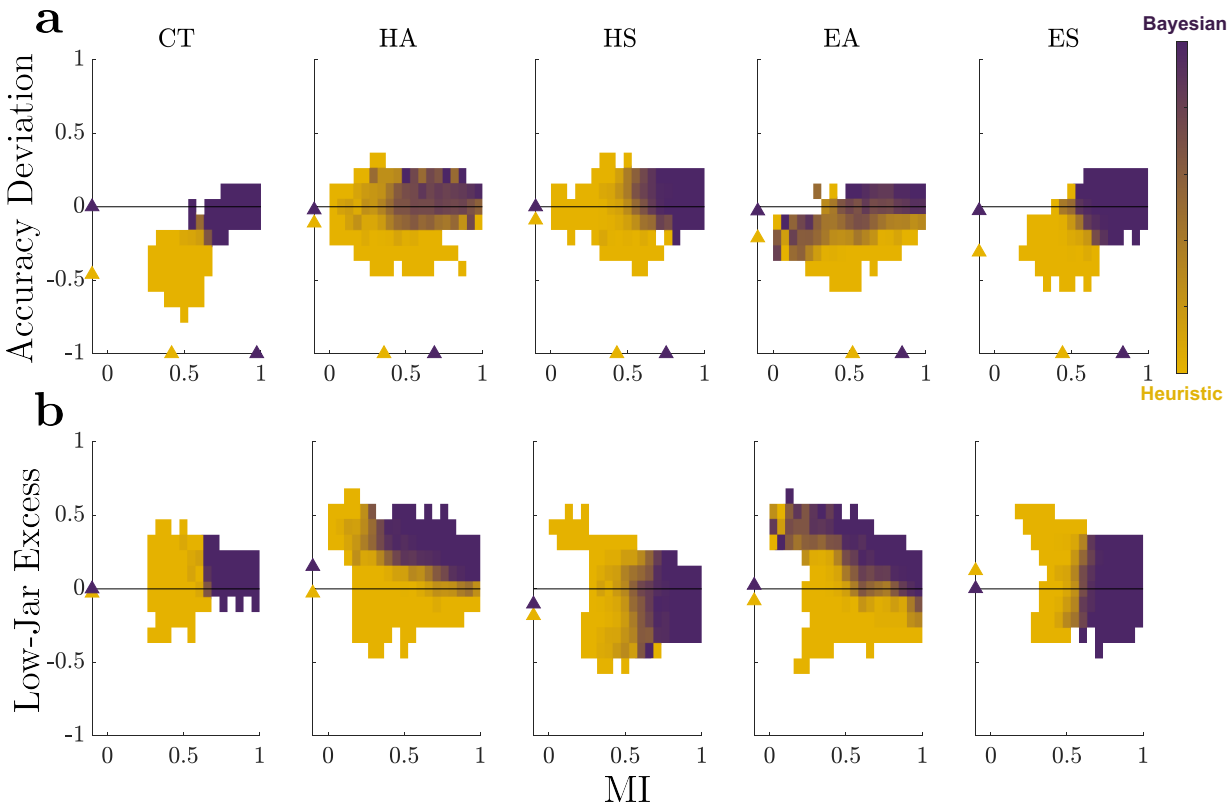

Figure 8: **Bayesian-based and heuristic models differ in their use of task information.** **a.** Maps identifying the regions of MI-accuracy deviation space that were most consistent with Bayesian (purple) and heuristic (yellow) model classes based on 1000 simulated subjects for each of the 7 models in Fig. 4. Color bar represents the fraction of Bayesian subjects at each point. In each case, filled triangles denote statistically significant differences between classes ($p < 0.05$) based on a Wilcoxon rank-sum test. **b.** Model maps of MI and low-jar excess. Colors as in **a**.

strategies (Fig. 7**b** and **c**). Together, these results suggest that: 1) lower accuracy and excessive biases can result from mis-tuning either strategy type, but particularly for the generally more-complex Bayesian-based strategies; and 2) there was an overall, across-subject strategic trade-off, whereby the use of heuristic strategies sacrificed accuracy for the benefit of greater simplicity and more limited biases.

These relationships between accuracy, strategy complexity, and low-jar biases were also predicted by model simulations and thus were not simply an unexplained behavioral idiosyncrasy. Specifically, we generated synthetic responses based on model fits from our subjects, assuming each subject used the decision-making models that best fit their actual responses. For each model, we randomly drew parameter values from those values that were within the 5% confidence intervals obtained from subject data fits. Using this procedure, we identified the ranges of MI, deviations from the sample-matched ideal observer accuracy, and deviations in low-jar biases that were most consistent with the responses generated by Bayesian-based and heuristic models (Fig. 8). In general, Bayesian-based models produced responses with more-complex strategies and higher accuracy compared to heuristic models for all task conditions (Fig. 8**a**). In addition, consistent with the subjects' behavioral patterns, low-jar biases in asymmetric blocks tended to exceed those of ideal observers (positive values) and were larger in magnitude for Bayesian models than for heuristic models across a range of strategic complexities (Fig. 8**b**).

Thus, individual differences in task performance could be characterized using data-driven metrics that, in turn, allowed us to better interpret the results of model fitting. Subjects who used the most information from a sample (highest MI) tended to match Bayesian-based strategies most similar to the ideal observer in terms of accuracy and low-jar biases. As strategy complexity decreased (lower MI), subjects using Bayesian-based strategies tended to have lower accuracy and stronger low-jar biases. In contrast, subjects using heuristic strategies generally showed lower MI, lower accuracy, and lower biases than those of subjects who used Bayesian strategies.

Discussion

How do people use rare, asymmetric evidence to guide decisions? We examined and compared normative and human decisions under these conditions, which have been used in a number of classical [29, 30] and recent [6] tasks, and report several new findings. First, even when the priors are symmetric, ideal observers exhibit choice asymmetries (which

for simplicity we have referred to as "biases" for both artificial and human observers, regardless of whether the choice asymmetries are beneficial or not) that favor the alternative producing fewer rare balls (events). These biases result from the scarcity of observations favoring one of the choices, when evidence is limited. Second, human subjects also showed these biases, albeit with strong individual differences in their magnitude. Third, these differences reflected two broad classes of decision strategies. The most prevalent "Bayesian-based" strategies were based on the computation of likelihoods used by true Bayesian observers, but with suboptimalities including mis-tuned weights associated with rare-ball observations and priors that deviated from their objective (symmetric) values. Subjects using these strategies had the best performance, on average, but also the largest biases. Other subjects used heuristic strategies that tended to give lower accuracies but also lower biases. Fourth, these strategy types also differed systematically in terms of complexity, as measured by MI between observations and decisions, such that Bayesian-based strategies tended to be more complex, more accurate, and more biased compared to heuristic strategies. Together, these findings yield new insights into sources and interpretations of suboptimalities in human decision-making, as discussed below.

In our task, a major factor governing both accuracy and choice biases was, unsurprisingly, the number of observations upon which a given decision was made. In general, decisions based on fewer observations tend to be less accurate [31,32] and more prone to biased inferences, such as for decisions based on illusory structure in random sequences [26, 33–35] and the related "hot-hand fallacy" [36]. However, when the evidence and priors are symmetric, the ideal observer is not expected show systematic biases towards a particular alternative, which is typically considered suboptimal (although such biases can arise from near-Bayesian decision strategies [34, 37]). In contrast, when evidence is limited and asymmetric, these kinds of biases can be expected even for an ideal observer. As we showed, people have a very strong tendency to exhibit these biases and did not try to suppress them given explicit instructions that the alternatives were equally likely.

Many of our subjects exhibited excessive biases beyond those expected by the ideal observer. These excessive biases came from several different sources, which for many subjects involved decision strategies that followed Bayesian decision principles. In some cases, these strategies appeared to be based on subjective priors with a preference for the low jar. These findings are distinct from previous work that examined choice biases in tasks with symmetric evidence but asymmetries in expected choice frequencies [5, 38–40], or reward outcomes [39, 41–44]. Under those conditions, biases based on asymmetric priors are common and, on average, tend to follow established, normative principles often formulated in the context of Signal Detection Theory [5] and/or sequential analysis [45]. In contrast, in our study subjects tended to either use inappropriate priors (e.g., subjects whose choices were best matched by the Prior Bayesian model with a prior biased towards the low jar) or neglect the symmetric prior altogether (e.g., subjects whose choices were best matched by heuristic models). These strategies could, in principle, reflect a relatively common form of recency bias that can cause an initial belief shift in the direction of the previous response [38, 39, 41, 42, 46, 47], and, more generally, is consistent with many previous findings of mis-tuned priors [1, 48–51].

Another major source of suboptimality for some of our subjects was the underweighting of evidence associated with the rare ball. This finding is related to description-experience gap theory, which distinguishes the tendency to overestimate the importance of rare events when their frequency is described and underestimate their importance when subjects learn their frequency through experience [7, 12, 16, 52]. In our task, event probabilities were both described and experienced across trials, which previously has been shown to promote better decisions [53]. Nevertheless, a substantial fraction of our subjects underweighted rare-ball evidence. This deviation from description-experience gap theory could reflect differences in our task design, which did not include the reward component typical of previous tasks used in studies that supported that theory [7, 14, 54]. This disagreement also likely reflects the general challenge of appropriately calibrating the quantitative value of probabilistic evidence associated with rare events.

The prevalence of mis-tuned priors and/or evidence used by subjects with Bayesian-based strategies also highlights certain costs inherent to using those strategies: although these strategies have the potential to optimize decisions and maximize accuracy, reward, and related factors, they are also computationally expensive [55] and difficult to tune appropriately [56]. These benefits and costs were evident in our data. Subjects who used Bayesian-based strategies tended to be more accurate, on average, but only the best performing of these subjects managed to achieve the highest levels of accuracy with minimal (normative) biases. The remainder of these subjects had degraded accuracy and excessive biases, as their use of relevant information from the current trial decreased. In contrast, subjects who used a heuristic strategy forfeited some accuracy but had generally lower biases that more closely matched those of the ideal observer, possibly with the additional benefit of reduced computational costs to implement those strategies (e.g., related to the lower task-related information they used in our information-bottleneck analyses) [2].

These results are consistent with resource-rational theories of decision-making that posit that people simultaneously seek to maximize reward while constraining strategy complexity [55, 57]. From this perspective, subjects who used Bayesian-based strategies sufficiently valued the potential benefits of higher accuracy (reward) to expend the additional computational costs associated with those strategies. In contrast, subjects who used heuristic strategies were content to sacrifice accuracy for simplicity. However, our information-theoretic measure was an imperfect way to assess this trade-off, for two primary reasons. The first is that this metric measured complexity only in terms of the use of information

from the current trial and not in terms of other (irrelevant) sources of information (e.g., from previous trials) that could have been used in a manner that increased the complexity of the decision process and also affected accuracy. The second is that this metric of information complexity does not necessarily correspond to computational complexity; i.e., the computational and memory-based resources needed to implement a particular strategy, which also likely contributes to trade-offs that can limit accuracy [58–61]. Moreover, we do not know the extent to which our subjects using low-complexity strategies had poorly tuned heuristics, were more distracted, employed fewer cognitive resources, and/or allocated less attention, as has been reported previously [62, 63]. Nevertheless, our findings add to a growing literature emphasizing that "optimality" in human decision-making is not a simple concept that can be related only to objective performance benchmarks (e.g., accuracy and biases) but also limitations on the capacity to achieve those benchmarks given the information available and the strategy used to process that information [64, 65].

Together these results provide new insights into the factors that govern the many forms of human decision-making that are neither truly optimal nor truly irrational. In this regime, optimal models still provide useful behavioral benchmarks and can be used to differentiate classes of suboptimal strategies used by different subjects and/or under different conditions. We showed that this approach can be used to understand strategies for solving rare-event decision problems. Future work could build on our findings by examining if and how they generalize to a wide variety of relevant real-world rare-event problems, including how people make decisions about financial investments that involve predictions of rare reward losses and gains [66–69], which of two areas is more prone to earthquakes [70], which of two states is more susceptible to a severe terrorist event [71], and how weather events relate to climate change [72].

Methods

Subject Protocols and Preregistration

We recruited 201 consenting subjects to perform the Jar-Discrimination Task on the Amazon Mechanical Turk crowdsourcing platform (95 female, 105 male, 1 non-disclosed). Subjects were compensated $4.50 for completing the task and were recruited only if they had a 95% or better approval rating and had performed at least 100 previous approved tasks. Location restrictions were limited to the United States. Human subject protocols were approved by the University of Pennsylvania Internal Review Board. The task was preregistered at osf.io prior to data acquisition (doi:10.17605/OSF.IO/J9XET). Analyses presented in Fig. 2, Fig. S2, Fig. S3, Fig. S10 and Fig. 5**a**,**b**, and the MLEs from Fig. 5**d** were performed exactly as listed in the preregistration.

Task Design

The goal of the task was to identify which of two jars was the source of a sample of balls shown to the observer. The jars were equally likely to be the source *a priori*. On each trial, subjects were shown a sample of 2, 5, or 10 red and/or blue balls randomly drawn with replacement and asked to determine which of the two jars displayed on the screen was the source of the sample (Fig. 1**a** and Fig. S1). Subjects were informed that the two jars were *a priori* equally likely to be the source. The ratios of colored balls in each jar were varied to create five blocks of trials and could be described by the proportion of balls of one color, termed the "rare-ball" color. The rare-ball color remained consistent throughout all blocks. Blocks were defined by the following rare-ball fractions for the high jar (containing more rare balls)/low jar (containing fewer rare balls): Control (0.9/0.1), Hard Asymmetric (HA; 0.2/0.1), Hard Symmetric (HS; 0.55/0.45), Easy Asymmetric (EA; 0.4/0.1), Easy Symmetric (ES; 0.7/0.3).

Before beginning the full task, subjects were shown a training slideshow and performed 24 trials using the control block. To continue to the full task, each subject was required to respond correctly on at least 80% of the control trials. Full sessions included randomized block orders for the remaining 4 test blocks interspersed with 12 control trials between test blocks. Each test block consisted of 42 trials, with randomly ordered but equally sampled values of: 1) the jar used for ball draws, and 2) sample length for each trial (2, 5, or 10 balls).

Prior to data acquisition, we used synthetic data generated by simulating the responses from the proposed models to confirm that models were identifiable given the task conditions and amount of data to be collected (Fig. S6). We determined the number of trials in a block by balancing: 1) model parameter identifiability, with 2) reasonable task-time length for human subjects (i.e., about 30 min per session). The jar ratios were selected based on generated synthetic responses of the ideal observer, such that overall accuracy was matched between the asymmetric and symmetric blocks at each difficulty (i.e., the hard asymmetric and hard symmetric tasks were matched in accuracy). Models were developed and fit to pilot data to ensure model and parameter identifiability (See Model Fitting and Comparison below).

Models

Bayesian Models

One class of models we considered depended on the probabilities of ball samples coming from the high or low jar that would be computed by a Bayesian observer.

Ideal Observer. In the simplest case without noise, an ideal Bayesian observer makes a decision based on a sample of n balls drawn from one of the jars, $\xi_{1:n}$, where $\xi_i = 1$ and $\xi_i = -1$ denote an observation of a rare and common ball color, respectively. The ideal observer uses these observations to update the log likelihood ratio (*belief*), $z_n = \log \frac{P(h=h_+|\xi_{1:n})}{P(h=h_-|\xi_{1:n})}$, between the probabilities that the sample came from a jar with a rare ball frequency of $h = h_+$ (high) or $h = h_-$ (low). We can write the update iteratively:

$$\begin{aligned} z_n &= \log \frac{P(\xi_n|h=h_+)}{P(\xi_n|h=h_-)} + z_{n-1} \\ &= \left\{ \begin{array}{cc} \log \frac{h_+}{h_-}, & \xi_n = +1 \\ \log \frac{1-h_+}{1-h_-}, & \xi_n = -1 \end{array} \right\} + z_{n-1} \equiv \Psi(\xi_n) + z_{n-1}. \end{aligned}$$

The most likely choice based on n ball draws is given by the sign of z_n ($z_n > 0 \mapsto$ choose the high jar; $z_n < 0 \mapsto$ choose the low jar). In all blocks, the probability that either jar was the source of the sample was 0.5, so that the ideal observer model had a flat prior, and $z_0 = \log \frac{P(h=h_+)}{P(h=h_-)} = 0$.

In symmetric environments, $h_+ = 1 - h_-$, so

$$\Psi(+1) = \log \frac{h_+}{1-h_+} = -\frac{1-h_+}{h_+} = -\Psi(-1),$$

and thus the magnitude of the belief increment is the same for either observation ($|\Psi(+1)| = |\Psi(-1)|$). When the environment is asymmetric, $h_- < 1 - h_+$, and different ball colors correspond to different evidence weights ($|\Psi(+1)| \neq |\Psi(-1)|$).

For n ball draws, we can compute the probability of the responses (choices) on a given trial, $r = h_-$ and $r = h_+$ for the low and high fraction jars as

$$\begin{aligned} P(r=h_+) &= P(r=h_+|h=h_+)P(h=h_+) + P(r=h_+|h=h_-)P(h=h_-) \\ P(r=h_-) &= P(r=h_-|h=h_+)P(h=h_+) + P(r=h_-|h=h_-)P(h=h_-) \end{aligned}$$

using binomial distributions. For example,

$$P(r=h_+) = 0.5 \left[\sum_{j=1}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) h_+^j h_-^{n-j} \right].$$

Conditioning on trial type, we can extend this analysis to obtain the minimum number of rare balls, B, that must be observed to produce a high jar response, given a sample of size n. This number is dependent on h_+ and h_-. When the jars are symmetric ($h_+ = 1 - h_-$), $B = n/2$. In asymmetric cases, $B < n/2$ if $h_+ + h_- < 1$. Thus,

$$\begin{aligned} P(r=h_+|h=h_+) &= \sum_{k>B}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) h_+^k (1-h_+)^{n-k} \\ P(r=h_+|h=h_-) &= \sum_{k>B}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) h_-^k (1-h_-)^{n-k} \end{aligned}$$

and

$$\begin{aligned} P(r=h_-|h=h_+) &= \sum_{k=1}^{\lfloor B \rfloor} \left(\begin{array}{c} n \\ k \end{array} \right) h_+^k (1-h_+)^{n-k} \\ P(r=h_-|h=h_-) &= \sum_{k=1}^{\lfloor B \rfloor} \left(\begin{array}{c} n \\ k \end{array} \right) h_-^k (1-h_-)^{n-k}. \end{aligned}$$

Noisy Bayesian. We extended the ideal observer model to include noisy belief updates, with means and variances of arbitrary magnitude. To do so we let $w_n \sim \mathcal{N}(0, a^2)$ be a normally distributed random variable with zero mean and variance a^2 and defined the weight updates by

$$z_n = \left\{ \begin{array}{cc} \rho, & \xi_n = +1 \\ -1, & \xi_n = -1 \end{array} \right\} + z_{n-1} + w_n,$$

where ρ is a free parameter representing the belief update in response to observing a a rare ball, $\xi_n = 1$. Because the sign of z_n is all that matters for determining a model observer's response, we normalized the update in response to a common ball to remove an unnecessary parameter.

Noisy Bayesian Model with set ρ. In this model, the belief updates in a Noisy Bayesian model with

$$\rho \equiv \rho_{\text{IO}} = \frac{\log h_+/h_-}{\log(1-h_+)/(1-h_-)},$$

are on average equal to those in a rescaled ideal Bayesian model. However, each belief update is perturbed additively by a Gaussian random variable with variance, a^2. If ρ is set to the optimal value ρ_{IO}, this variance is the only free parameter.

Prior Bayesian Model. We found that the Noisy Bayesian model that included a bias in the prior probability of the two choices (3 free parameters) was not well identified from the Noisy Bayesian model (see Model Fitting and Comparison below and Fig. S9). Thus, instead we modified the Noisy Bayesian with fixed ρ to include a free parameter z_0 for the prior. An observer using this model uses potentially unequal prior probabilities,

$$z_0 \equiv \log \frac{\tilde{P}(h = h_+)}{\tilde{P}(h = h_-)} \neq 0,$$

where $\tilde{P}(h)$ represents the observer's assumed prior probability, which may differ from the the true prior probability that a jar with rare ball fraction h is a source of the sample. A positive (negative) value of z_0 implies that the observer believes *a priori* that the high (low) jar is more likely to be the source of a sample.

Windowing Model. We extended the Noisy Bayesian model to include a free parameter, m, that determines the number of ball observations used to make a decision on any given trial. If an observer uses fewer observations than n the number of balls in a sample ($m < n$), then the last m of the sample beginning from the final/ rightmost displayed ball determine the belief update on a given trial. If the samples on a trial contain fewer than m balls ($m > n$), then the rightmost $m - n$ balls from the previous trial are used to update the belief, where n is the number of balls in the current trial's sample.

Heuristic Models

The other class of models (heuristic) did not depend on the likelihood functions associated with drawing a ball of a certain color from either jar.

Rare-Ball Model. The probability of choosing either jar in the Rare Ball Model depends only on whether one or more rare balls are observed in a sample in the current trial (N),

$$P_{\text{response}} = \left\{ \begin{array}{c} P(r_N = h_+|\text{rare}^N) = P_{\text{rare}} \\ P(r_N = h_-|\text{rare}^N) = 1 - P_{\text{rare}} \\ P(r_N = h_-|\text{no rare}^N) = P_{\text{no}} \\ P(r_N = h_+|\text{no rare}^N) = 1 - P_{\text{no}} \end{array} \right\}.$$

Here r_N is the response on the current trial N, $(\text{rare}^N) \equiv (\max(\xi_{1:n}^N) = 1)$ corresponds to observing one or more rare balls, and $(\text{no rare}^N) \equiv (\max(\xi_{1:n}^N) = -1)$ to observing no rare balls in the current trial.

History-Dependent Rare-Ball Model. In this model, the probability of a choice depends on observing a rare ball in the sample, *and* the choice r_{N-1} on the previous trial. Thus, the model has four free parameters:

$$P_{response|high} = \left\{ \begin{array}{c} P(r_N = h_+|\text{rare}^N, r_{N-1} = h_+) = P_{\text{rare—high}} \\ P(r_N = h_-|\text{rare}^N, r_{N-1} = h_+) = 1 - P_{\text{rare—high}} \\ P(r_N = h_-|\text{no rare}^N, r_{N-1} = h_+) = P_{\text{no—high}} \\ P(r_N = h_+|\text{no rare}^N, r_{N-1} = h_+) = 1 - P_{\text{no—high}} \end{array} \right\},$$

$$P_{response|low} = \left\{ \begin{array}{c} P(r_N = h_+|\text{rare}^N, r_{N-1} = h_-) = P_{\text{rare—low}} \\ P(r_N = h_-|\text{rare}^N, r_{N-1} = h_-) = 1 - P_{\text{rare—low}} \\ P(r_N = h_-|\text{no rare}^N, r_{N-1} = h_-) = P_{\text{no—low}} \\ P(r_N = h_+|\text{no rare}^N, r_{N-1} = h_-) = 1 - P_{\text{no—low}} \end{array} \right\}.$$

Guess Model. In this model, the probability of each choice is fixed, and independent of the sample. The Guess Model includes one free parameter that determines the probability of choosing the high jar:

$$P(r_N = h_+) = P_{\text{guess}}, \qquad P(r_N = h_-) = 1 - P_{\text{guess}},$$

regardless of any observations within a trial.

Model Fitting and Comparison

We fit model parameters to data using Bayesian maximum-likelihood estimation. We obtained the posteriors over the parameters by considering the vectors of responses, $r_{1:42}$, and observation samples, $\boldsymbol{\xi}_{1:42}$, across all 42 trials in a block ($\boldsymbol{\xi}_{1:60}$ for the control block that had 60 trials). For instance, to infer the noise variance, a^2, and rare ball weight, ρ, in the Noisy Bayesian Model, we applied Bayes' rule and then computed the probability of a response r_N in a given trial conditioned on observations $\boldsymbol{\xi}_N$ as

$$p(a, \rho|r_N, \boldsymbol{\xi}_N) = \frac{p(r_N|a, \rho, \boldsymbol{\xi}_N)p(a, \rho)}{p(r_N|\boldsymbol{\xi}_N)}.$$

Because the denominator provides only a normalization of the probability densities of a and ρ, the primary contributions are the probability of a response r_N given the parameters and observations, and the prior over the parameters, $p(a, \rho)$. We explain the choice of priors below. Because all models were defined in terms of either simple binary random variables or thresholded Gaussians, we could evaluate the associated likelihood functions analytically. For instance, in the case of the Noisy Bayesian model, for a trial with 5 balls, and a sample containing 4 common and 1 rare ball ($\boldsymbol{\xi}_N = (-1, -1, +1, -1, -1)$), the probability of choosing the high jar, $r_N = H$, is

$$P(r_N = h_+|a, \rho, \boldsymbol{\xi}_N) = \frac{1}{\sqrt{2\pi a^2}} \int_0^\infty \exp\left[-\frac{(z + 4 - \rho)^2}{2a^2}\right] dz = \frac{1}{2}\left[1 - \text{erf}\left(\frac{4 - \rho}{\sqrt{2a^2}}\right)\right].$$

For models in which responses are independent across trials, we used the trial-wise response probabilities to compute the posteriors given responses and samples in a block of trials,

$$p(a, \rho|r_{1:42}, \boldsymbol{\xi}_{1:42}) = \frac{p(a, \rho)}{p(r_{1:42}|\boldsymbol{\xi}_{1:42})} \prod_{j=1}^{42} p(r_j|a, \rho, \boldsymbol{\xi}_j).$$

The maximum of this posterior is the maximum likelihood estimate of the model parameters. The interval of parameters containing at least 95% of the maximum likelihood estimate were included as credible intervals for the model fits.

To design the human task and determine whether the models are identifiable from the given data, we performed model comparisons on synthetic data. We first used the Noisy Bayesian model to determine the minimum number of trials (42) needed to fit synthetic data and produce a task with a reasonable task duration for online data acquisition (30 min or less) (Fig. S6). However, for this model, parameters produced with a flat prior were not always identifiable, given the amount of data that we could reasonably expect to collect in a block. This problem resulted from dependencies between the noise variance, a^2, and rare ball weight, ρ, parameters (Fig. S6 HA). To account for this effect, we used pilot data from 20 subjects to create an informative prior based on the subjects' posteriors. The informative priors were computed as a smoothed version of averaged posteriors produced by the pilot subject's fits to the Noisy Bayesian model. The averaged posterior was smoothed with respect to each parameter (Fig. S7). To weaken the posterior with respect to ρ, the averaged marginal posterior was filtered using a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where the mean μ was set at the maximum value of the averaged marginal posterior and the variance σ^2 was set such that the median mean squared error (MSE) of the parameter fits ρ for 100 synthetic Noisy Bayesian datasets was below one. The averaged marginal posterior with respect to the noise parameter a was smoothed using the function $(x+c)/(1+cL)$ where x is the marginal posterior and c and L are scaling constants selected such that the averaged posterior was smooth (no jagged edges) but did not

impact the accuracy of the rare ball parameter fitting (values provided in Fig. S6). Given that a low-noise parameter a was identified for most pilot subjects and that higher values of a could correspond with underweighting the value of ρ, we prioritized accurately identifying ρ.

To confirm that the new informative priors produced realistic fits for all of our Bayesian models, we applied the informative priors to model fits for data of 100 synthetically generated subjects with randomly selected model parameters for each model per block and found that the credible intervals usually contained the true parameters (median MSE ρ range 0.01–0.82, median MSE a range 0.006–0.69) (Fig. S8). The fits of the synthetic subjects also matched the averaged posteriors from the pilot data, with a strong preference for low-noise parameter values and values at or below the true rare-ball weight. Thus, using informative priors did limit identifiability at high noise variance and rare-ball weight values, but we ensured that our models could be correctly identified near or below the values predicted by the ideal observer, as was suggested by the pilot data.

We then generated model responses from 100 randomly sampled versions for each of the candidate models (sampling from the informative priors) to confirm that each model could be appropriately selected when compared to other models. We performed model comparison and selection using log Bayes factors, comparing the likelihood that a particular dataset came from one of two models by computing the log likelihood ratio of the marginal likelihoods for any given pair of models,

$$\log \mathrm{BF} = \log \frac{P(D|M_2)}{P(D|M_1)} = \log \frac{P(M_2|D)P(M_1)}{P(M_1|D)P(M_2)}.$$

Here D is the data from a block of trials ($r_{1:42}$ and $\boldsymbol{\xi}_{1:42}$), and M_1 and M_2 are two models from the list we described above. For example, to compare the Noisy Bayesian Model to the Prior Bayesian Model for a given block, we integrate the probability of responses conditioned on observations and parameters against the priors over the model parameters:

$$\log \mathrm{BF} = \log \frac{\int_{-\infty}^{\infty}\int_{0}^{\infty} p(r_{1:42}|a, z_0, \boldsymbol{\xi}_{1:42})p(a, z_0)dadz_0}{\int_{0}^{\infty}\int_{0}^{\infty} p(r_{1:42}|a, \rho, \boldsymbol{\xi}_{1:42})p(a, \rho)dad\rho}.$$

For all comparisons, we used the Noisy Bayesian model as the baseline model (in the denominator of the Bayes factor); i.e., model 1 (M_1). We found that two candidate models were not identifiable (one assumed an asymmetric repetition bias, the other included a biased prior and free parameter for rare ball weight; Bayes factors correctly selected the true model $< 80\%$ of the time) and thus were excluded from our analyses.

To determine the model that was most consistent with a human subject's responses on a particular block, we computed the $\log$ Bayes factors between each alternative model and the Noisy Bayesian Model. Positive values of the $\log$ Bayes factor provided evidence in favor of a particular alternative model over the noisy Bayesian model, with evidence growing with the magnitude of the factor (we chose $|\log \mathrm{BF}| > 1$ to indicate sufficiently strong evidence in favor of a model [27]). The most-likely model was selected based on the maximal log Bayes factor value across all alternative models. If no values were >0, the Noisy Bayesian model was selected.

Rate-Distortion Theory

We applied rate-distortion theory to compare the subjects' accuracy (fraction correct) to the maximal accuracy bound obtainable by an ideal observer constrained to a fixed amount of mutual information (MI) between an observer's response, r and the observation on a trial. We describe this observation as a random variable $(|\xi|, n)$, where n is the size of a sample, and $|\xi|$ is the number of rare balls in the sample, as:

$$I_B^i(|\xi|, n; r) = \sum_{n\in\{2,5,10\}} \sum_{|\xi|\in\{0,\ldots,n\}} \sum_{r=h_\pm} P_B^i(|\xi|, n, r) \log_2 \frac{P_B^i(|\xi|, n, r)}{P_B^i(|\xi|, n)P_B^i(r)},$$

where i is a subject or model, B is the block. We computed the probabilities P_B^i empirically: Subject estimates were computed from all response and observation data for the 42 trials within a block, so any particular observation sample not seen was not included in the sum. Each subject's trials within a block were bootstrapped by uniformly resampling the data 1000 times to obtain a distribution of MI and accuracy estimates for the block.

To define the accuracy bound for a normative observer, we computed MI in the limit of many samples (not just 42 per block), allowing for a calculation directly using probability mass functions. As such, we considered all possible samples $\boldsymbol{\xi}_B^i$, disregarding ball order, ($n+1$ possible counts for trials with $n = 2, 5, 10$ ball draws) in Ξ and responses r_B^i in R :

$$I_\pi(\boldsymbol{\xi}_B^i, r_B^i) = \sum_{\boldsymbol{\xi}_B^i\in\Xi} P(\boldsymbol{\xi}_B^i) \sum_{r_B^i\in R} \pi(r_B^i|\boldsymbol{\xi}_B^i) \log_2 \frac{\pi(r_B^i|\boldsymbol{\xi}_B^i)}{p(r_B^i)}$$

where $\pi(r_B^i|\boldsymbol{\xi}_B^i)$ is the policy used to generate responses from observations across the block. Note, that this is simply given by the standard ideal observer model defined above when fixing the MI to unity. However, for values of MI less than one, we employed an optimization procedure, which we describe below, in order to obtain the optimal policy that uses a fixed MI budget.

Computing the Optimal Bound. The rate-distortion bound can be computed according to a constrained-optimization problem in which we identify the maximum possible accuracy for a given level of MI in the limit of many trials. In the ideal observer case, the policy applied to compute MI and accuracy is:

$$\pi(r = h_+|\xi^c) = \begin{cases} 1, & \xi^c > B \\ 0.5, & \xi^c = B, \\ 0, & \xi^c < B. \end{cases}$$

(and $\pi(r = h_-|\xi^c) = 1 - \pi(r = h_+|\xi^c)$) where $\xi^c \in \{0, 1, 2, ..., n\}$ is the count of rare balls observed and B is the number required to trigger a high jar response. Note this provides a specific accuracy bound for a fixed value of MI, corresponding to the ideal observer. Additionally, we must compute the predictive accuracy using the value function applied to a particular policy π

$$V_\pi = \sum_{\xi^c} p(\xi^c) \sum_{r \in \{h_+, h_-\}} \pi(r|\xi^c) Q(\xi^c, r),$$

which sums over all possible combinations of unordered sample counts ($\xi^c = 0, 1, 2, ..., n$ rare balls for $n = 2, 5, 10$ balls in a trial) for which we can always compute the trial specific value function from the ideal observer $Q(\xi^c, r) = 1/(1 + \mathrm{e}^{-z_n})$, where z_n is the ideal observer's log likelihood ratio.

Thus, to bound accuracy for a given MI ($I_\pi \equiv C$), we maximized the value function according to the best policy that uses the prescribed MI:

$$V^* = \max_\pi V_\pi \quad \text{s.t.} \quad I_\pi \equiv C,$$

which generates the optimal predicted bounds. This maximization problem was solved using MATLAB's constrained optimization package (`fmincon`) with a constraint given by $I_\pi \equiv C$ and V_π as the objective function.

Model Mapping. To identify relevant regions of MI-accuracy space for each model class, 1000 synthetic sets of responses were generated for each model in each class using parameters fit in the credible interval for subjects matching that model. Computing the MI and accuracy values for each synthetic observer, the density of each model class was binned in MI-accuracy space. The fraction of Bayesian observers in each bin ($\#_{\text{Bayesian}}/(\#_{\text{Bayesian}} + \#_{\text{heuristic}})$) was computed to identify regions that were most consistent with each class.

Statistics

Population statistics were computed by uniformly bootstrapping 1000 times from each data set, using the same number of samples as the original sample, to identify the mean and confidence intervals. Subject p-value statistics were computed using bootstrapping and then applying the Benjamini-Hochberg correction for multiple comparisons [73]. Correlations were computed using Spearman's correlation, with correlation coefficient r and p-value p reported in figure captions. Differences between medians were computed using a two-sided Wilcoxon rank-sum test. We defined significance as $p < 0.05$.

Data Availability

The human data set used in Figures 2, 3, 5, 6, and 7 will be made available on github.com/teissa/RareEvents following acceptance of the manuscript for publication.

Code Availability

The code developed for all proposed models has been made available on github.com/teissa/RareEvents.

References

[1] Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124–1131 (1974).

[2] Gigerenzer, G. & Brighton, H. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science* **1**, 107–143 (2009).

[3] Hilbert, M. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin* **138**, 211 (2012).

[4] Yeshurun, Y., Carrasco, M. & Maloney, L. T. Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Res.* **48**, 1837–51 (2008).

[5] Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics*, vol. 1 (Wiley New York, 1966).

[6] Odoemene, O., Pisupati, S., Nguyen, H. & Churchland, A. K. Visual evidence accumulation guides decision-making in unrestrained mice. *Journal of Neuroscience* **38**, 10143–10155 (2018).

[7] Hertwig, R., Barron, G., Weber, E. U. & Erev, I. Decisions from experience and the effect of rare events in risky choice. *Psychological Science* **15** (2004).

[8] Barron, G. & Yechiam, E. The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgement and Decision Making* **4**, 447–460 (2009).

[9] Camilleri, A. R. & Newell, B. R. When and why rare events are underweighted: a direct comparison of the sampling, partial feedback, full feedback and description choice paradigm. *Psychon Bull Rev* (2011).

[10] Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–292 (1979).

[11] Hertwig, R. & Erev, I. The description-experience gap in risky choice. *Trends in Cog. Sci.* **13**, 517–523 (2009).

[12] Wulff, D. U., Mergenthaler-Canseco, M. & Hertwig, R. A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin* **144**, 140–176 (2018).

[13] Fox, C. R. & Hadar, L. "Decisions from experience"= sampling error+ prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making* **1**, 159 (2006).

[14] Barron, G. & Erev, I. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making* **16** (2003).

[15] Lieder, F., Griffiths, T. L. & Hsu, M. Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review* **125**, 1 (2018).

[16] Hau, R., Pleskac, T., Kiefer, J. & Hertwig, R. The description-experience gap in risky choice: the role of sample size and experienced probabilities. *J. Behav. Decis. Making* **21** (2008).

[17] Rakow, T., Demes, K. A. & Newell, B. R. Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes* **106**, 168–179 (2008).

[18] Ungemach, C., Chater, N. & Stewart, N. Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science* **20** (2009).

[19] Jessup, R. K., Bishara, A. J. & Busemeyer, J. R. Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science* **19**, 1015–1022 (2008).

[20] Glöckner, A., Hilbig, B., Henninger, F. & Fiedler, S. The reversed description-experience gap: Disentangling sourced of presentation format effects in risky choice. *J. Exp. Psych.* **145**, 486–508 (2016).

[21] Plonsky, O., Teodorescu, K. & Erev, I. Reliance on small samples, the wavy recency effect, and simlarity-based learning. *Psychological Review* **122**, 621–647 (2015).

[22] Szollosi, A., Liang, G., Konstantinidis, E., Donkin, C. & Newell, B. R. Simultaneous underweighting and overestimation of rare events: unpacking a paradox. *J. Exp. Psych* **148**, 2207–2217 (2019).

[23] Kellen, D., PAchur, T. & Hertwig, R. How (in)variant are subjective representations of descrived and experienced risk and rewards? *Cognition* **157**, 126–138 (2016).

[24] Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature communications* **8**, 1–11 (2017).

[25] Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M. & Gardner, J. L. Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences* **113**, E3548–E3557 (2016).

[26] Meyniel, F., Maheu, M. & Dahaene, S. Human inferences about sequences: A minimal transition probability model. *Plos Comp. Biol.* **12** (2016).

[27] Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the american statistical association* **90**, 773–795 (1995).

[28] Berger, T. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications* (2003).

[29] Mountcastle, V. B., Talbot, W. H., Sakata, H. & Hyvärinen, J. Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys. neuronal periodicity and frequency discrimination. *Journal of neurophysiology* **32**, 452–484 (1969).

[30] Mountcastle, V., Steinmetz, M. & Romo, R. Frequency discrimination in the sense of flutter: psychophysical measurements correlated with postcentral events in behaving monkeys. *Journal of Neuroscience* **10**, 3032–3044 (1990).

[31] Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30** (2007).

[32] Heitz, R. P. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front. Neurosci.* **8** (2014).

[33] Sun, Y. & Wang, H. Perception of randomness: On the time of streaks. *Cognitive psychology* **61**, 333–342 (2010).

[34] Williams, J. J. & Griffiths, T. L. Why are people bad at detecting randomness? A statistical argument. *J. Exp. Psychol.* **39**, 1473–1490 (2013).

[35] Oskarsson, A. T., Van Boven, L., McClelland, G. H. & Hastie, R. What's next? Judging sequences of binary events. *Psychological bulletin* **135**, 262 (2009).

[36] Gilovich, T., Vallone, R. & Tversky, A. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* **17** (1985).

[37] Cassey, P., Hawkins, G. E., Donkin, C. & Brown, S. D. Using alien coins to test whether simple inference is Bayesian. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2015).

[38] Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E. & Shadlen, M. N. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *Journal of Neuroscience* **31**, 6339–6352 (2011).

[39] Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W. & Forstmann, B. U. Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience* **32**, 2335–2343 (2012).

[40] White, C. N. & Poldrack, R. A. Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **40**, 385 (2014).

[41] Leite, F. P. & Ratcliff, R. What cognitive processes drive response biases? a diffusion model analysis. *Judgment & Decision Making* **6** (2011).

[42] Gao, J., Tortell, R. & McClelland, J. L. Dynamic integration of reward and stimulus information in perceptual decision-making. *PloS one* **6**, e16749 (2011).

[43] Fan, Y., Gold, J. I. & Ding, L. Ongoing, rational calibration of reward-driven perceptual biases. *Elife* **7**, e36018 (2018).

[44] Afacan-Seref, K., Steinemann, N. A., Blangero, A. & Kelly, S. P. Dynamic interplay of value and sensory information in high-speed decision making. *Current Biology* **28**, 795–802 (2018).

[45] Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review* **113**, 700 (2006).

[46] Kim, T., Kabir, M. & Gold, J. I. Coupled decision processes update and maintain saccadic priors in a dynamic environment. *J. Neurosci.* **37**, 3632–3645 (2017).

[47] Nguyen, K. P., Josić, K. & Kilpatrick, Z. Optimizing sequential decisions in the drift-diffusion model. *J. Math Psychol.* **88** (2019).

[48] Goodie, A. S. & Fantino, E. Learning to commit or avoid the base-rate error. *Nature* **380**, 247–249 (1996).

[49] Goodie, A. S. & Fantino, E. What does and does not alleviate base-rate neglect under direct experience. *J. Behavioral Decision Making* **12**, 307–335 (1999).

[50] Welsh, M. B. & Navarro, D. J. Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes* **119**, 1–14 (2012).

[51] Bar-Hillel, M. The base-rate fallacy in probability judgments. *Acta Psychologica* **44**, 211–233 (1980).

[52] Hau, R., Pleskac, T. & Hertwig, R. Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *J. Behav. Decis. Making* **23** (2010).

[53] Fantino, E. & Navarro, A. Description–experience gaps: Assessments in other choice paradigms. *J. Behav. Dec. Making* **25** (2012).

[54] Weber, E. U., Shafir, S. & Blais, A. Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychol. Rev.* **111**, 430–445 (2004).

[55] Lieder, F. & Griffiths, T. L. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* **43** (2020).

[56] Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).

[57] Tavoni, G., Balasubramanian, V. & Gold, J. I. What is optimal in optimal inference? *Curr. Op. Behav. Sci.* **29** (2019).

[58] Bossaerts, P. & Murawski, C. Computational complexity and human decision-making. *Trends Cog. Sci.* **21** (2017).

[59] Bossaerts, P., Yadav, N. & Murawski, C. Uncertainty and computational complexity. *Phil. Trans. Roy. Soc. Lond.Series B* **374** (2019).

[60] Kool, W., Gershman, S. J. & Cushman, F. A. Planning complexity registers as a cost in metacontrol. *J. Cog. Neurosci.* **30** (2018).

[61] Polonio, L., Di Guida, S. & Coricelli, G. Strategic sophistication and attention in games: An eye-tracking study. *Game and Econ. Behav.* **94** (2015).

[62] Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *bioRxiv* (2020).

[63] Gershman, S. J. & Lai, L. The reward-complexity trade-off in schizophrenia. *bioRxiv* (2020).

[64] Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Ann. Rev. Psychol.* **62** (2011).

[65] Simon, H. A. A behavioral model of rational choice. *Quart. J. Econom.* **69** (1955).

[66] Bradbury, M. A. S., Hens, T. & Zeisberger, S. Improving investment decisions with simulated experience. *Rev. Finance* **19** (2015).

[67] M.Hogarth, R. & Soyer, E. Communicating forecasts: The simplicity of simulated experience. *J. Business Research* **68** (2015).

[68] M.Hogarth, R. & Soyer, E. Providing information for decision making: Contrasting description and simulation. *J. Applied Research in Memory and Cognition* **4** (2015).

[69] Kaufmann, C., Weber, M. & Haisley, E. Public understanding of climate change in the united states. *Management Science* **59** (2012).

[70] Livina, V. N., Havlin, S. & Bunde, A. Memory in the occurrence of earthquakes. *Physical review letters* **95**, 208501 (2005).

[71] Clauset, A., Young, M. & Gleditsch, K. S. On the frequency of severe terrorist events. *Journal of Conflict Resolution* **51**, 58–87 (2007).

[72] Weber, E. U. & Stern, P. C. Public understanding of climate change in the united states. *American Psychologist* **66**, 315–328 (2011).

[73] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

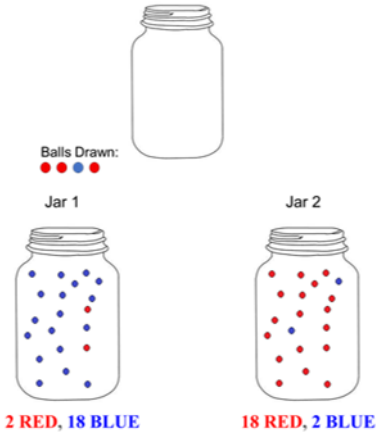

Figure S1: Example of the screen viewed by subjects on Amazon Mechanical Turk. A prompt at the bottom of the screen indicated to the subject to select the jar from which the sample was drawn.

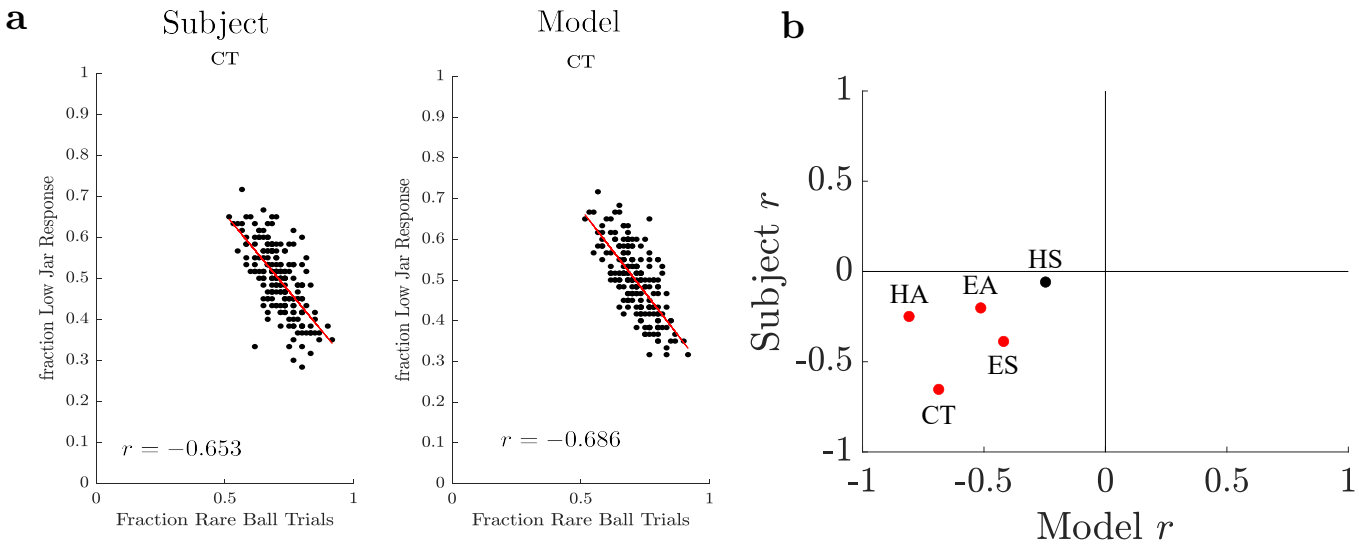

Figure S2: **a** Example of the fraction of rare-ball trials compared to the low-jar response frequency for individual subjects, and matched ideal observers (black dots). Non-parametric correlation coefficients (r) and linear regressions (red traces) shown. **b** Correlation coefficients r for the model and subjects in each block. Blocks where both the subjects and model had significant correlations ($p < 0.05$) are noted in red.

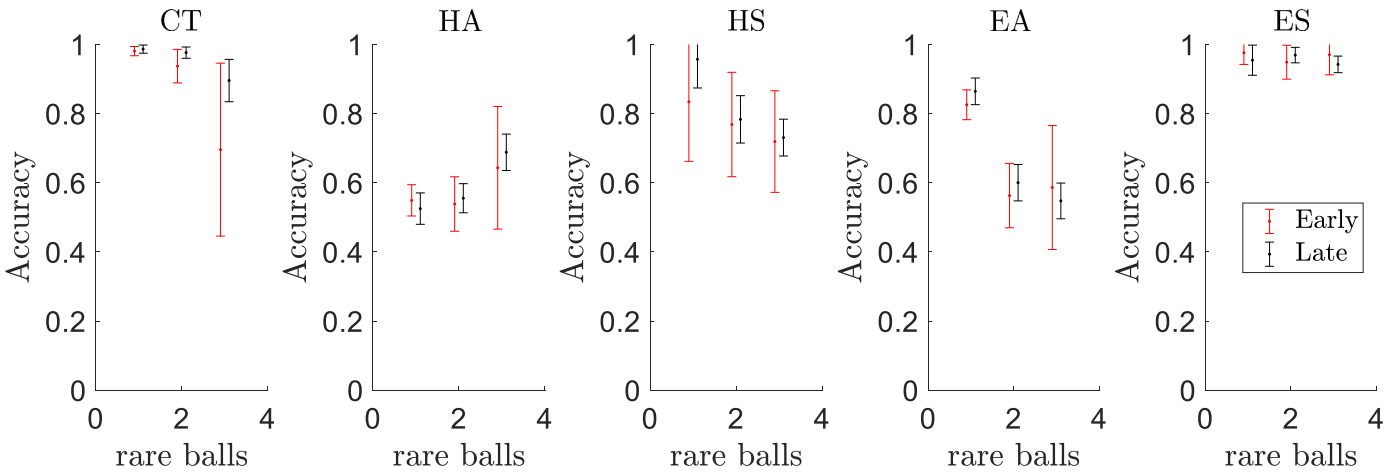

Figure S3: Fraction correct for 10-ball trials with 1, 2, or 3 rare balls, where the final rare ball occurred early (in the first 5 balls, observing from left to right) or late (in the last 5 balls) in the sequence of balls shown on the current trial. We found no systematic difference between the early and late trials, as evidenced by the bootstrapped population means and 95% confidence intervals.

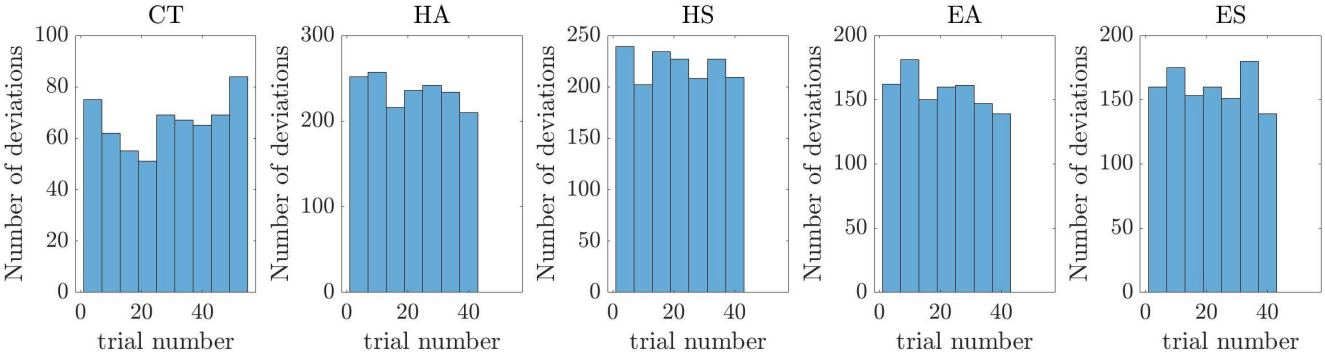

Figure S4: Histogram of frequency of subject deviations in response from the ideal observer based on trial number within a block, binned for every 6 trials. T-test of deviations in the first half versus the second half of the block using bootstrapped data showed no significant difference for any block.

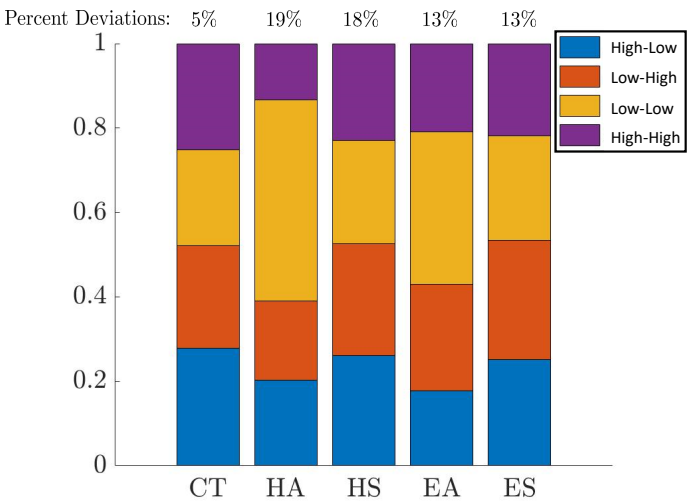

Figure S5: Fraction of ideal-observer response-deviation trials corresponding to each sequence (pre-deviation trial - deviation trial) within each block. Percentage of deviations per block reported at the top.

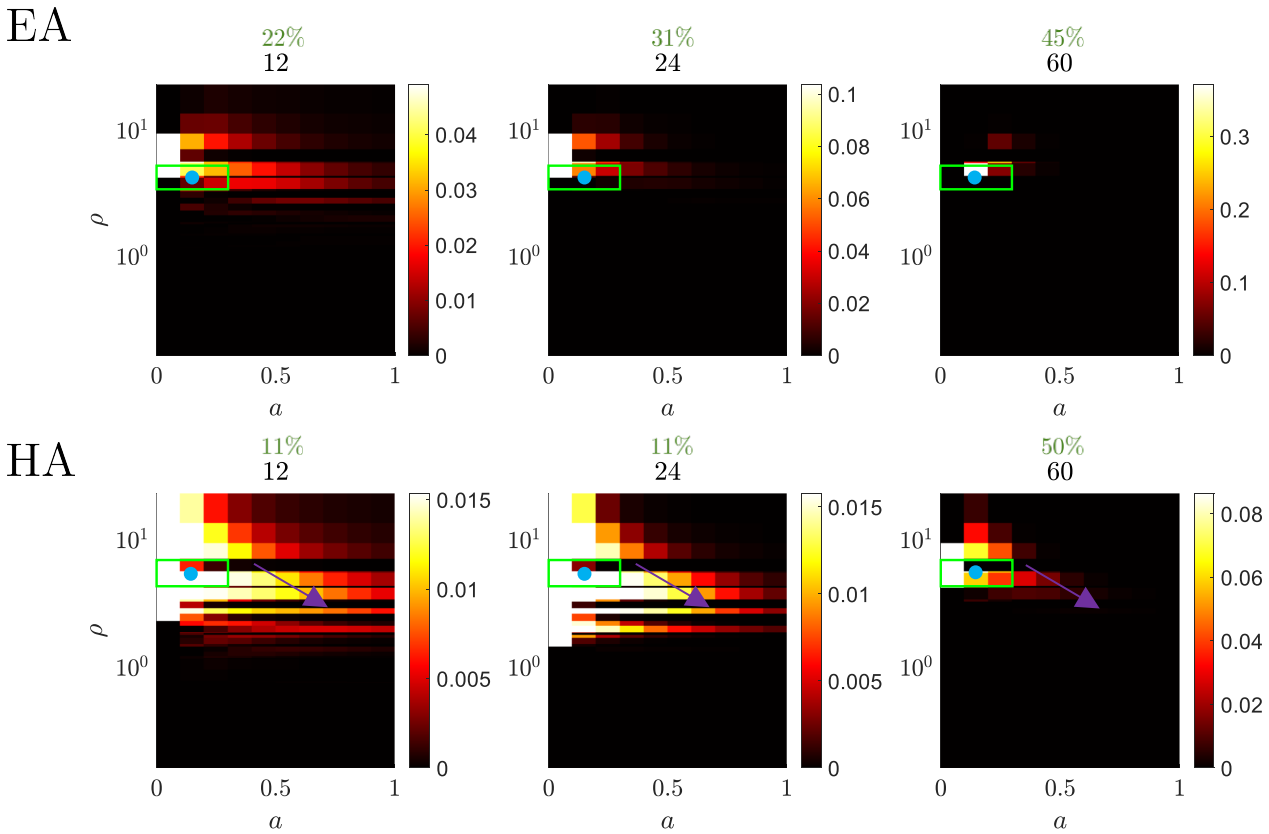

Figure S6: Example Bayesian parametric posteriors for the rare-ball weight ρ and noise variance a of the noisy Bayesian model using Bayesian sequential updating based on a flat prior over $0 \leq a \leq 1$ and $0 < \rho \leq 24.16$ (computed from $0.01 \leq h_{\pm} \leq 1$). Fits are based on synthetic data collected from varied block lengths (12, 24, and 60 trials, columns) of the asymmetric blocks based on responses produced using the ideal observer's ρ and a low level of noise ($a = 0.1$). True parameters are shown as blue dots. By 60 trials, the parameters are well identified, with over 40% of the posterior falling within a range of one parameter value (green box, corresponding percentages shown in green). Because a flat prior is used, there can be a high likelihood for alternative scenarios in which there is a trade-off between higher noise and lower ρ values, as shown by the arrows in the HA fits.

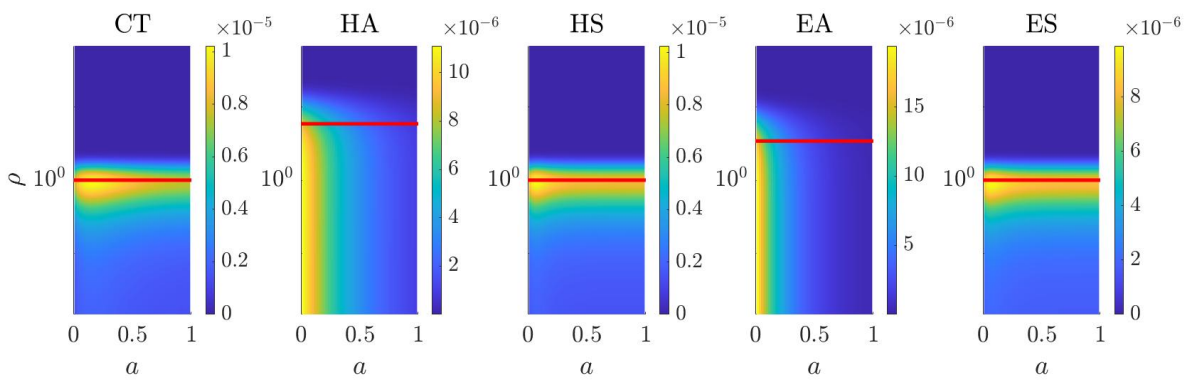

Figure S7: Examples of the weakened informative prior computed from pilot data of 20 subjects. Posteriors were computed for each of the 20 pilot subjects based on the Noisy Bayesian model with a flat prior and then averaged to produce a population posterior for each block. The averaged posterior was then smoothed to create an informative prior (symmetric blocks: L=2, c=5, asymmetric blocks: l=1, c=2). Red line shows the rare ball weighting ρ for the ideal observer in each block.

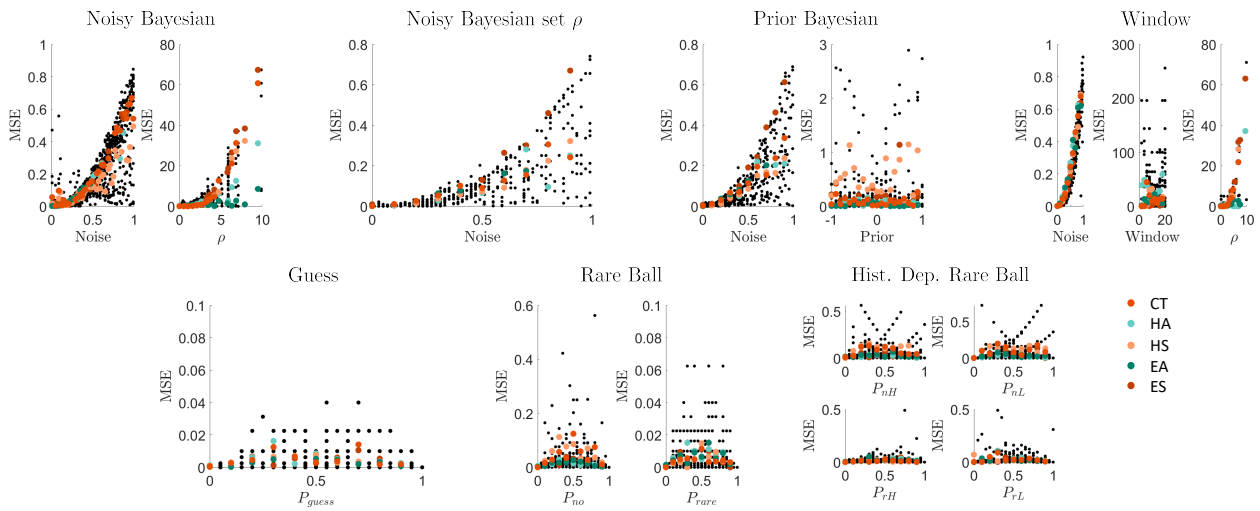

Figure S8: MSE of parameter fits for each model per block based on 100 synthetic datasets with randomly selected parameters using the informative prior developed in Fig. S7 for the Bayesian-based models (top row) and heuristic models (bottom row), as indicated. Black dots represent MSE for synthetic data with a given parameter value (x-axis, other model parameters randomly selected). Colored dots show binned averages for each block. Note that ρ is well fit for values near or below the ideal observer's rare ball weight, as was seen in the pilot human data. Thus, asymmetric averages (green dots) lay closer to zero at higher ρ values.

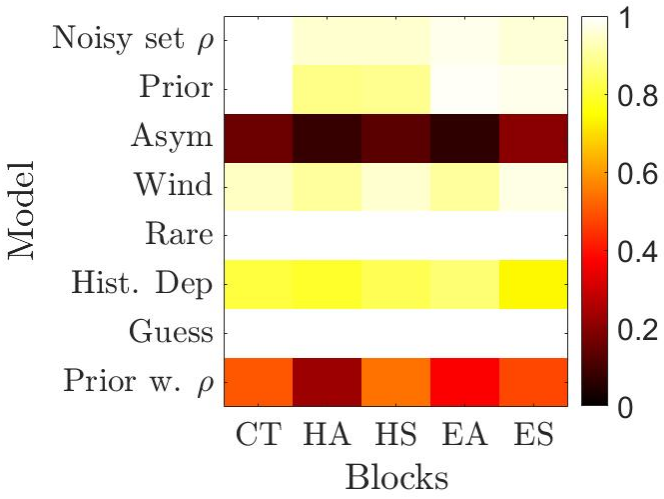

Figure S9: Fraction of times an alternative model was correctly identified as compared to the Noisy Bayesian model. Data from 100 synthetic subjects were produced using the human task structure (4 blocks with 42 trials, control block with 60 trials). Models are labeled and defined as in Fig. 4, except for two models that were not identifiable above 80% for all blocks ("Asym", which assumed an asymmetric repetition bias following a low-jar response, and "Prior w. ρ", which was the noisy Bayesian model with biased prior) and thus were not used in the analyses. The Noisy Bayesian model includes noise and a variable rare ball weight ρ. The Noisy Bayesian set ρ model sets ρ to the ideal observer's rare-ball weight (ρ_{IO}). The Prior Bayesian model includes a jar bias (prior) but ρ matching the ideal observer. The Windowing model applies the Noisy Bayesian model to a selected window of data beginning at the rightmost ball in the sample (the final ball, assuming subjects view the sample moving from left to right).The Rare-Ball model has a preference for either response based on the presence or lack of a rare balls. The History-Dependent Rare-Ball model extends the Rare Ball model to include the previous response. The Guess model guesses the high jar with some probability.

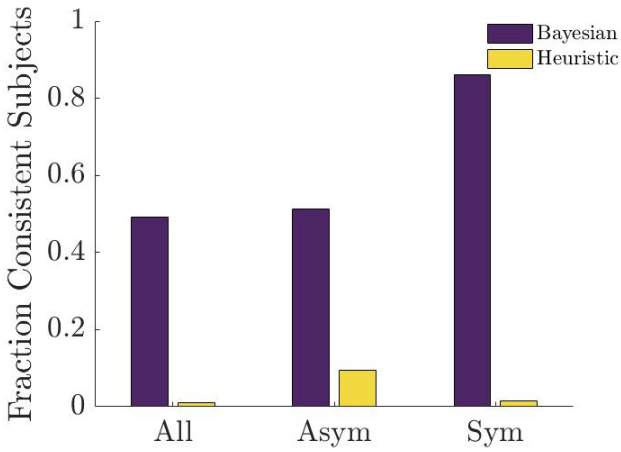

Figure S10: Fraction of subjects who were well matched to models in the same class (Bayesian or heuristic) across all blocks (All), asymmetric blocks (Asym) or symmetric blocks (Sym). Bootstrapped fractions of Bayesian and heuristic models are all significanty different from one another ($p < 0.05$).