RESEARCH ARTICLE

# Suboptimal human inference can invert the bias-variance trade-off for decisions with asymmetric evidence

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

**1** Department of Applied Mathematics, University of Colorado Boulder, Boulder, Colorado, United States of America, **2** Department of Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** Department of Mathematics, University of Houston, Houston, Texas, United States of America, **4** Department of Biology and Biochemistry, University of Houston, Houston, Texas, United States of America, **5** Institute of Cognitive Science, University of Colorado Boulder, Boulder, Colorado, United States of America

These authors contributed equally to this work.
* tahra.eissa@colorado.edu

## Abstract

Solutions to challenging inference problems are often subject to a fundamental trade-off between: 1) bias (being systematically wrong) that is minimized with complex inference strategies, and 2) variance (being oversensitive to uncertain observations) that is minimized with simple inference strategies. However, this trade-off is based on the assumption that the strategies being considered are optimal for their given complexity and thus has unclear relevance to forms of inference based on suboptimal strategies. We examined inference problems preregistered, online study. The participants tended to use suboptimal decision strategies that reflected an inversion of the classic bias-variance trade-off: some used complex, nearly normative strategies with mistuned evidence weights that corresponded to relatively high choice biases but lower choice variance, whereas others used simpler heuristic strategies that corresponded to lower biases but higher variance. These relationships illustrate structure in suboptimality that can be used to identify systematic sources of human errors.

## Introduction

Understanding how the brain makes inferences about the world requires first understanding the diversity of strategies individuals use to solve inference problems. One useful approach for

Bayesian strategies tended to have higher bias and lower variance, whereas subjects who

S1 Text ªTask and Recruitmentº for additional details). 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 ªrare-ballº color. The rare-ball color remained consistent throughout all blocks. Blocks were

were recruited only if they had a 95% or better from one of the jars, $\xi_{1:n}$, where $\xi_i = 1$ ($\xi_i = -1$) denote an observation of a rare (common) ball color. The ideal observer uses these observations to update the log likelihood ratio (*belief*), $z_n$ ˆ log $\frac{P…h\,ˆ\,h_{‡}\,jx_{1:n}†}{P…h\,ˆ\,h\quad jx_{1:n}†}$, 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 belief as:

$$z_n \quad ˆ \quad \underset{j\,ˆ\,1}{X} \log \frac{P…x_j jh_{‡}\,†}{P…x_j jh\quad †} \quad ˆ$$

and thus the magnitude of the belief increment is the same for either observation ($|C(+1)| = |C(-1)|$). When the environment is asymmetric, $h_- < 1 - h_+$, and different ball colors correspond to different evidence weights ($|C(+1)| 6 |C(-1)|$).

For $n$ ball draws, we can compute the probability of the responses (choices) on a given trial, $r = h_-$ and $r =$

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 $h_+ +$

likely to be the source of a sample. Thus, fits using this model had two free parameters: *a*

Here $LLR_b$ is the true LLR of each observed set [redacted]balls as computed using the ideal observer model. We fit the following parameters: 1) $\alpha$, th apse rate; 2) $\phi$, the LLR value at which each choice (high or low jar) is equally likely; and 3) the slope around the point $\phi$. Bias was defined as a non-zero value of $\phi$, so that positive egative) values correspond to biases towards (away) from the low jar. Noise was defined as 1/||, so that shallower functions correspond to higher noise.

Variance was defined as the weighted average of the absolute value of the residuals (mean absolute error),

$$v = \frac{1}{x}\sum_{i=1}^{x} n_i |P(r = h_+)_{b,i} - \rho_{b,i}|$$

where $x$ is the number of LLR values for a block, $n_i$ is the number of trials at a given LLR value, $\rho_{b,i}$ is the logistic fit for a given block-LLR, and $P(r = h_+)_{b,i}$ is the probability of a high jar response from the observer for a given block-LLR. Larger values of $v$ reflected more variance.

Our interpretation is based on the idea that noise is driven by either errors in the internal representation of the LLR or post-decision choice variability, whereas variance reflects strategies that are independent of the LLR. Based on the two model classes studied here (Bayesian and Heuristic), we find that models that rely on the LLR (Bayesian models) and the subjects best fit by them are fit with some noise but substantially less variance compared to models and subjects that use a pattern-based approach that does not depend on the LLR (Heuristic models). While there is correlation between the two metrics, heuristic subjects show substantially larger values for noise, which reflect the the poor logistic fits to these responses, and the conclusions of our analyses are comparable using either metric (see Supplementary Materials S5 Text ªNoise Versus Varianceº, S10 and S11 Figs, for more details).

## Model fitting and comparison

Parameter fitting. 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 ($\boldsymbol{\xi}$

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,

$$p(a, \mathrm{r} | r_{1:42}, \xi_{1:42}) = \frac{p(a, \mathrm{r})}{p(r_{1:42} | \xi_{1:42})} \prod_{j=1}^{42} p(r_j | a, \mathrm{r}, \xi_j).$$

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 ite computing the log likelihood ratio of the marginal likelihoods for any given pair of models,

$$\log\ldots\mathrm{BF}\dagger\ \hat{}\ \log\frac{P\ldots D\mathrm{j}M_2\dagger}{P\ldots D\mathrm{j}M_1\dagger}\ \hat{}\ \log\frac{P\ldots M_2\mathrm{j}D\dagger P\ldots M_1\dagger}{P\ldots M_1\mathrm{j}D\dagger P\ldots M_2\dagger}:$$

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, te

The MI with the inclusion of

## Algorithmic complexity

As in [9], algorithmic complexity isdescribedj 1.7mplexity

asymmetric blocks in Fig 3A and 3B but focus on asymmetric blocks in the remainder of the manuscript. Results from symmetric blocks can be found in Supplementary Materials S6 Text ªSymmetric Resultsº, for comparison purposes (S12 Fig).

Overall, the subjects' accuracy tended to be above chance (bootstrapped means and 95% confidence intervals were significantly above 0.5 for population data from each of the five blocks) and in many cases was qualitatively similar to that of the ideal observer under matched conditions (Fig 3A). Moreover, for asymmetric conditions both the ideal observer and the subjects had choice asymmetries in favor of the low jar that deviated from the prior (Fig 3B, bootstrapped means and 95% confidence intervals of low-jar responses significantly above 0.5).

However, the subjects also exhibited numerous suboptimalities in the asymmetric blocks. These suboptimalities included errors attributable to bias and variance (Fig 3C and 3D) that varied in magnitude across individual subjects but, in general, were larger than expected, given the responses of the ideal observer (Fig 3E and 3F). Although bias varied in magnitude and sign, most cases corresponded to an accentuation of choice asymmetry favoring the low jar. Likewise, variance ranged from zero, corresponding to choices that exactly matched the best-fitting logistic psychometric function, to near one, corresponding to choice patterns that deviated substantially from the best-fitting psychometric function. These effects were amplified by short sample lengths and task difficulty (see Supplementary Materials S4 Text ªChoice-Asymmetry Analysesº and S9 Fig for details).

## Formal model comparison

To relate these human behavioral patterns to particular inference strategies, we fit Bayesian-based and heuristic models separately to each individual subject's responses per block. We used Bayes factors to select the model that best matched each subject's responses on a given block and further confirmed the fits by cross-validating the subject responses with the best-fit model (S8 Fig). We then determined the bias-variance trends for each subject's best-fitting model based on the subjects' psychometric fits (details on model selection and fitting can be found in the Methods and Supplementary Materials S2 Text ªModel Fittingº and S3 Text ªSubject Model Fittingº, S6 and S7 Figs).

Three models we used were Bayesian-based (Fig 4A). The first model assumed that the observer makes decisions based on a noisy version of the log-likelihood, in which noise was a normally distributed random variable with zero mean and a free parameter for variance, and $\rho$ was a free parameter representing the belief update in response to observing a rare ball (ªNoisy Bayesianº). When $\rho > 1$, the model weighted a rare-ball observation more strongly than an observation of a common ball. For the second model, we set $\rho$ to the ideal observer's rare-ball weight. Without noise, this version di (()Tj0.05md9brizwytbk14 -1.29 1.7235m2 1627-1.2983d (35 assumption that the observer chooses the high jar with some probability whenever one or more rare balls are observed (ªRare Ballº). This assumption is equivalent to fixing the threshold parameter in the Variable Rare Ball model to 1. The third model described a simple guessing complexity. The first approach was purely data-driven, allowing us to avoid making assumptions about the specific, algorithmic form of each strategy. This approach was based on the idea that efficient inference strategies solve an ªinformation bottleneckº problem [10], which is closely related to lossy data compression and rate-distortion theory [11]; i.e., maximizing predictive accuracy for a fixed information budget. Specifically, for this approach we computed two quantities using data separately from each subject and block: 1) strategic complexity, measured as the mutual information (MI) between the subject's observations (the samples of balls observed on each trial) and their choices in the given block (Fig 6A), where larger values implied that the known ball sample reduced uncertainty in a subject's choice; and 2) strategic effectiveness, measured as the proximity of the subject's accuracy to the maximum achievable accuracy given their strategic complexity (termed the ªoptimal accuracy boundº; for details see the ªComplexity Analysesº S7 Text of the Supplemental Materials), where smaller values implied that the strategy was being used more effectively to generate correct choices for a given level of complexity. Note, high complexity does not necessarily imply high accuracy since complex strategies could use irrelevant information and/or be ineffective, increasing the distance to the maximal achievable accuracy.

Thrue, Tmon mod (ming)Tj j 1.3832 0 Td irrelevant block (Note, mod (more-athe9388increasin9180j 37

In general, subjects who used more-complex strategies (i.e., those who used more information from the current trial to make choices) were more accurate, withchoicesgj 6.0945

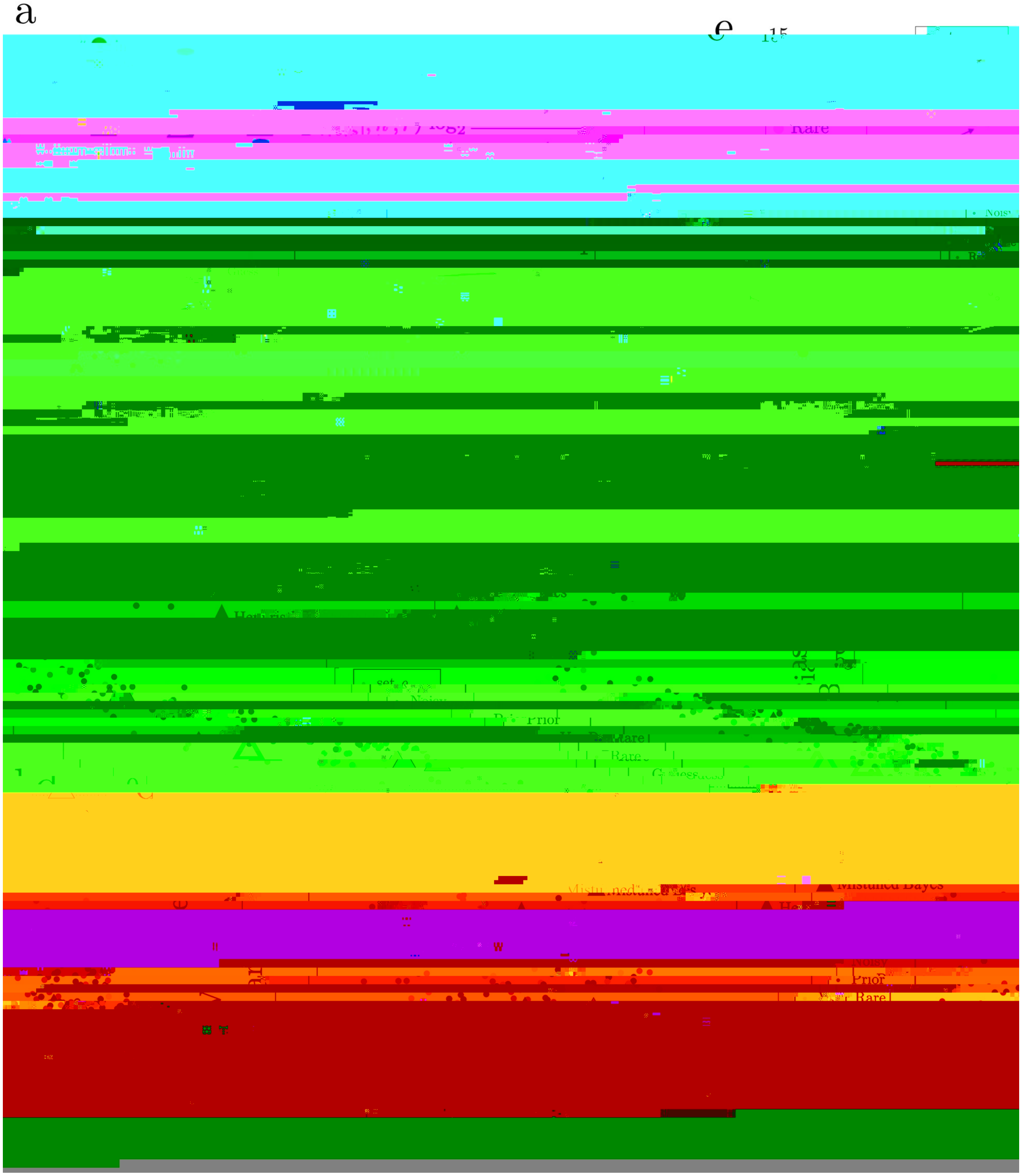

Fig 6. More complex but suboptimal human strategies exhibited more bias. a. Mutual information (MI) between the number of rare balls in a sample ($|\xi|$), the sample length ($n$), and the response ($r$) for each subject a0l4to

the task. Based on our assignments, this metric showed a sample-length dependent scaling in Bayesian complexity, but still confirmed that measures of complexity for the Bayesian models were much larger than those of heuristics (Fig 6E). These model-based results support the idea that the observed patterns of bias and variance are inherent to the relationship between the strategies described by these models and not simply idiosyncrasies of the subjects' behavioral patterns, with errors in more-complex Bayesian-like strategies leading to increased biases, but less-complex strategies based on the pattern of observations leading to increased variance (details of this analysis can be found in Supplementary Materials S7 Text ªComplexity Analysesº, S16 Fig).

## Discussion

How do people's error trends depend on the inference strategies they use? We examined the properties of errors made by human subjects performing a two-alternative forced-choice task with asymmetric evidence [7, 16, 17]. The evidence took the form of two colors of balls drawn from jars, such that one (ªrareº) color was drawn less often than the other. Similar to ideal observers, most subjects exhibited a choice asymmetry favoring the option that produced fewer rare balls. In addition, subjects fell into two categories depending on the type of strategy that best described their responses. Subjects described by heuristic strategies, which were based on less information and fewer algorithmic operations, displayed substantially more choice variability but comparable choice asymmetry to the ideal observer. In contrast, subjects described by more-complex, mistuned Bayesian strategies displayed minimal increases in choice variability but much more bias than the ideal observer. These effects reflected the nature of the suboptimalities introduced by each strategy type: the heuristic strategies we considered did not take into account specific task features responsible for choice asymmetries and thus tended to add variability, whereas the Bayesian-like strategies that we considered did attempt to model those features explicitly but, when implemented suboptimally (mistuned) by the subjects, tended to exacerbate asymmetries inherent in such decision rules.

### Inversion of the bias-variance trade-off

These findings provide new insights into the generalizability of bias-variance trade-offs that are well established in machine learning and related fields [2, 3] and can be used to account for individual differences in human behavior under certain conditions [1, 4]. Bias-variance trade-offs can be conceptualized in terms of fitting various functions that differ in complexity (e.g., polynomial order) to noisy data whose generative source is unknown. Typically, simpler (e.g., linear) models tend to have higher bias, because they miss higher-order (e.g., nonlinear) features of the generative source, but lower variance, because their best-fitting parameters are relatively stable across different data instances. In contrast, more complex (e.g., high-order polynomial) models tend to have lower bias, because they can capture complex features of the data, but higher variance, because the specific features they capture can differ across different data instances.

Critically, this traditional

optimality. Specifically, we considered two broad classes of strategies that could result in suboptimalities either from the model used or a mistuning of the parameters. In the context of asymmetric evidence, these suboptimalities introduced errors that could invert the bias-variance trade-off. However, this inversion only manifested when considering the relationship of complexity across model classes in asymmetric contexts. In contrast, decreases in complexity within a model class in asymmetric contexts produced increases in both bias and variance, regardless of model class. Therefore, our results suggest that the inversion of the bias-variance trade-off arises in particular situations, such as when suboptimal strategies are used in asymmetric environments, and may choice biases in tasks with symmetric evidence but asymmetries in expected choice frequencies [30±33] or reward outcomes [32, 34±37]. 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 [30] and/or sequential analysis [38]. 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 [31, 32, 34, 35, 39, 40], and, more generally, is consistent with many previous findings of mistuned priors [41±45]. Alternatively, while our Prior Bayesian model described changes in choice asymmetry that were attributed to biased priors without impacts to the ideal evidence weights, it is plausible that the ideal observer model and its mistuned Bayesian variants could be implemented by a competitive neural network model with plastic synapses that could represent the evidence asymmetry of rare balls observations and their responses [49]. Moreover, the presence and amplitude of rewards shapes task attention [50], which could be reflected in strategy usage.

In this task, suboptimality took three forms: 1) underweighting rare balls; 2) biased priors in favor of the low jar; and 3) applying heuristics, which occurred predominantly in harder tasks. We hypothesize that underweighting may be the result of weighting biases in favor of symmetric weights, rather than a mistuning relative to the ideal-observers weights, given that subject's rare-ball parameters showed comparable values for both easy and hard asymmetric blocks. Likewise, the mistuning of subjects' priors in favor of the low jar may reflect a recency bias, in which previous low-jar responses encourage subjects to repeat their choice [51, 52]. Finally, the use of heuristic strategies in more complex tasks (e.ge

S7 Text. Complexity analyses.
(DOCX)

S1 Fig. Example of the screen viewed by subjects on Amazon Mechanical Turk. The details of the current set of jars were available to participants on every trial. A prompt at the bottom of the screen indicated to the subject to select the jar from which the sample was drawn.
(TIF)

S2 Fig. Inattentive subjects. Accuracy for each subjects' interspersed control trials to test for attentiveness (3 interspersed blocks of 12 trials). Inattentive subjects were defined as those whose accuracy was 50% or lower on two or more interspersed control blocks (3 subjects identified, red lines). These subjects were excluded from all further analyses.
(TIF)

S3 Fig. Trial identification. Examples of the Bayesian parametric posteriors of the Noisy Bayesian model with a flat prior over the noise variance 0 $a$ 1 and the rare-ball weight 0 < $\rho$ 24.16 (computed from jars with rare-ball probabilities 0.01 $h_{\pm}$ 1). Posteriors are based on synthetic responses from a Noisy Bayesian model whose assumes that $\rho$ equals the ideal observer's rare-ball

of rare balls that must be drawn (e.g., 1 up to 2) to trigger a ªhighº response, generating a sawtooth-shaped response fraction function of ball number. Right: The overall (correct and incorrect trials) low-jar response probability for the ideal observer shows a general decrease in choice asymmetry as sample size increases. However, the effect is accompanied by the sawtooth structure depicted in the center panels.
(TIF)

S10 Fig. Noise variance comparison. Top: Estimated noise and variance from psychometric functions fit to individual subject data (points). Noise and variance showed a significant correlation in all blocks: Control (CT), Hard Asymmetric (HA), Hard Symmetric (HS), Easy Asymmetric (EA), Easy Symmetric (ES) (Spearman's Correlation, $p < 0.05$). Center: Same data as in the top row, but color coded by each subject's best-fit models for each block. In general, heuristic subjects had the largest values of variance and noise. Triangles represent medians for each model group. Filled triangles differ significantly from the Nearly Ideal subjects (two-sided Wilcoxon rank-sum test, $p < 0.05$). 2.6702 0 Td (rank-sum)Trman's

variance, consistent with general trends of better (less variable) performance associated with more-complex strategies.
(TIF)

S14 Fig. Mutual information with previous response. Across-group bias-variance relationships were robust to a measure of mutual information (MI) that took into account not just the balls observed on the current trial (i.e., relevant information, as in Fig 6A)) but also the previous choice (i.e., irrelevant information), for the two asymmetric blocks (columns, as indicated). a: Accuracy versus MI. The bound is the maximum accuracy attainable by the idea

**11.** Berger T. Rate-distortion theory. Wiley Encyclopedia of Telecommunications. 2003;.

**12.** Bossaerts P, Murawski C. Computational Complexity and Human Decision-Making. Trends Cog Sci. 2017; 21. https://doi.org/10.1016/j.tics.2017.09.005

**13.** Bossaerts P, Yadav N, Murawski C. Uncertainty and computational complexity. Phil Trans Roy Soc LondSeries B. 2019; 374. https://doi.org/10.1098/rstb.2018.0138

**14.** Kool W, Gershman SJ, Cushman FA. Planning Complexity Registers as a Cost in Metacontrol. J Cog Neurosci. 2018; 30. https://doi.org/10.1162/jocn_a_01263

**15.** Balasubramanian V. Bayesian inference, and the geometry of the space of probability distributions. In: in Advances in Minimum Description

**36.** Fan Y, Gold JI, Ding L. Ongoing, rational calibration of reward-driven perceptual biases. Elife. 2018; 7: e36018. https://doi.org/10.7554/eLife.36018

**37.** Afacan-Seref K, Steinemann NA, Blangero A, Kelly SP. Dynamic interplay of value and sensory information in high-speed decision making. Current Biology. 2018; 28(5):795±802. https://doi.org/10.1016/j.cub.2018.01.071

**38.** Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review.