How Occam's razor guides human decision-making

Eugenio Piasini[1,2,†,*], Shuze Liu[2,†], Pratik Chaudhari[2], Vijay Balasubramanian[2,3,‡], and Joshua I. Gold[2,‡]

[1]International School for Advanced Studies (SISSA), Trieste, Italy
[2]University of Pennsylvania, Philadelphia PA
[3]Santa Fe Institute, Santa Fe NM
[†,‡]Equal contribution
[*]epiasini@sissa.it

Abstract

Occam's razor is the principle that, all else being equal, simpler explanations should be preferred over more complex ones. This principle is thought to play a role in human perception and decision-making, but the nature of our presumed preference for simplicity is not understood. Here we use preregistered behavioral experiments informed by formal theories of statistical model selection to show that, when faced with uncertain evidence, human subjects exhibit preferences for particular, theoretically grounded forms of simplicity of the alternative explanations. These forms of simplicity can be understood in terms of geometrical features of statistical models treated as manifolds in the space of the probability distributions, in particular their dimensionality, boundaries, volume, and curvature. The simplicity preferences driven by these features, which are also exhibited by artificial neural networks trained to optimize performance on comparable tasks, generally improve decision accuracy, because they minimize over-sensitivity to noisy observations (i.e., overfitting). However, unlike for artificial networks, for human subjects these preferences persist even when they are maladaptive with respect to the task training and instructions. Thus, these preferences are not simply transient optimizations for particular task conditions but rather a more general feature of human decision-making. Taken together, our results imply that principled notions of statistical model complexity have direct, quantitative relevance to human and machine decision-making and establish a new understanding of the computational foundations, and behavioral benefits, of our predilection for inferring simplicity in the latent properties of our complex world.

Occam's razor formalized as model selection

To make decisions in the real world, we must often choose between multiple, plausible explanations for noisy, sparse data. When evaluating such competing explanations, Occam's razor says that we should consider not just how well they account for the observed data, but also their potentially excessive flexibility in describing alternative, and potentially irrelevant, data that have not been observed ((*1*); e.g., "a ghost did it!", Figure 1a). In cognitive science, simplicity, or parsimony, has long been proposed as an organizing principle in mental function (*2*), from the early concept of Prägnanz in Gestalt psychology (*3*), to a number of "minimum principles" for vision (*4*), to theories that posit a central role for data compression in cognition (*5*). However, despite evidence that human decision-makers can exhibit simplicity preferences under certain task conditions (*6–10*), we lack a principled understanding of what, exactly, constitutes the "simplicity" that is favored (or, equivalently, "complexity" that is disfavored) and how we balance that preference with the evidence provided by the observed data when we make decisions.

To provide this understanding, we turn to an approach based on Bayesian statistics (*11*, *12*), which allows us to measure the complexity of an explanation for data on an absolute scale. Our process is formalized as a model-selection problem: given a set X of N observations and a set of possible statistical models $\mathcal{M}_1, \mathcal{M}_2, \ldots$, we seek the model $\mathcal{M}$ that in some sense is the best for the data X. In this context, Occam's razor can be interpreted as requiring the goodness-of-fit of a model to be penalized by some measure of its flexibility, or complexity, when comparing it against other models. Bayesian statistics offers a natural characterization of such a measure of complexity and specifies the way in which it should be traded off

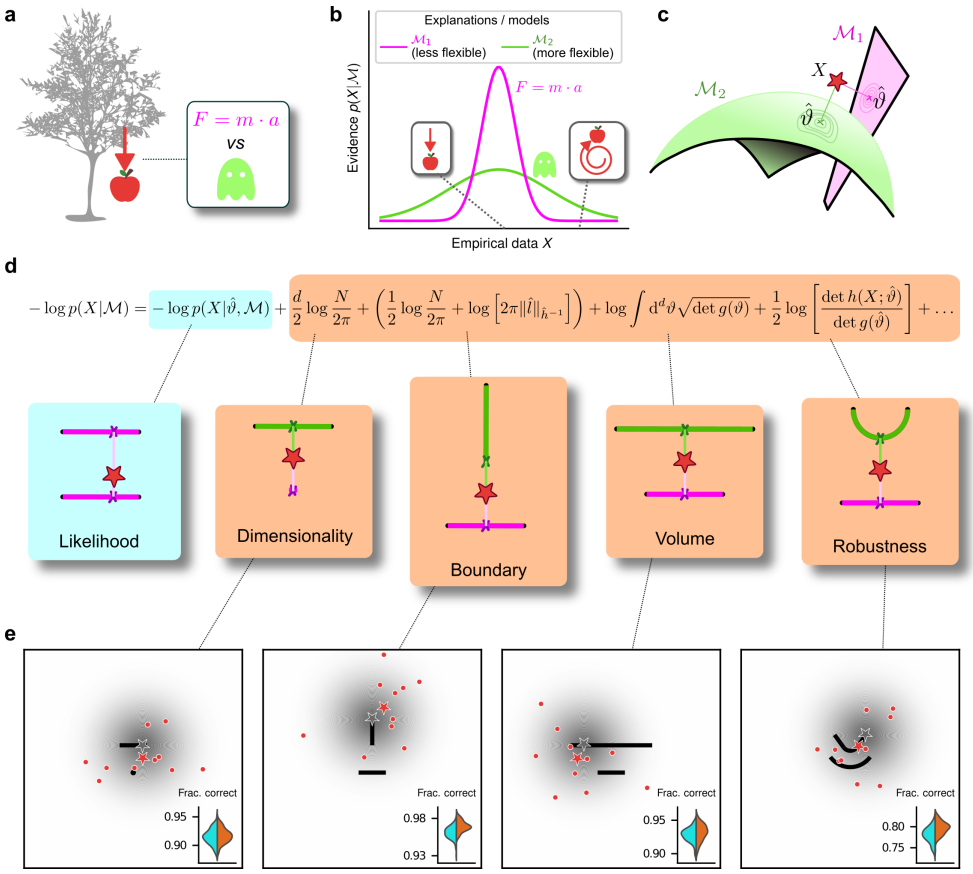

Figure 1: *Formalizing Occam's razor as Bayesian model selection to understand simplicity preferences in human decision-making.* **a**: Occam's razor prescribes an aversion to complex explanations (models). In Bayesian model selection, model complexity is a measure of the flexibility of a model, or its capacity to account for a broad range of empirical observations. In this example, we observe an apple falling from a tree (left) and compare two possible explanations: 1) classical mechanics, and 2) the intervention of a ghost. **b**: Schematic comparison of the evidence of the two models in **a**. Classical mechanics (pink) explains a narrower range of observations than the ghost (green), which is a valid explanation for essentially any conceivable phenomenon (e.g., both a falling and spinning-upward trajectory, as in the insets). Absent further evidence, Occam's razor posits that the simpler model (classical mechanics) is preferred, because its hypothesis space is more concentrated around the sparse, noisy data and thus avoids "overfitting" to noise. **c**: A geometrical view of the model-selection problem. Two alternative models are represented as geometrical manifolds, and the maximum-likelihood point $\hat{\vartheta}$ for each model is represented as the projection of the data (red star) onto the manifolds. **d**: Systematic expansion of the log evidence of a model $\mathcal{M}$. $\hat{\vartheta}$ is the maximum-likelihood point on model $\mathcal{M}$ for data X, N is the number of observations, d is the number of parameters of the model, $\hat{l}$ is the likelihood gradient evaluated at $\hat{\vartheta}$, h is the observed Fisher information matrix, and g is the expected Fisher information matrix (see Methods). $g(\vartheta)$ captures how distinguishable elements of $\mathcal{M}$ are in the neighborhood of ϑ. When $\mathcal{M}$ is the true source of the data X, $h(X;\vartheta)$ can be seen as a noisy version of $g(\vartheta)$, estimated from limited data. $\hat{h}^{-1}$ is a shorthand for $h(X;\hat{\vartheta})^{-1}$, and $\|\hat{l}\|_{\hat{h}^{-1}} = \sqrt{\hat{l}^T\hat{h}^{-1}\hat{l}}$ is the length of $\hat{l}$ measured in the metric defined by $\hat{h}^{-1}$. The ellipsis collects terms that decrease as N grows. Each term of the expansion represents a distinct geometrical feature of the model (*11*): dimensionality penalizes models with many parameters; boundary (a novel contribution of this work) penalizes models for which $\hat{\vartheta}$ is on the boundary; volume counts the number of distinguishable probability distributions contained in $\mathcal{M}$; and robustness captures the shape (curvature) of $\mathcal{M}$ near $\hat{\vartheta}$. **e**: Psychophysical task with variants designed to probe each geometrical feature in **d**. For each trial, a random location on one model was selected (gray star), and data (red dots) were sampled from a Gaussian centered around that point (gray shading). The red star represents the empirical centroid of the data, by analogy with **c**. The maximum-likelihood point can be found by projecting the empirical centroid onto one of the models. Subjects saw the models (black lines) and data (red dots) only and were required to choose which model was best for the data. Insets: task performance for the given task variant, for a set of 100 simulated ideal Bayesian observers (orange) versus a set of 100 simulated maximum-likelihood observers (i.e., choosing based only on whichever model was the closest to the empirical centroid of the data on a given trial; cyan).

against goodness-of-fit to maximize decision accuracy, typically because the increased flexibility provided by increased complexity tends to cause errors by overfitting to noise in the observations (*12–14*).

Specifically, according to this framework models should be compared based on their evidence or marginal likelihood $p(X|M) = \int \mathrm{d}\vartheta w(\vartheta) p(X|\mathcal{M}, \vartheta)$, where ϑ represents model parameters and $w(\vartheta)$ their associated prior (Figure 1b). Under mild regularity assumptions and with sufficient data, the (log) evidence can be written as the sum of the maximum log likelihood of $\mathcal{M}$ and several penalty factors (Figure 1d). These penalty factors, which are found even when the prior probabilities of the models under consideration are equal (i.e., independent of the data, all are equally likely to be the correct model), can be interpreted as providing quantitatively defined preferences against certain models according to specific forms of complexity that they embody (*11*, *12*). If the prior over parameters $w(\vartheta)$ is taken to be uninformative (*15*), each penalty factor can be shown to capture a distinct geometric property of the model (*11*), including dimensionality (number of parameters), boundary (a novel term, detailed below), volume, and shape (Figure 1c,d). This approach, which we call the Fisher Information Approximation (FIA), generalizes the well-known Bayesian Information Criterion (BIC) for model selection (*16*, *17*). Its effectiveness has been demonstrated by using it to identify worse-fitting, but better-generalizing, psychophysical models describing the relationship between physical variables (e.g., light intensity) and their psychological counterparts (e.g., brightness) (*18*). Similar quantitative definitions of statistical model complexity or model-selection prescriptions can be obtained with different theoretical approaches, such as the Minimum Description Length (*19*, *20*), Minimum Message Length (*21*), and Predictive Information (*22*) frameworks, testifying to the generality of this approach (*23*).

A limitation of these existing approaches is that they typically assume that the maximum-likelihood solution is in the interior of the parameter space of a given model (*11*). In contrast, because models are just approximations of the true processes in the real world that generated a given set of observations, those observations may fall outside of the parameter space of a given model. In these cases (or even when the observations are based on samples generated by the model but are corrupted by noise to fall outside of the model's parameter space), the maximum-likelihood solution for that model, given those data, may fall on the boundary of the model's parameter space. To account for this condition, we extended the FIA to deal with the simple case of a linear boundary in parameter space (see Methods). When the maximum-likelihood solution is on such a boundary, an additional penalty term appears in the FIA, which we denote "boundary" (Figure 1d). This extended FIA, consisting of dimensionality, boundary, volume, and robustness terms, provides a quantitative framework for assessing simplicity preferences in simple decision tasks, as we detail below.

Humans exhibit theoretically grounded simplicity preferences

We designed a simple decision-making task to relate the FIA complexity terms to the potential preferences exhibited by both human and artificial decision-makers. For each trial, $N = 10$ simultaneously presented observations (red dots in Figure 1e) were sampled from a 2D Normal ("generative") distribution centered somewhere within one of two possible shapes (black shapes in Figure 1e). The identity of the shape generating the data (top versus bottom) was chosen at random with equal probability. Likewise, the location of the center of the Normal distribution within the selected shape was sampled uniformly at random, in a way that did not depend on the model parametrization, by using Jeffrey's prior (*15*). Given the observations, the subjects decided which shape (model) was more likely to contain the center of the generative distribution. We used four task variants, each designed to primarily probe one of the distinct geometrical features that are penalized in Bayesian model selection (i.e., a Bayesian observer is expected to have a particular, quantitative preference away from the more-complex alternative in each pair; Figure 1d and e). In our task, the FIA provided a good approximation of the exact Bayesian posterior (Supplementary Information section C.1 and Supplementary Figure S1).

For our human studies, we used the on-line research platforms Pavlovia, to implement the task, and Prolific, to recruit subjects. Following our preregistered approaches (*24–26*), we collected data from 202 subjects, divided into four groups that each performed one of the four separate versions of the task depicted in Figure 1e (each group comprised 50 subjects; see Supplementary Text for details on subject demographics). We provided instructions that used the analogy of seeds from a flower located in one of two flowerbeds, to provide an intuitive framing of the key concepts of noisy data generated by a particular instance of a parametric model from one of two model families. To minimize the possibility that

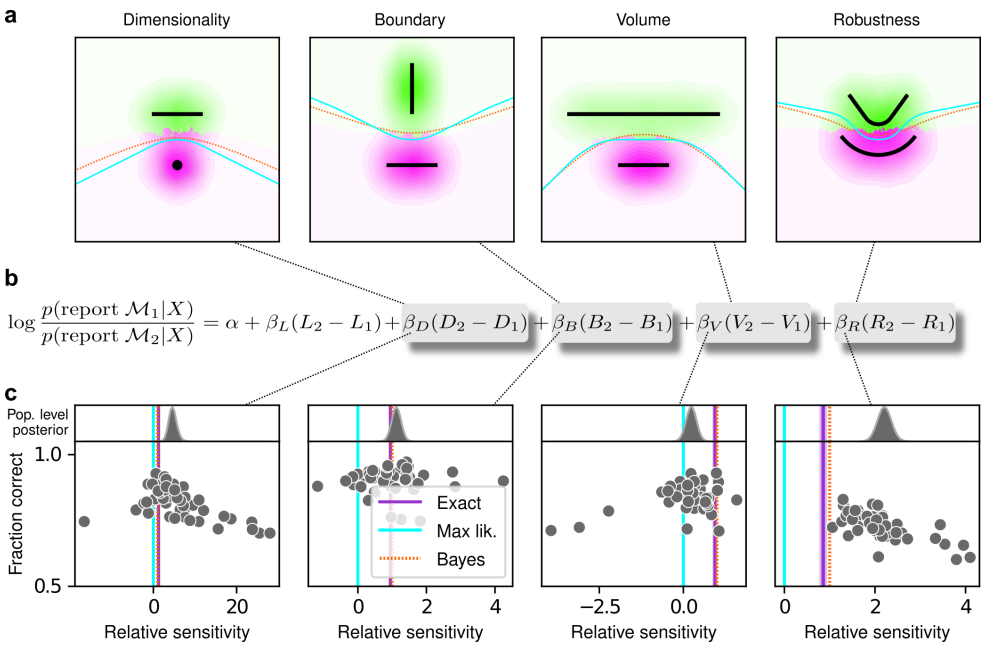

Figure 2: *Humans exhibit theoretically grounded simplicity preferences*. **a**: Summary of human behavior. Hue (pink/green): k-nearest-neighbor interpolation of the model choice, as a function of the empirical centroid of the data. Color gradient (light/dark): marginal density of empirical data centroids for the given model pair, showing the region of space where data were more likely to fall. Cyan solid line: decision boundary for an observer that always chooses the model with highest maximum likelihood. Orange dashed line: decision boundary for an ideal Bayesian observer. The subjects' choices tended to reflect a preference for the simpler model, particularly near the center of the screen, where the evidence for the alternatives was weak. For instance, in the left panel there is a region where data were closer to the line than to the dot, but subjects chose the dot (the simpler, lower-dimensional "model") more often than the line. **b**: Subject sensitivity to each geometrical feature characterizing model complexity was estimated via hierarchical logistic regression (see Methods, section A.5, Supplementary Figure S2 and Supplementary Text section C.3), using as predictors a constant to account for an up/down choice bias, the difference in likelihoods for the two models ($L_2 - L_1$) and the difference in each FIA term for the two models ($D_2 - D_1$, etc). Following the hierarchical regression scheme, the subject-level sensitivities were in turn modeled as being sampled from a population-level distribution. The mean of this distribution is our population-level estimate for the sensitivity. **c**: Overall accuracy versus estimated relative FIA sensitivity for each task condition, as indicated. Points are data from individual subjects. Each fitted FIA coefficient was normalized to the likelihood coefficient and thus could be interpreted as a relative sensitivity to the associated FIA term. For each term, an ideal Bayesian observer would have a relative sensitivity of one (dashed orange lines), whereas an observer that relied on only maximum-likelihood estimation (i.e., choosing "up" or "down" based on only the model that was the closest to the data) would have a relative sensitivity of zero (solid cyan lines). Top, gray: Population-level estimates (posterior distribution of population-level relative sensitivity given the experimental observations). Bottom: each gray dot represents the task accuracy of one subject (y axis) versus the posterior mean estimate of the relative sensitivity for that subject (x axis). Purple: relative sensitivity of an ideal observer that samples from the exact Bayesian posterior (not the approximated one provided by the FIA). Shading: posterior mean ± 1 or 2 stdev., estimated by simulating 50 such observers.

subjects would simply learn from implicit or explicit feedback over the course of each session to make more optimal (i.e., simplicity-preferring) choices of flowerbeds, we: 1) used conditions for which the difference in performance between ideal observers that penalized model complexity according to the FIA and simulated observers that used only model likelihood was ∼1% (depending on the task type; Figure 1e, insets), which translates to ∼5 additional correct trials over the course of an entire experiment; and 2) provided feedback only at the end of each block of 100 trials, not each trial. We used hierarchical (Bayesian) logistic regression to measure the degree to which each subject's choices were affected by model likelihood (distance from the data to a given model) and each of the FIA-derived geometrical features characterizing model complexity (see Methods, section A.5). We defined each subject's sensitivity to each FIA term as a normalized quantity, relative to their likelihood sensitivity (i.e., by dividing the logistic coefficient associated with a given FIA term by the logistic coefficient associated with the likelihood).

The human subjects were sensitive to all four forms of model complexity (Figure 2). Specifically, the estimated normalized population-level sensitivity for human subjects (posterior mean ± st. dev., where zero implies no sensitivity and one implies Bayes-optimal sensitivity) was 4.66±0.96 for dimensionality, 1.12±0.10 for boundary, 0.23±0.12 for volume, and 2.21±0.12 for robustness (note that, following our preregistered plan, we emphasize parameter estimation using Bayesian approaches (*27–29*) here and throughout the main text, and we provide complementary null hypothesis significance testing in the Supplementary Text, Section C.7 and Supplementary Table S4). Formal model comparison (WAIC; see Supplementary Text, section C.7.1 and Supplementary Tables S2 and S3) confirmed that their behavior was better described by taking into account the geometric penalties defined by the theory of Bayesian model selection, rather than by relying on only the minimum distance between model and data (i.e., the maximum-likelihood solution).

The subjects also exhibited substantial individual variability in performance that included ranges of sensitivities to each FIA term that spanned optimal and sub-optimal values. This variability was large compared to the uncertainty associated with subject-level sensitivity estimates (Supplementary Text, section C.5 and Supplementary Figure S3) and impacted performance in a manner that highlighted the usefulness of appropriately tuned (i.e., close to Bayes optimal) simplicity preferences: accuracy tended to decline for subjects with FIA sensitivities further away from the theoretical predictions (Figure 2c; posterior mean ± st. dev. of Spearman's rho between accuracy and $|\beta - 1|$, where β is the sensitivity: dimensionality, -0.69±0.05; boundary, -0.21±0.11; volume, -0.10±0.10; robustness, -0.54±0.10). The sub-optimal sensitivities exhibited by many subjects did not appear to result simply from a lack of task engagement, because FIA sensitivity did not correlate with errors on easy trials (posterior mean ± st. dev. of Spearman's rho between lapse rate, estimated with an extended regression model detailed in Methods, section A.5.1, and the absolute difference from optimal sensitivity for: dimensionality, 0.08±0.12; boundary, 0.15±0.12; volume, -0.04±0.13; robustness, 0.15±0.14; see Supplementary Text section C.6 and Supplementary Figure S4). Likewise, sub-optimal FIA sensitivity did not correlate with weaker likelihood sensitivity for the boundary (rho=-0.13±0.11) and volume (-0.06±0.11) terms, although stronger, negative relationships with the dimensionality (-0.35±0.07) and robustness terms (-0.56±0.10) suggest that the more extreme and variable simplicity preferences under those conditions (and lower performance, on average; see Figure 2c) reflected a more general difficulty in performing those versions of the task.

Human simplicity preferences are robust to task demands

To better understand the optimality, variability, and generality of the simplicity preferences exhibited by our human subjects, we compared their performance to that of artificial neural networks (ANNs) trained to optimize performance on this task. We used a novel ANN architecture that we designed to perform statistical model selection, in a form applicable to the task described above (Figure 3a,b). On each trial, the network took as input two images representing the models to be compared, and a set of coordinates representing the observations on that trial. The output of the network was a decision between the two models, encoded as a softmax vector. We analyzed 50 instances of the ANN that differed only in the random initialization of their weights and in the examples seen during training, using the same logistic-regression approach we used for the human subjects.

The ANN was designed as follows. The input stage consisted of two pretrained VGG16 convolutional neural networks (CNNs), each of which took in a pictorial representation of one of the two models under consideration. VGG was chosen as a popular architecture that is often taken as a benchmark for

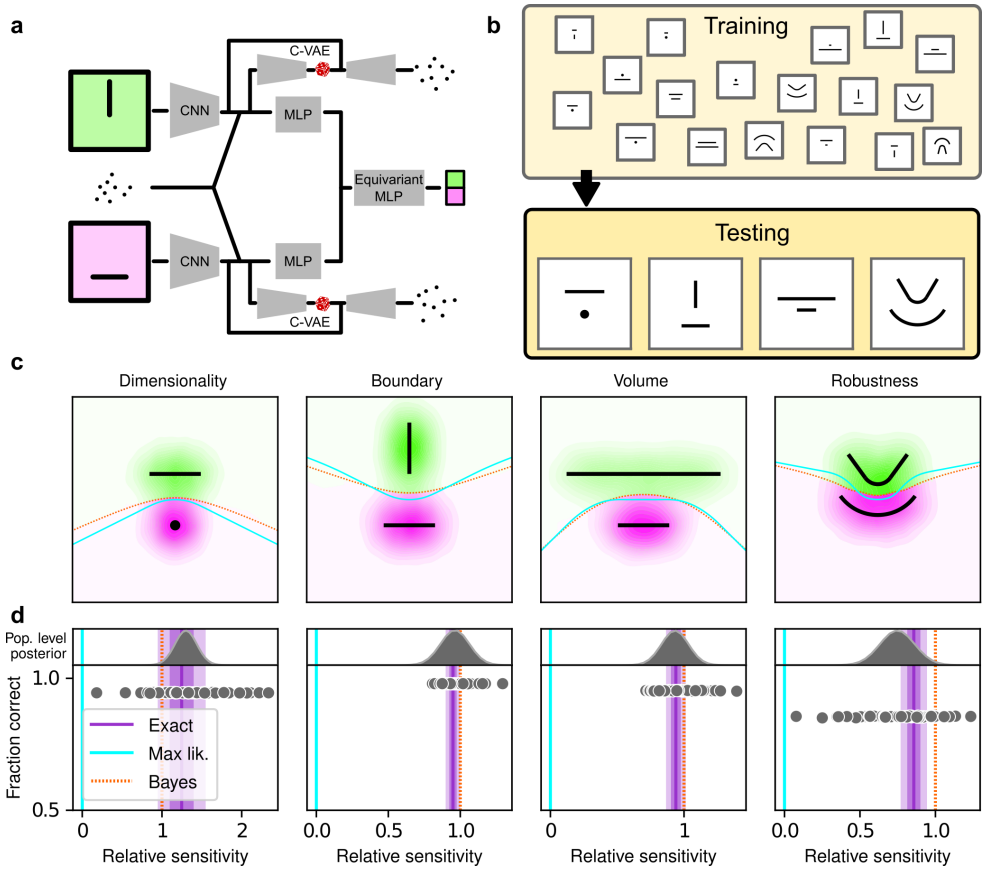

Figure 3: *Deep neural networks exhibit theoretically grounded simplicity preferences.* **a**: A novel deep neural-network architecture for statistical model selection. The network (see text and Methods for details) takes two images as input, each representing a model, and a set of 2D coordinates, each representing a datapoint. The output is a softmax-encoded choice between the two models. **b**: Each network was trained on multiple variants of the task, including systematically varied model length or curvature, then tested using the same configurations as for the human studies. **c**: Summary of network behavior, like Figure 2a. Hue (pink/green): k-nearest-neighbor interpolation of the model choice, as a function of the empirical centroid of the data. Color gradient (light/dark): marginal density of empirical data centroids for the given model pair, showing the region of space where data were more likely to fall. Cyan solid line: decision boundary for an observer that always chooses the model with highest maximum likelihood. Orange dashed line: decision boundary for an ideal Bayesian observer. **d**: Estimated relative sensitivity to geometrical features characterizing model complexity. As for the human subjects, each fitted FIA coefficient was normalized to the likelihood coefficient and thus could be interpreted as a relative sensitivity to the associated FIA term. For each term, an ideal Bayesian observer would have a relative sensitivity of one (dashed orange lines), whereas an observer that relied on only maximum-likelihood estimation (i.e., choosing"up" or "down" based on only the model that was the closest to the data) would have a relative sensitivity of zero (solid cyan lines). Top: population-level estimate (posterior distribution of population-level relative sensitivity given the experimental observations; see Methods, section A.5 for details). Bottom: each gray dot represents the task accuracy of one of 50 trained networks (y axis) versus the posterior mean estimate of the relative sensitivity for that network (x axis). Purple: relative sensitivity of an ideal observer that samples from the exact Bayesian posterior (not the approximated one provided by the FIA). Shading: posterior mean $\pm$ 1 or 2 stdev., estimated by simulating 50 such observers.

comparisons with the human visual system (*30*, *31*). The CNNs were composed by a number of convolutional layers, whose weights were kept frozen at their pretrained values, followed by three fully-connected layers, whose weights were allowed to change during training (see Methods for details). The output of the CNNs were each fed into a multilayer perceptron (MLP) consisting of linear, rectified-linear (ReLU), and batch-normalization layers. The MLP outputs were then concatenated and fed into an Equivariant MLP, which enforces equivariance of the network output under position swap of the two models through a custom parameter-sharing scheme (*32*). The network also contained two conditional variational autoencoder (C-VAE) structures, which sought to replicate the data-generation process conditioned on each model and therefore encouraged the fully connected layers upstream to learn model representations that captured task-relevant features.

After training, the ANNs performed the task substantially better than the human subjects, with higher overall accuracies that included higher likelihood sensitivities (Supplementary Text, section C.4 and Supplementary Table S1) and simplicity preferences that more closely matched the theoretically optimal values (Figure 3d). In fact, these simplicity preferences were closer to those expected from simulated observers that use the exact Bayesian model posterior rather than the FIA-approximated one, indicating an imperfect approximation of the FIA to the exact Bayesian posterior rather than suboptimal network behavior. These simplicity preferences varied slightly in magnitude across the different networks, but unlike for the human subjects this variability was relatively small (compare ranges of values in Figures 2c and 3d, plotted on different x-axis scales) and not related systematically to any differences in the generally high accuracy rates for each condition (Figure 3e; posterior mean ± st. dev. of Spearman's rho between accuracy and $\beta - 1$, where β is the sensitivity: dimensionality, -0.14±0.10; boundary, 0.08±0.11; volume, -0.12±0.11; robustness, -0.08±0.11). These results imply that the stochastic nature of the task gives rise to some variability in simplicity biases even after extensive training to optimize performance accuracy, but this source of variability cannot by itself account for the range of sensitivities (and suboptimalities) exhibited by the human subjects.

These results, combined with the fact that we did not provide trial-by-trial feedback to the subjects while they performed the task, suggest that the human simplicity preferences we measured were not simply learned optimizations for these particular task conditions but rather are a more inherent (and variable) part of how we make decisions under uncertainty. However, because we provided each subject with instructions that echoed Bayesian-like reasoning (see Methods) and a brief training set with feedback before their testing session, we cannot rule out from this dataset alone that at least some aspects of the simplicity preferences we measured from the human subjects depended on those specific instructions and training conditions. We therefore ran a second experiment to rule out this possibility. For this experiment, we used the same task variants as above but a different set of instructions and training, designed to encourage subjects to pick the model with the maximum likelihood, thus disregarding model complexity. Specifically, the visual cues were the same as in the original experiment, but the subjects were asked to report which of the two shapes on the screen was closest to the center-of-mass of the dot cloud. We ensured that the subjects recruited for this "maximum-likelihood" task had not participated in the original, "generative" task. We also trained and tested ANNs on this version of the task, using the maximum-likelihood solution as the correct answer.

Despite this major difference in instructions and training, the human subjects exhibited similar simplicity preferences on the generative and maximum-likelihood tasks, suggesting that humans have a general predilection for simplicity even without relevant instructions or incentives (Figure 4, left). Specifically, despite some quantitative differences, the distributions of relative sensitivities showed the same basic patterns for both tasks, with a general increase of relative sensitivity from volume (0.19±0.08 for the maximum-likelihood task; compare to values above), to boundary (0.89±0.10), to robustness (2.27±0.15), to dimensionality (2.29±0.41). In stark contrast to the human data and to ANNs trained on the true generative task, ANN sensitivity to model complexity on the maximum-likelihood task was close to zero for all four terms (Figure 4, right).

To summarize the similarities and differences between how humans and ANNs used simplicity biases to guide their decision-making behaviors for these tasks, and their implications for performance, Figure 5 shows overall accuracy for each set of conditions we tested. Specifically, for each network or subject, task configuration, and instruction set, we computed the percentage of correct responses with respect to both the generative task (i.e., for which theoretically optimal performance depends on simplicity biases) and the maximum-likelihood task (i.e., for which theoretically optimal performance does not depend on sim-

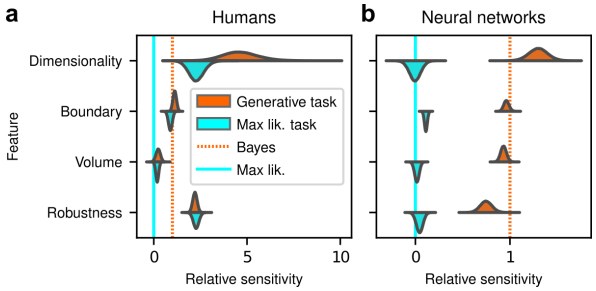

Figure 4: *Humans, but not artificial neural networks, exhibit simplicity preferences even when they are suboptimal.* **a**: Relative sensitivity of human subjects to the geometric complexity terms (population-level estimates, as in Figure 2c, top) for two task conditions: 1) the original, "generative" task where subjects were implicitly instructed to solve a model-selection problem (same data as in Figure 2c, top; cyan); and 2) a "maximum-likelihood" task variant, where subjects were instructed to report which of two models has the highest likelihood (shortest distance from the data; orange). The two task variants were tested on distinct subject pools of roughly the same size (202 subjects for the generative task, 201 for the maximum-likelihood task, in both cases divided in four groups of roughly 50 subjects each). Solid cyan lines: relative sensitivity of a maximum-likelihood observer. Orange dashed lines: relative sensitivity of an ideal Bayesian observer. **b**: Same comparison and format, but for two distinct populations of 50 deep neural networks trained on the two variants of the task (orange is the same data as in Figure 3d, top).

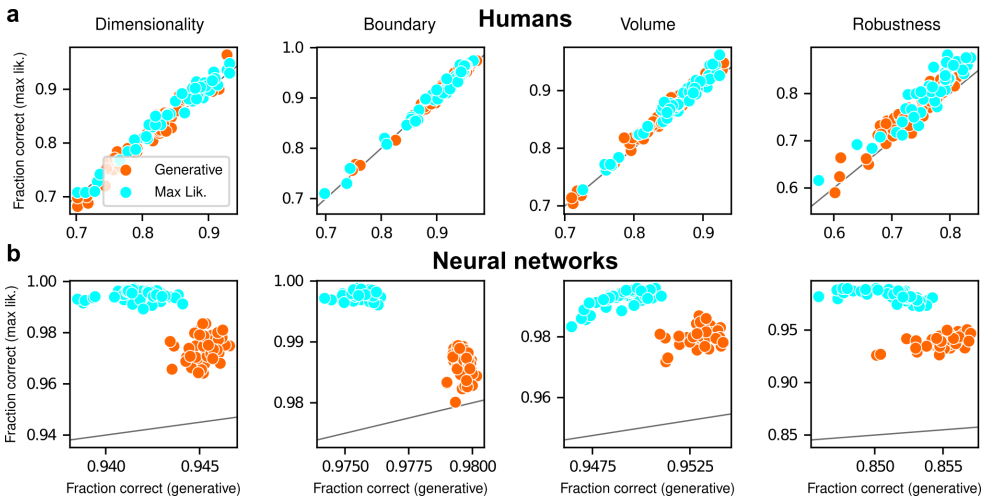

Figure 5: *Humans and artificial neural networks have different patterns of accuracy reflecting their different use of simplicity preferences.* Each panel shows accuracy with respect to maximum-likelihood solutions (i.e., the model closest to the centroid of the data; ordinate) versus with respect to generative solutions (i.e., the model that generated the data; abscissa). The gray line is the identity. Columns correspond to the four task variants associated with the four geometric complexity terms, as indicated. **a**: Data from individual human subjects (points), instructed to find the generative (orange) or maximum-likelihood (cyan) solution. Subject performance was higher when evaluated against maximum-likelihood solutions than it was when evaluated against generative solutions, for all groups of subjects (two-tailed paired t-test, generative task subjects: dimensionality, t-statistic 2.21, p-value 0.03; boundary, 6.21, 1e-7; volume, 9.57, 8e-13; robustness, 10.6, 2e-14. Maximum-likelihood task subjects: dimensionality, 5.75, 5e-7; boundary, 4.79, 2e-6; volume, 10.8, 2e-14; robustness, 12.2, 2e-16). **b**: Data from individual ANNs (points), trained on the generative (orange) or maximum-likelihood (cyan) task. Network performance was always highest when evaluated against maximum-likelihood solutions, compared to generative solutions (all dots are above the identity line).

plicity biases). Because the maximum-likelihood solutions are deterministic (they depend only on which model the data centroid is closest to, and thus there exists an optimal, sharp decision boundary that leads to perfect performance) and the generative solutions are not (they depend probabilistically on the likelihood and bias terms, so it is generally impossible to achieve perfect performance), performance on the former is expected to be higher than on the latter. Accordingly, both ANNs and (to a lesser extent) humans tended to perform better when assessed relative to maximum-likelihood solutions. Moreover, the ANNs tended to exhibit behavior that was consistent with optimization to the given task conditions: networks trained to find maximum-likelihood solutions did better than networks trained to find generative solutions for the maximum-likelihood task, and networks trained to find generative solutions did better than networks trained to find maximum-likelihood solutions for the generative task. In contrast, humans tended to adopt similar strategies regardless of the task conditions, in all cases using Bayesian-like simplicity biases.

Put briefly, ANNs exhibited simplicity preferences only when trained to do so, whereas human subjects exhibited them regardless of their instructions and training.

Discussion

Simplicity has long been regarded as a key element of effective reasoning and rational decision-making, and it has been proposed as a foundational principle in philosophy (*1*), psychology (*2*, *5*), statistical inference (*11–14*, *20*, *21*, *23*, *33*, *34*), and more recently machine learning (*35–38*). Accordingly, multiple studies have identified preferences for simplicity in human cognition (*6*, *8*, *9*), such as a tendency to prefer smoother (simpler) curves as the inferred, latent source of noisy observed data (*7*, *10*). However, the quantitative form and magnitude of this preference have never been identified. In this work, we showed that the simplicity preference is closely related to a specific mathematical formulation of Occam's razor, situated at the convergence of Bayesian model selection and information theory (*11*). This formulation enabled us to go beyond the mere detection of a preference for simple explanations for data and to measure precisely the strength of this preference in artificial and human subjects under a variety of theoretically motivated conditions.

Our study makes several novel contributions. The first is theoretical: we derived a new term of the Fisher Information Approximation (FIA) in Bayesian model selection that accounts for the possibility that the best model is on the boundary of the model family. This boundary term is important because it can account for the possibility that, because of the noise in the data, the best value of one parameter (or of a combination of parameters) takes on an extreme value. This condition is related to the phenomenon of "parameter evaporation" that is common in real-world models for data (*39*). Moreover, boundaries for parameters are particularly important for studies of perceptual decision-making, in which sensory stimuli are limited by the physical constraints of the experimental setup and thus reasoning about unbounded parameters would be problematic for subjects. For example, imagine designing an experiment that requires subjects to report the location of a visual stimulus. In this case, an unbounded set of possible locations (e.g., along a line that stretches infinitely far in the distance to the left and to the right) is clearly untenable. Our "boundary" term formalizes the impact of considering the set of possibilities as having boundaries, which tend to increase local complexity because they tend to reduce the number of local hypotheses close to the data (see Figure 1b).

The second contribution of this work relates to ANNs: these networks can learn to use or ignore the simplicity preferences in an optimal way (i.e., according to the magnitudes prescribed by the theory), depending on how they are trained. These results are different from, and complementary to, recent work that has focused on the idea that implementation of simple functions could be key to generalization in deep neural networks (*35–38*). Here we have shown that effective learning can take into account the complexity of the hypothesis space, rather than that of the decision function, in producing normative simplicity preferences. On the one hand, these results do not seem surprising, because ANNs, and deep networks in particular, are powerful function approximators that perform well in practice on a vast range of inference tasks(*40*). Accordingly, our ANNs trained with respect to the true generative solutions were able to make effective decisions, including simplicity preferences, about the generative source of a given set of observations. Likewise, our ANNs trained with respect to maximum-likelihood solutions were able to make effective decisions, without simplicity preferences, about the maximum-likelihood match for a given set of observations. On the other hand, these results provide new insights into how ANNs might be analyzed to better understand the kinds of solutions they produce for particular problems. In particular, assessing the presence or absence of these kinds of

simplicity preferences might help identify if and/or how well an ANN is likely to avoid overfitting to training data and provide more generalizable solutions.

The third, and most important, contribution of this work relates to human behavior: people tend to use simplicity preferences when making decisions, and unlike ANNs these preferences do not seem to be simply the consequences of learning specific task demands but rather an inherent part of how we interpret uncertain information. This tendency has important implications for the kinds of computations our brains must use to solve these kinds of tasks, and how those computations appear to differ from those implemented by the ANNs we used. From a theoretical perspective, the difference between a Bayesian solution (i.e., one that includes the simplicity preferences) and a maximum-likelihood solution (i.e., one that does not include the simplicity preferences) to these tasks is that the latter considers only the single best-fitting model from each family, whereas the former integrates over all possible models in each family. Our finding that ANNs can converge on either solution when trained appropriately indicates that both are, in principle, learnable. However, our finding that people tend to use the Bayesian solution even when instructed to use the maximum-likelihood solution suggests that we naturally do not make decisions based simply on the single best or archetypical instance within a family of possibilities but rather integrate across that family. Put more concretely in terms of our task, when told to identify the shape closest to the data points, subjects were likely uncertain about which exact location on each shape was closest and thus integrated over the possibilities — thus inducing simplicity preferences as prescribed by the Bayesian solution. These findings will help motivate and inform future studies to identify where and how the brain implements and stores these integrated solutions to relevant decision problems.

Another key feature of our findings that merits further study is the magnitude and variability of preferences exhibited by the human subjects. On average, human sensitivity to each geometrical model feature was: 1) larger than zero, 2) at least slightly different from the optimal value (e.g., larger for dimensionality and robustness, smaller for volume), 3) different for distinct features and different subjects; and 4) independent of instructions and training. What is the source of this diversity? One hypothesis is that people may weigh more heavily the model features that are easier or cheaper to compute. In our experiments, the most heavily weighted feature was model dimensionality. In our mathematical framework, this feature corresponds to the number of degrees of freedom of a possible explanation for the observed data and thus can be relatively easy to assess. By contrast, the least heavily weighted feature was model volume. This feature involves integrating over the whole model family (to count how many distinct states of the world can be explained by a certain hypothesis, one needs to enumerate them) and thus can be very difficult to compute. The other two terms, boundary and robustness, are intermediate in terms of human weighting and computational difficulty: they are harder to compute than dimensionality, because they depend on the data and on the properties of the model at the maximum likelihood location, but are also simpler than the volume term, because they are local quantities that do not require integration over the whole model manifold. This intuition leads to new questions about the relationship between the complexity of the explanations being compared and the complexity of the decision-making process itself, calling into question notions of bounded rationality and diminishing returns in optimal inference (*41*, *42*). Answering such questions is beyond the scope of the present work but merits further study.

Another potentially intriguing future direction is a comparison with other formal approaches to the emergence of simplicity that can lead to different predictions. Recent studies have argued that Jeffrey's prior (upon which our geometric approach is based) could give an incomplete picture of the complexity of a class of models that occur commonly in the natural sciences, which contain many combinations of parameters that do not affect model behavior, and proposed instead the use of data-dependent priors (*43*, *44*). The two methods lead to different results, especially in the data-limited regime (*45*). It would be useful to understand the relevance of these differences to human and machine decision-making.

In summary, our work reveals the direct, quantitative relevance of formal notions of model complexity for human behavior. By relying on a combination of theoretical advances, computational modeling and behavioral experiments, we have established a novel set of normative reference points for decision making under uncertainty. Our findings therefore open up a new arena in which human cognition could be measured against optimal inferential processes, potentially leading to new insights into the constraints affecting information processing in the brain.

Data availability. All experimental data collected in this work is available at

doi:10.17605/OSF.IO/R6D8N.

Code availability. All data and code needed to reproduce the experiments (including running the online psychophysics tasks and training and testing the neural networks), and to analyze the data and produce all figures is available at doi:10.17605/OSF.IO/R6D8N.

Ethics. Human subject protocols were approved and determined to be Exempt by the University of Pennsylvania Internal Review Board (IRB protocol 844474). Subjects provided written consent on-line before they began the task.

Acknowledgments. We thank Kamesh Krishnamurthy for discussions, and acknowledge the financial support of R01 NS113241 (EP), R01 EB026945 (JIG and VB), as well as a hardware grant from the NVIDIA Corporation (EP). The HPC Collaboration Agreement between SISSA and CINECA granted access to the Marconi100 cluster.

Author contribution. Conceptualization: EP VB JG. Methodology: EP SL PC VB JG. Software: EP SL. Formal analysis: EP SL. Investigation: EP SL. Resources: EP VB JG. Data curation: EP SL. Writing - original draft: EP JG. Writing - editing and reviewing: EP SL PC VB JG. Supervision: VB JG. Project administration: JG. Funding acquisition: VB JG.

References

1. A. Baker, in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta (Metaphysics Research Lab, Stanford University, Summer 2022, 2022), (2022; https://plato.stanford.edu/archives/sum2022/entries/simplicity/).
2. J. Feldman, The Simplicity Principle in Perception and Cognition: The Simplicity Principle. *Wiley Interdisciplinary Reviews: Cognitive Science* **7**, 330–340, ISSN: 19395078, (2022; https://onlinelibrary.wiley.com/doi/10.1002/wcs.1406) (Sept. 2016).
3. K. Koffka, *Principles of Gestalt Psychology* (Mimesis international, Milano [etc., 2014), ISBN: 978-88-575-2393-4.
4. G. Hatfield, The Status of the Minimum Principle in the Theoretical Analysis of Visual Perception. *Psychological Bulletin* **97**, 155, ISSN: 1939-1455, (2022; https://psycnet.apa.org/fulltext/1985-19203-001.html) (Aug. 1, 1985).
5. N. Chater, P. Vitányi, Simplicity: A Unifying Principle in Cognitive Science? *Trends in Cognitive Sciences* **7**, 19–22, ISSN: 1364-6613, (2022; https://www.sciencedirect.com/science/article/pii/S1364661302000050) (Jan. 1, 2003).
6. E. M. Pothos, N. Chater, A Simplicity Principle in Unsupervised Human Categorization. *Cognitive Science* **26**, 303–343, ISSN: 1551-6709, (2022; https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog2603_6) (2002).
7. T. Genewein, D. A. Braun, Occam's Razor in Sensorimotor Learning. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20132952 (May 2014).
8. S. Gershman, Y. Niv, Perceptual Estimation Obeys Occam's Razor. *Frontiers in Psychology* **4**, ISSN: 1664-1078, (2022; https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00623) (2013).
9. D. R. B. Little, R. Shiffrin, Simplicity Bias in the Estimation of Causal Functions. *Proceedings of the Annual Meeting of the Cognitive Science Society* **31**, (2022; https://escholarship.org/uc/item/3d85q7zt) (2009).
10. S. Johnson, A. Jin, F. Keil, presented at the Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36).
11. V. Balasubramanian, Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions. *Neural Computation* **9**, 349–368 (1997).
12. D. J. C. MacKay, Bayesian Interpolation. *Neural Computation* **4**, 415–447 (May 1992).
13. H. Jeffreys, *Theory of Probability* (Clarendon Press, 1939), 380 pp.
14. S. F. Gull, in *Maximum-Entropy and Bayesian Methods in Science and Engineering: Foundations*, ed. by G. J. Erickson, C. R. Smith (Springer Netherlands, Dordrecht, 1988), pp. 53–74, ISBN: 978-94-009-3049-0, (2022; https://doi.org/10.1007/978-94-009-3049-0_4).

15. E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Apr. 1, 2003), 753 pp., ISBN: 0-521-59271-2.

16. G. Schwarz, Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461–464, ISSN: 0090-5364, 2168-8966, (2022; https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-2/Estimating-the-Dimension-of-a-Model/10.1214/aos/1176344136.full) (Mar. 1978).

17. A. A. Neath, J. E. Cavanaugh, The Bayesian Information Criterion: Background, Derivation, and Applications. *WIREs Computational Statistics* **4**, 199–203, ISSN: 1939-0068, (2022; https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.199) (2012).

18. I. J. Myung, V. Balasubramanian, M. A. Pitt, Counting Probability Distributions: Differential Geometry and Model Selection. *Proceedings of the National Academy of Sciences* **97**, 11170–11175, (2023; https://www.pnas.org/doi/abs/10.1073/pnas.170283897) (Oct. 10, 2000).

19. J. Rissanen, Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory* **42**, 40–47, ISSN: 1557-9654 (Jan. 1996).

20. P. D. Grünwald, *The Minimum Description Length Principle* (MIT press, 2007), ISBN: 978-0-262-07281-6.

21. C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length* (Springer, New York, 2005), 429 pp., ISBN: 978-0-387-23795-4.

22. W. Bialek, I. Nemenman, N. Tishby, Predictability, Complexity and Learning. *Neural Computation*, 2409–2463 (2001).

23. A. D. Lanterman, Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection. *International Statistical Review* **69**, 185–212, ISSN: 1751-5823, (2023; https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2001.tb00456.x) (2001).

24. E. Piasini, V. Balasubramanian, J. I. Gold, *Preregistration Document* (https://doi.org/10.17605/OSF.IO/2X9H6).

25. E. Piasini, V. Balasubramanian, J. I. Gold, *Preregistration Document Addendum* (https://doi.org/10.17605/OSF.IO/5HDQZ).

26. E. Piasini, S. Liu, V. Balasubramanian, J. I. Gold, *Preregistration Document Addendum* (https://doi.org/10.17605/OSF.IO/826JV).

27. R. McElreath, *Statistical Rethinking* (CRC Press, 2016), ISBN: 978-1-4822-5344-3.

28. J. K. Kruschke, *Doing Bayesian Data Analysis* (Academic Press, ed. 2, 2015), ISBN: 978-0-12-405888-0.

29. A. Gelman *et al.*, *Bayesian Data Analysis* (CRC Press, ed. 3, 2014), ISBN: 978-1-4398-4095-5.

30. M. Schrimpf *et al.*, Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like? *bioRxiv*, 407007, (http://biorxiv.org/content/early/2020/01/02/407007.abstract) (Jan. 1, 2020).

31. P. Muratore, S. Tafazoli, E. Piasini, A. Laio, D. Zoccolan, presented at the Advances in Neural Information Processing Systems 22, arXiv: 2205.13816 (cs, q-bio), (2022; http://arxiv.org/abs/2205.13816).

32. S. Ravanbakhsh, J. Schneider, B. Póczos, presented at the Proceedings of the 34th International Conference on Machine Learning, International Conference on Machine Learning, pp. 2892–2901, (2023; https://proceedings.mlr.press/v70/ravanbakhsh17a.html).

33. C. de Mulatier, P. P. Mazza, M. Marsili, "Statistical Inference of Minimally Complex Models," Sept. 27, 2021, arXiv: 2008.00520 (physics, q-bio, stat), (2022; http://arxiv.org/abs/2008.00520).

34. R. Xie, M. Marsili, "Occam Learning," Oct. 24, 2022, arXiv: 2210.13179 (cond-mat), (2022; http://arxiv.org/abs/2210.13179).

35. G. De Palma, B. Kiani, S. Lloyd, presented at the Advances in Neural Information Processing Systems, ed. by H. Wallach *et al.*, vol. 32, (https://proceedings.neurips.cc/paper/2019/file/feab05aa91085b7a8012516bc3533958-Paper.pdf).

36. G. Valle-Perez, C. Q. Camargo, A. A. Louis, presented at the International Conference on Learning Representations.

37. P. Chaudhari *et al.*, Entropy-SGD: Biasing Gradient Descent into Wide Valleys*. *Journal of Statistical Mechanics: Theory and Experiment* **2019**, 124018, ISSN: 1742-5468, (2023; https://dx.doi.org/10.1088/1742-5468/ab39d9) (Dec. 2019).

38. R. Yang, J. Mao, P. Chaudhari, presented at the Proceedings of the 39th International Conference on Machine Learning, International Conference on Machine Learning, pp. 25166–25197, (2023; https://proceedings.mlr.press/v162/yang22k.html).

39. M. K. Transtrum *et al.*, Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. *The Journal of Chemical Physics* **143**, 010901, ISSN: 0021-9606, (2022; https://aip.scitation.org/doi/10.1063/1.4923066) (July 7, 2015).

40. Y. Bengio, Y. Lecun, G. Hinton, Deep Learning for AI. *Communications of the ACM* **64**, 58–65, ISSN: 0001-0782, (2022; https://doi.org/10.1145/3448250) (June 21, 2021).

41. G. Tavoni, V. Balasubramanian, J. I. Gold, What Is Optimal in Optimal Inference? *Current Opinion in Behavioral Sciences*, Artificial Intelligence **29**, 117–126, ISSN: 2352-1546, (2022; https://www.sciencedirect.com/science/article/pii/S235215461930066X) (Oct. 1, 2019).

42. G. Tavoni, T. Doi, C. Pizzica, V. Balasubramanian, J. I. Gold, Human Inference Reflects a Normative Balance of Complexity and Accuracy. *Nature Human Behaviour* **6**, 1153–1168, ISSN: 2397-3374, (2022; http://www.nature.com/articles/s41562-022-01357-z) (8 Aug. 2022).

43. H. H. Mattingly, M. K. Transtrum, M. C. Abbott, B. B. Machta, Maximizing the Information Learned from Finite Data Selects a Simple Model. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 1760–1765, ISSN: 1091-6490, pmid: 29434042 (Feb. 2018).

44. K. N. Quinn, M. C. Abbott, M. K. Transtrum, B. B. Machta, J. P. Sethna, *Information Geometry for Multiparameter Models: New Perspectives on the Origin of Simplicity*, Sept. 22, 2022, arXiv: 2111.07176 (cond-mat, physics:physics), (2022; http://arxiv.org/abs/2111.07176).

45. M. C. Abbott, B. B. Machta, "Far from Asymptopia," May 6, 2022, arXiv: 2205.03343 (physics, stat), (2022; http://arxiv.org/abs/2205.03343).

Supplementary Materials

A Methods **14**
A.1 Derivation of the boundary term in the Fisher Information Approximation 14
A.2 Behavioral experiments with human subjects 20
A.3 Detailed model definitions and computation of FIA terms 21
A.4 Numerical experiments with artificial neural networks 27
A.5 Experimental data analysis 29

B Supplementary Figures and Tables **32**

C Supplementary Text **41**
C.1 Numerical comparison of the extended FIA vs exact Bayes 41
C.2 Subject demographics 41
C.3 Posterior predictive checks 41
C.4 Details on raw estimated sensitivities 41
C.5 Uncertainty in subject-level sensitivities 42
C.6 Lapse-rate analysis 42
C.7 Outcome of significance tests specified in the preregistration documents 42

References **43**

A Methods

A.1 Derivation of the boundary term in the Fisher Information Approximation

Here we generalize the derivation of the Fisher Information Approximation given by Balasubramanian (*11*) to the case where the maximum-likelihood solution for a model lies on the boundary of the parameter space. Apart from the more general assumptions, the following derivation follows closely the original one, with some minor notational changes. This derivation appeared in preliminary form in (*46*).

A.1.1 Set-up and hypotheses

The problem we consider here is that of selecting between two models (say $\mathcal{M}_1$ and $\mathcal{M}_2$), after observing empirical data $X = \{x_i\}_{i=1}^N$. N is the sample size and $\mathcal{M}_1$ is assumed to have d parameters, collectively indexed as ϑ taking values in a compact domain Θ. As a prior over ϑ we take Jeffrey's prior:

$$w(\vartheta) = \frac{\sqrt{\det g(\vartheta)}}{\int \mathrm{d}^d\vartheta \sqrt{\det g(\vartheta)}} \tag{1}$$

where g is the (expected) Fisher Information of the model $\mathcal{M}_1$:

$$g_{\mu\nu}(\vartheta) = \mathbb{E}\left[-\frac{\partial^2 \ln p(x|\vartheta)}{\partial\vartheta^\mu \partial\vartheta^\nu}\right]_\vartheta \tag{2}$$

The Bayesian posterior

$$\mathbb{P}(\mathcal{M}_1|X) = \frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(X)} \int \mathrm{d}^d w(\vartheta)\mathbb{P}(X|\vartheta) \tag{3}$$

then becomes, after assuming a flat prior over models and dropping irrelevant terms:

$$\mathbb{P}(\mathcal{M}_1|X) = \frac{\int_\Theta \mathrm{d}^d\vartheta \sqrt{\det g}\, \exp\left[-N(-\frac{1}{N}\ln \mathbb{P}(X|\vartheta))\right]}{\int \mathrm{d}^d\vartheta \sqrt{\det g}} \tag{4}$$

Just as in (*11*), we now make a number of regularity assumptions: 1. $\ln \mathbb{P}(X|\vartheta)$ is smooth; 2. there is a unique global minimum $\hat{\vartheta}$ for $\ln \mathbb{P}(X|\vartheta)$; 3. $g_{\mu\nu}(\vartheta)$ is smooth; 4. $g_{\mu\nu}(\hat{\vartheta})$ is positive definite; 5. $\Theta \subset \mathbb{R}^d$ is compact; and 6. the values of the local minima of $\ln \mathbb{P}(X|\vartheta)$ are bounded away from the global minimum by some $\epsilon > 0$. Importantly, unlike in (*11*), we do not assume that $\hat{\vartheta}$ is in the interior of Θ.

The shape of Θ. Because we are specifically interested in understanding what happens at a boundary of the parameter space, we add a further assumption that, while being not very restrictive in spirit, allows us to derive a particularly interpretable result. In particular, we assume that Θ is specified by a single linear constraint of the form:

$$D_\mu \vartheta^\mu + d \geq 0 \tag{5}$$

Without loss of generality, we also take the constraint to be expressed in Hessian normal form, namely, $\|D_\mu\| = 1$. For clarity, note this assumption on the shape of Θ is used only from subsubsection A.1.3 onward.

A.1.2 Preliminaries

We now proceed to set up a low-temperature expansion of Equation 4 around the saddle point $\hat{\vartheta}$. We start by rewriting the numerator in Equation 4 as:

$$\int_\Theta \mathrm{d}^d\vartheta \exp\left[-N\left(-\frac{1}{2N}\ln\det g - \frac{1}{N}\ln\mathbb{P}(X|\vartheta)\right)\right] \tag{6}$$

The idea of the Fisher Information Approximation is to expand the integrand in Equation 6 in powers of N around the maximum likelihood point $\hat{\vartheta}$. To this end, we define three useful objects:

$$\tilde{I}_{\mu_1\cdots\mu_i} := -\frac{1}{N}\nabla_{\mu_1}\cdots\nabla_{\mu_i}\ln\mathbb{P}(X|\vartheta)\Big|_{\hat{\vartheta}} = -\frac{1}{N}\sum_{j=1}^{N}\nabla_{\mu_1}\cdots\nabla_{\mu_i}\ln\mathbb{P}(x_j|\vartheta)\Big|_{\hat{\vartheta}}$$

$$F_{\mu_1\cdots\mu_i} := \nabla_{\mu_1}\cdots\nabla_{\mu_i}\ln\det g(\vartheta)\Big|_{\hat{\vartheta}}$$

$$\psi := -\frac{1}{2N}\ln\det g - \frac{1}{N}\ln\mathbb{P}(X|\vartheta)$$

We immediately note that:

$$\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\Big|_{\hat{\vartheta}} = \tilde{I}_{\mu_1\cdots\mu_i} - \frac{1}{2N}F_{\mu_1\cdots\mu_i}$$

which is useful to compute

$$\begin{aligned}\psi(\vartheta) &= \psi(\hat{\vartheta}) + \nabla_\mu\psi\Big|_{\hat{\vartheta}}(\vartheta^\mu - \hat{\vartheta}^\mu) + \frac{1}{2}\nabla_\mu\nabla_\nu\psi\Big|_{\hat{\vartheta}}(\vartheta^\mu - \hat{\vartheta}^\mu)(\vartheta^\nu - \hat{\vartheta}^\nu) + \dots \\ &= \sum_{i=0}^{\infty}\frac{1}{i!}\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\Big|_{\hat{\vartheta}}(\vartheta^{\mu_1} - \hat{\vartheta}^{\mu_1})\cdots(\vartheta^{\mu_i} - \hat{\vartheta}^{\mu_i}) \\ &= \sum_{i=0}^{\infty}\frac{1}{i!}\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\Big|_{\hat{\vartheta}}\prod_{k=1}^{i}(\vartheta^{\mu_k} - \hat{\vartheta}^{\mu_k})\end{aligned}$$

It is also useful to center the integration variables by introducing

$$\phi := \sqrt{N}(\vartheta - \hat{\vartheta}) \tag{7}$$

$$\mathrm{d}^d\phi = N^{d/2}\,\mathrm{d}^d\vartheta \tag{8}$$

so that

$$\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\Big|_{\hat{\vartheta}}\prod_{k=1}^{i}(\vartheta^{\mu_k} - \hat{\vartheta}^{\mu_k}) = N^{-i/2}\left(\tilde{I}_{\mu_1\cdots\mu_i} - \frac{1}{2N}F_{\mu_1\cdots\mu_i}\right)\phi^{\mu_1}\cdots\phi^{\mu_i} \tag{9}$$

and Equation 6 becomes:

$$
\begin{aligned}
\int \mathrm{d}^d\vartheta \exp[-N\psi] &= N^{-d/2}\int \mathrm{d}^d\phi \exp\left[-N\sum_{i=0}^{\infty}\frac{1}{i!}N^{-i/2}\left(\tilde{I}_{\mu_1\cdots\mu_i}-\frac{1}{2N}F_{\mu_1\cdots\mu_i}\right)\phi^{\mu_1}\cdots\phi^{\mu_i}\right] \\
&= N^{-d/2}\int \mathrm{d}^d\phi \exp\left\{-N\left(-\frac{1}{N}\ln\mathbb{P}(X|\hat{\vartheta})-\frac{1}{2N}\ln\det g(\hat{\vartheta})\right)+\right. \\
&\quad \left.-N\left[\sum_{i=1}^{\infty}\frac{1}{i!}N^{-i/2}\left(\tilde{I}_{\mu_1\cdots\mu_i}-\frac{1}{2N}F_{\mu_1\cdots\mu_i}\right)\phi^{\mu_1}\cdots\phi^{\mu_i}\right]\right\} \\
&= N^{-\frac{d}{2}}\exp\left[-\left(-\ln\mathbb{P}(X|\hat{\vartheta})-\frac{1}{2}\ln\det g(\hat{\vartheta})\right)\right]\times \\
&\quad \times\int \mathrm{d}^d\phi \exp\left\{-N\left[\frac{1}{\sqrt{N}}\tilde{I}_\mu\phi^\mu+\frac{1}{2N}\tilde{I}_{\mu\nu}\phi^\mu\phi^\nu+\right.\right. \\
&\quad \left.\left.+\frac{1}{N}\sum_{i=1}^{\infty}N^{-\frac{i}{2}}\left(\frac{1}{(i+2)!}\tilde{I}_{\mu_1\cdots\mu_{i+2}}\phi^{\mu_1}\cdots\phi^{\mu_{i+2}}-\frac{1}{2i!}F_{\mu_1\cdots\mu_i}\phi^{\mu_1}\cdots\phi^{\mu_i}\right)\right]\right\}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}(\mathcal{M}_1|X) &= N^{-\frac{d}{2}}\exp\left[-\left(-\ln\mathbb{P}(X|\hat{\vartheta})-\frac{1}{2}\ln\det g(\hat{\vartheta})+\ln\int \mathrm{d}^d\vartheta\sqrt{\det g}\right)\right]\times \\
&\quad \times\int \mathrm{d}^d\phi \exp\left[-\sqrt{N}\tilde{I}_\mu\phi^\mu-\frac{1}{2}\tilde{I}_{\mu\nu}\phi^\mu\phi^\nu+\right. \\
&\quad \left.-\sum_{i=1}^{\infty}N^{-\frac{i}{2}}\left(\frac{1}{(i+2)!}\tilde{I}_{\mu_1\cdots\mu_{i+2}}\phi^{\mu_1}\cdots\phi^{\mu_{i+2}}-\frac{1}{2i!}F_{\mu_1\cdots\mu_i}\phi^{\mu_1}\cdots\phi^{\mu_i}\right)\right]\Bigg\} \\
&= N^{-\frac{d}{2}}\exp\left[-\left(-\ln\mathbb{P}(X|\hat{\vartheta})-\frac{1}{2}\ln\det g(\hat{\vartheta})+\ln\int_\Theta \mathrm{d}^d\vartheta\sqrt{\det g}\right)\right]\cdot Q
\end{aligned}
\tag{10}
$$

where

$$
Q=\int_\Phi \mathrm{d}^d\phi \exp\left[-\sqrt{N}\tilde{I}_\mu\phi^\mu-\frac{1}{2}\tilde{I}_{\mu\nu}\phi^\mu\phi^\nu-G(\phi)\right] \tag{11}
$$

and

$$
G(\phi)=\sum_{i=1}^{\infty}N^{-\frac{i}{2}}\left(\frac{1}{(i+2)!}\tilde{I}_{\mu_1\cdots\mu_{i+2}}\phi^{\mu_1}\cdots\phi^{\mu_{i+2}}-\frac{1}{2i!}F_{\mu_1\cdots\mu_i}\phi^{\mu_1}\cdots\phi^{\mu_i}\right) \tag{12}
$$

where $G(\phi)$ collects the terms that are suppressed by powers of N.

Our problem has been now reduced to computing Q by performing the integral in Equation 11. Now our assumptions come into play for the key approximation step. For the sake of simplicity, assuming that N is large we drop $G(\phi)$ from the expression above, so that Q becomes a simple Gaussian integral with a linear term:

$$
Q=\int_\Phi \mathrm{d}^d\phi \exp\left[-\sqrt{N}\tilde{I}_\mu\phi^\mu-\frac{1}{2}\phi^\mu\tilde{I}_{\mu\nu}\phi^\nu\right] \tag{13}
$$

A.1.3 Choosing a good system of coordinates

Consider now the Observed Fisher Information at the maximum likelihood, $\tilde{I}_{\mu\nu}$. As long as it is not singular, we can define its inverse $\Delta^{\mu\nu}=(\tilde{I}_{\mu\nu})^{-1}$. If $\tilde{I}_{\mu\nu}$ is positive definite, then the matrix representation of $\tilde{I}_{\mu\nu}$ has a set of d positive eigenvalues, which we denote by $\{\sigma_{(1)}^{-2},\sigma_{(2)}^{-2},\ldots,\sigma_{(d)}^{-2}\}$. The matrix representation of $\Delta^{\mu\nu}$ has eigenvalues $\{\sigma_{(1)}^{2},\sigma_{(2)}^{2},\ldots,\sigma_{(d)}^{2}\}$, and is diagonal in the same choice of coordinates as $\tilde{I}_{\mu\nu}$. We denote by

U the (orthogonal) diagonalizing matrix; i.e., U is such that

$$U\Delta U^T = \begin{bmatrix} \sigma^2_{(1)} & 0 & \cdots & 0 \\ 0 & \sigma^2_{(2)} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2_{(d)} \end{bmatrix} \quad , \quad U^T U = UU^T = \mathbb{I} \tag{14}$$

We define also the matrix K as the product of the diagonal matrix with elements $1/\sigma_{(k)}$ along the diagonal and U:

$$K = \begin{bmatrix} 1/\sigma_{(1)} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{(2)} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1/\sigma_{(d)} \end{bmatrix} U \tag{15}$$

Note that

$$\det K = (\det \Delta^{\mu\nu})^{-1/2} = \sqrt{\det \tilde{I}_{\mu\nu}}$$

and that K corresponds to a sphering transformation, in the sense that

$$K\Delta K^T = \mathbb{I} \quad \text{or} \quad K^{\mu}{}_{\kappa}\Delta^{\kappa\lambda}K^{\nu}{}_{\lambda} = \delta^{\mu\nu} \tag{16}$$

and therefore, if we define the inverse

$$P = K^{-1}$$

we have

$$P^T(\tilde{I}_{\mu\nu})P = \mathbb{I} \quad \text{or} \quad P^{\kappa}{}_{\mu}\tilde{I}_{\kappa\lambda}P^{\lambda}{}_{\nu} = \delta_{\mu\nu} \tag{17}$$

We can now define a new set of coordinates by centering and sphering, as follows:

$$\xi^{\mu} = K^{\mu}{}_{\nu}\left(\phi^{\nu} + \sqrt{N}\Delta^{\nu\kappa}\tilde{I}_{\kappa}\right) \tag{18}$$

Then,

$$\mathrm{d}^d\xi = \sqrt{\det \tilde{I}_{\mu\nu}}\,\mathrm{d}^d\phi \tag{19}$$

and

$$\phi^{\mu} = P^{\mu}{}_{\nu}\xi^{\nu} - \sqrt{N}\Delta^{\mu\nu}\tilde{I}_{\nu} \tag{20}$$

In this new set of coordinates,

$$\begin{aligned} -\sqrt{N}\tilde{I}_{\nu}\phi^{\nu} - \frac{1}{2}\phi^{\mu}\tilde{I}_{\mu\nu}\phi^{\nu} = \\ &= -\left(\sqrt{N}\tilde{I}_{\nu} + \frac{1}{2}\phi^{\mu}\tilde{I}_{\mu\nu}\right)\phi^{\nu} \\ &= -\left(\sqrt{N}\tilde{I}_{\nu} + \frac{1}{2}P^{\mu}{}_{\kappa}\xi^{\kappa}\tilde{I}_{\mu\nu}\frac{1}{2}\sqrt{N}\Delta^{\mu\kappa}\tilde{I}_{\kappa}\tilde{I}_{\mu\nu}\right)\phi^{\nu} \\ &= -\sqrt{N}\tilde{I}_{\nu}P^{\nu}{}_{\lambda}\xi^{\lambda} + N\Delta^{\nu\lambda}\tilde{I}_{\lambda}\tilde{I}_{\nu} - \frac{1}{2}P^{\mu}_{\kappa}\xi^{\kappa}\tilde{I}_{\mu\nu}P^{\nu}{}_{\lambda}\xi^{\lambda} + \frac{\sqrt{N}}{2}P^{\mu}{}_{\kappa}\xi^{\kappa}\tilde{I}_{\mu\nu}\Delta^{\nu\lambda}\tilde{I}_{\lambda} + \\ &\quad + \frac{\sqrt{N}}{2}\Delta^{\mu\kappa}\tilde{I}_{\kappa}\tilde{I}_{\mu\nu}P^{\nu}{}_{\lambda}\xi^{\lambda} - \frac{N}{2}\Delta^{\mu\kappa}\tilde{I}_{\kappa}\tilde{I}_{\mu\nu}\Delta^{\nu\lambda}\tilde{I}_{\lambda} \\ &= \frac{N}{2}\tilde{I}_{\nu}\Delta^{\nu\lambda}\tilde{I}_{\lambda} - \frac{1}{2}\xi^{\kappa}\delta_{\kappa\lambda}\xi^{\lambda} \end{aligned} \tag{21}$$

where we have used Equation 17 as well as the fact that $\Delta^{\mu\nu} = \Delta^{\nu\mu}$ and that $\Delta^{\mu\kappa}\tilde{I}_{\kappa\nu} = \delta^{\mu}{}_{\nu}$ by definition.

Therefore, putting Equation 19 and Equation 21 together, Equation 13 becomes

$$Q = \frac{\exp\left[\frac{N}{2}\tilde{I}_{\mu}\Delta^{\mu\nu}\tilde{I}_{\nu}\right]}{\sqrt{\det \tilde{I}_{\mu\nu}}}\int_{\Xi}\mathrm{d}^d\xi\,\exp\left[-\frac{1}{2}\xi_{\mu}\delta^{\mu\nu}\xi_{\nu}\right] \tag{22}$$

The problem is reduced to a (truncated) spherical gaussian integral, where the domain of integration Ξ will depend on the original domain Θ but also on $\tilde{I}_{\mu}$, $\tilde{I}_{\mu\nu}$ and $\hat{\vartheta}$. To complete the calculation, we now need to make this dependence explicit.

A.1.4 Determining the domain of integration

We start by combining Equation 7 and Equation 20 to yield:

$$\vartheta^\mu = \frac{1}{\sqrt{N}} P^\mu{}_\nu \xi^\nu - \Delta^{\mu\nu}\tilde{I}_\nu + \hat{\vartheta}^\mu \tag{23}$$

By substituting Equation 23 into Equation 5 we get

$$D_\mu \left(\frac{P^\mu{}_\nu \xi^\nu}{\sqrt{N}} - \Delta^{\mu\nu}\tilde{I}_\nu + \hat{\vartheta}^\mu \right) + d \geq 0$$

which we can rewrite as

$$\tilde{D}_\mu \xi^\mu + \tilde{d} \geq 0 \tag{24}$$

with

$$\tilde{D}_\mu := \frac{1}{\sqrt{N}} D_\nu P^\nu{}_\mu \tag{25}$$

and

$$\begin{aligned} \tilde{d} &:= d + D_\mu \hat{\vartheta}^\mu - D_\mu \Delta^{\mu\nu}\tilde{I}_\nu \\ &= d + D_\mu \hat{\vartheta}^\mu - \langle D_\mu, \tilde{I}_\mu \rangle_\Delta \end{aligned} \tag{26}$$

where by $\langle \cdot, \cdot \rangle_\Delta$ we mean the inner product in the inverse observed Fisher information metric. Now, note that whenever $\tilde{I}_\mu$ is not zero, it will be parallel to D_μ. Indeed, by construction of the maximum-likelihood point $\hat{\vartheta}$, the gradient of the log likelihood can only be orthogonal to the boundary at $\hat{\vartheta}$, and pointing towards the outside of the domain. Therefore $\tilde{I}_\mu$, which is defined as minus the gradient, will point inward. At the same time, D_μ will also always point toward the interior of the domain because of the form of the constraint we have chosen in Equation 5. Because by assumption $\|D_\mu\| = 1$, we have that

$$\tilde{I}_\mu = \|\tilde{I}_\nu\| D_\mu$$

and

$$\langle D_\mu, \tilde{I}_\mu \rangle_\Delta = \|D_\nu\|_\Delta \cdot \|\tilde{I}_\nu\|_\Delta$$

so that

$$\tilde{d} = d + D_\mu \hat{\vartheta}^\mu - \|D_\mu\|_\Delta \cdot \|\tilde{I}_\mu\|_\Delta \tag{27}$$

Now, the signed distance of the boundary to the origin in ξ-space is

$$l = -\frac{\tilde{d}}{\|\tilde{D}_\mu\|}$$

where the sign is taken such that l is negative when the origin is included in the integration domain. But noting that

$$K^\mu{}_\kappa \Delta^{\kappa\lambda} K^\nu{}_\lambda = \delta^{\mu\nu} \quad \Rightarrow \quad \Delta^{\mu\nu} = P^\mu{}_\kappa \delta^{\kappa\lambda} P^\nu{}_\lambda$$

we have

$$\begin{aligned} \|\tilde{D}_\mu\| = \sqrt{\tilde{D}_\mu \delta^{\mu\nu} \tilde{D}_\nu} &= \sqrt{\frac{1}{N} D_\kappa \left(P^\kappa{}_\mu \delta^{\mu\nu} P^\lambda{}_\nu \right) D_\lambda} \\ &= \sqrt{\frac{1}{N} D_\kappa \Delta^{\kappa\lambda} D_\lambda} = \frac{\|D_\mu\|_\Delta}{\sqrt{N}} \end{aligned}$$

and therefore

$$l = -\sqrt{N} \frac{\tilde{d}}{\|D_\mu\|} \tag{28}$$

Finally, by plugging Equation 27 into Equation 28 we obtain

$$\begin{aligned} l &= -\sqrt{N}\left[\frac{d + D_\mu \hat{\vartheta}^\mu}{\|D_\mu\|_\Delta} - \|\tilde{I}_\mu\|_\Delta\right] \\ &=: \sqrt{2}(s - m) \end{aligned} \tag{29}$$

where m and s are defined for convenience like so:

$$m := \sqrt{\frac{N}{2}}\frac{d + D_\mu \hat{\vartheta}^\mu}{\|D_\mu\|_\Delta} \quad (\geq 0) \tag{30}$$

$$s := \sqrt{\frac{N}{2}}\|\tilde{I}_\mu\|_\Delta \quad (\geq 0) \tag{31}$$

We note that m is a rescaled version of the margin defined by the constraint on the parameters (and therefore is never negative by assumption), and s is a rescaled version of the norm of the gradient of the log likelihood in the inverse observed Fisher metric (and therefore is nonnegative by construction).

A.1.5 Computing the penalty

We can now perform a final change of variables in the integral in Equation 22. We rotate our coordinates to align them to the boundary, so that

$$\tilde{D}_\mu = (\|\tilde{D}_\mu\|, 0, 0, \ldots, 0)$$

Note that we can always do this as our integrand is invariant under rotation. In this coordinate system, Equation 22 factorizes:

$$\begin{aligned} Q &= \frac{\exp\left[\frac{N}{2}\tilde{I}_\mu \Delta^{\mu\nu}\tilde{I}_\nu\right]}{\sqrt{\det \tilde{I}_{\mu\nu}}}\int_{\mathbb{R}^{d-1}} \mathrm{d}^{d-1}\xi \exp\left[-\frac{\xi_\mu \delta^{\mu\nu}\xi_\nu}{2}\right]\int_l^\infty \mathrm{d}\zeta \exp\left[-\frac{\zeta^2}{2}\right] \\ &= \sqrt{\frac{(2\pi)^d}{\det \tilde{I}_{\mu\nu}}}\exp\left[\frac{N}{2}\|\tilde{I}\|_\Delta^2\right]\frac{1}{\sqrt{\pi}}\int_l^\infty \frac{\mathrm{d}\zeta}{\sqrt{2}}\exp\left[-\frac{\zeta^2}{2}\right] \\ &= \sqrt{\frac{(2\pi)^d}{\det \tilde{I}_{\mu\nu}}}\exp\left(s^2\right)\frac{1}{\sqrt{\pi}}\int_{l/\sqrt{2}}^\infty \mathrm{d}\zeta \exp\left[-\zeta^2\right] \\ &= \sqrt{\frac{(2\pi)^d}{\det \tilde{I}_{\mu\nu}}}\exp\left(s^2\right)\frac{\mathrm{erfc}(s-m)}{2} \end{aligned} \tag{32}$$

where $\mathrm{erfc}(\cdot)$ is the complementary error function (*47*, section 7.1.2).

Finally, plugging Equation 32 into Equation 10 and taking the log, we obtain the extended FIA:

$$-\ln \mathbb{P}(\mathcal{M}_1|E) \simeq \ln \mathbb{P}(E|\hat{\vartheta}) + \frac{d}{2}\ln\frac{N}{2\pi} + \ln\int_\Theta \mathrm{d}^d\vartheta\sqrt{\det g} + \frac{1}{2}\ln\left[\frac{\det \tilde{I}_{\mu\nu}}{\det g_{\mu\nu}}\right] + B \tag{33}$$

where

$$B := \ln(2) - \ln\left[\exp\left(s^2\right)\mathrm{erfc}(s-m)\right] \tag{34}$$

can be interpreted as a penalty arising from the presence of the boundary in parameter space.

A.1.6 Interpreting the penalty

We now take a closer look at Equation 34. One key observation is that, by construction, at most one of m and s is ever nonzero. This is because in the interior of the manifold, $m > 0$ by definition, but $s = 0$ because the gradient of the likelihood is zero at $\hat{\vartheta}$; and on the boundary, $m = 0$ by definition, and s can be either zero or positive.

Interior of the manifold When $\hat{\vartheta}$ is in the interior of the parameter space Θ, then $\tilde{I}_\mu = 0 \Rightarrow s = 0$ and Equation 34 simplifies to

$$B = \ln(2) - \ln(\operatorname{erfc}(-m)) \tag{35}$$

but since N is large we have $m \gg 0$, $\operatorname{erfc}(-m) \to 2$ and $B \to 0$, so our result passes the first sanity check: we recover the expression in (*11*).

Boundary of the manifold When $\hat{\vartheta}$ is on the boundary of Θ, $m = 0$ and $s \geq 0$. Equation 34 becomes

$$B = \ln(2) - \ln\left[\exp\left(s^2\right)\operatorname{erfc}(s)\right] = \ln(2) - \ln(w(is)) \tag{36}$$

where w is the Feddeeva function (*47*, p. 7.1.3):

$$w(z) = e^{-z^2}\operatorname{erfc}(-iz)$$

This function is tabulated and can be computed efficiently. However, it is interesting to analyze its limiting behavior, as follows.

As a consistency check, when s is small we have at fixed N, to first order:

$$\begin{aligned} B &\simeq \ln(2) - \ln\left(1 - \frac{2s}{\sqrt{\pi}}\right) \\ &\simeq \ln(2) + \frac{2s}{\sqrt{\pi}} = \ln(2) + \sqrt{\frac{2N}{\pi}}\|\tilde{I}_\mu\|_\Delta \end{aligned} \tag{37}$$

and $B = \ln(2)$ when $\tilde{I}_\mu = 0$, as expected.

However, the real case of interest is the behavior of the penalty when N is assumed to be large, which is consistent with the fact that we derived Equation 32 as an asymptotic expansion of Equation 11. In this case, using the asymptotic expansion for the Feddeeva function (*47*, section 7.1.23):

$$\exp\left[s^2\right]\operatorname{erfc}(s) \sim \frac{1}{s\sqrt{\pi}}\left[1 + \sum_{m=1}^{\infty}(-1)^m \frac{1 \cdot 3 \cdots (2m-1)}{(2s^2)^m}\right]$$

To leading order, we obtain

$$\begin{aligned} B &\simeq \ln(2) + \ln\left(s\sqrt{\pi}\right) \\ &= \ln(2) + \ln\left(\sqrt{\frac{N\pi}{2}}\|\tilde{I}_\mu\|_\Delta\right) \end{aligned}$$

which we can rewrite as

$$\boxed{B \simeq \frac{1}{2}\ln\frac{N}{2\pi} + \ln\left[2\pi\|\tilde{I}_\mu\|_\Delta\right]} \tag{38}$$

We can summarize the above by saying that a new penalty term of order $\ln N$ arose due to the presence of the boundary. Interestingly, comparing Equation 38 with Equation 33 we see that the first term in Equation 38 is analogous to counting an extra parameter dimension in the original Fisher Information Approximation.

A.2 Behavioral experiments with human subjects

The behavioral task required subjects to view a screen showing two curves (one on the upper half, the other on the lower half of the screen) and 10 dots and decide, based on different instructions (see below for details), which curve was the more likely source of the observed dots. There were four task types that differed in terms of the shapes of the curves, corresponding to the different terms of the FIA (see Figure 1 in the main text and Figure S1): *dimensionality*, *boundary*, *volume*, and *robustness*. In each case, the curves represent two parametric statistical models of the form:

$$p(x|t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\|x - \mu(t)\|^2}{2\sigma^2}\right] \tag{39}$$

where x is a location on the 2D plane visualized on the screen, and $\mu(t)$, $t \in [0,1]$ is a parametrization of the curve. In other words, the curves represent Gaussians of unit isotropic variance whose mean μ can be located at any point along them. The dots shown to the subjects were sampled iid from one of the two models, selected at random with uniform probability. The location of the true mean of the Gaussian generating the dots (i.e., the value of t in the expression above) was randomly sampled from Jeffrey's prior for the selected model. All dots shown within a trial come from the same distribution (same model and same true mean). In the "generative" version of the task, the subjects had to report which curve (model) the dots are more likely to come from. In the "maximum-likelihood" version, the subjects had to report which curve was closest to the empirical centroid of the dot cloud. In both versions of the task, they pressed the "up" or "down" keys on their keyboard to select the curve in the upper or lower part of the screen, respectively.

Each model pairing was designed to emphasize a different term of the FIA. In the dimensionality variant, models have different dimensionality ($d = 0$ for the point and $d = 1$ for the line). In the boundary variant, both models have the same dimensionality and volume and are both flat so that their robustness terms are always identically zero. However, they are oriented such that, for ambiguous data falling around the midpoint between the two models, the influence of the boundary of the vertical model is stronger than that of the horizontal model. In the volume variant, models have the same dimensionality but different volume (length). In the robustness variant, models have the same dimensionality and volume, but their curvature is such that one of them bends away from the region of data space that is more likely to contain ambiguous stimuli, whereas the other bends around it (and therefore the robustness term for these models has opposite sign for data points that fall in that region).

A single run of the task consisted of a brief tutorial followed by 500 trials, divided in 5 blocks of 100 trials each. For each trial, the chosen curve pairing was presented, randomly flipped vertically to dissociate a fixed preference for one of the two models from a fixed preference for reporting "up" or "down". At the end of each block, the subject received feedback on their overall performance during that block. Subjects received a fixed compensation for taking part in the experiment.

We ran both experiments (generative and maximum-likelihood) on the online platform Pavlovia (`pavlovia.org`). For each task type, we collected data from at least 50 subjects who passed a pre-established performance threshold (60% correct for the robustness task variant and 70% correct for the other variants; these thresholds were chosen based on pilot data, and were fixed at preregistration (*24–26*)). We discarded the data collected from all other subjects. For the generative task, the final dataset included 52 subjects for the robustness task variant and 50 subjects for each of the other task variants. For the maximum-likelihood task, the final dataset included 51 subjects for the dimensionality task variant and 50 subjects for each of the others.

A.3 Detailed model definitions and computation of FIA terms

In this section, we report the detailed mathematical form of the models we used for the psychophysics experiment. Each model is defined by specifying the form of the function μ in Equation 39. Given this function, we then derive the analytical solution to the maximum-likelihood problem for any value of $X = \{x_i\}_{i=1}^N$, and finally the expressions for the likelihood (L), dimensionality (D), boundary (B), volume (V) and robustness (R) terms in the FIA for the model pairings we use in the experiment.

We also show that the (expected) Fisher information is constant for all models considered:

$$g(t) \equiv \frac{T^2}{\sigma^2} \tag{40}$$

so that Jeffrey's prior is simply the uniform probability distribution over the $[0,1]$ interval:

$$w(t) = \mathbb{1}_{[0,1]}(t) \tag{41}$$

A.3.1 Fisher information and robustness term for curved exponential families

In the following, we compute the observed Fisher information for each of our models. To do so, it is convenient to have a general expression for the Hessian of the log likelihood and for the observed and expected Fisher information for curved exponential families.

The general form of a curved exponential family is:

$$p(x|u) = \exp\left[C(x) + \vartheta^i(u)F_i(x) - \psi(\vartheta(u))\right] \tag{42}$$

where $\vartheta(u) : \mathbb{R}^d \to \mathbb{R}^k$, $k \geq d$, is a smooth parametrization. The Hessian of the log-likelihood is:

$$\begin{aligned}
\partial_a\partial_b \log p(x|u) &= F_i(x)\partial_a\partial_b\vartheta^i(u) - \partial_a\partial_b\psi(\vartheta(u)) \\
&= F_i(x)\frac{\partial^2\vartheta^i}{\partial u^a \partial u^b} - \frac{\partial}{\partial u^a}\left(\frac{\partial\psi}{\partial\vartheta^i}\frac{\partial\vartheta^i}{\partial u^b}\right) \\
&= F_i(x)\frac{\partial^2\vartheta^i}{\partial u^a \partial u^b} - \frac{\partial\vartheta^j}{\partial u^a}\frac{\partial\psi}{\partial\vartheta^j\partial\vartheta^i}\frac{\partial\vartheta^i}{\partial u^b} - \frac{\partial\psi}{\partial\vartheta^i}\frac{\partial^2\vartheta^i}{\partial u^a\partial u^b} \\
&= \frac{\partial^2\vartheta^i}{\partial u^a \partial u^b}\left[F_i(x) - \mathbb{E}_u[F_i]\right] - \frac{\partial\vartheta^j}{\partial u^a} g_{ji}\frac{\partial\vartheta^i}{\partial u^b}
\end{aligned} \tag{43}$$

where we note that $g_{ij} = -\mathrm{Cov}_u[F]_{ji}$ (remember that by g_{ij} we indicate the Fisher information of the ambient family). Therefore, the (expected) Fisher information is:

$$g_{ab} = \mathbb{E}_u\left[-\partial_a\partial_b \log p(x_i|u)\right] = \frac{\partial\vartheta^j}{\partial u^a} g_{ji}\frac{\partial\vartheta^i}{\partial u^b} \tag{44}$$

and the oberved Fisher information is:

$$\begin{aligned}
h_{ab} &= -\frac{1}{N}\sum_{i=1}^{N}\partial_a\partial_b \log p(x_i|u) \\
&= g_{ab} + \frac{\partial^2\vartheta^i}{\partial u^a\partial u^b}\left[\mathbb{E}_u[F_i] - \frac{1}{N}\sum_{n=1}^{N}F_i(x_n)\right]
\end{aligned} \tag{45}$$

As a corollary, we note that $h_{ab} = g_{ab}$ whenever $\vartheta(\cdot)$ is an affine transformation, that is when

$$\vartheta^i(u) = A^i_a u^a + B^i \tag{46}$$

For some constant A^i_b and B^i. In this case (which corresponds to autoparallel submanifolds in the exponential connection, (*48*, Theorem 1.1)), the robustness term in the FIA is identically zero:

$$\vartheta^i(u) = A^i_a u^a + B^i \Rightarrow R(X;u) \equiv 0 \tag{47}$$

A.3.2 General properties of curved 2D Gaussian models

Our models of interest, defined through Equation 39, are a special case of curved exponential families. They are all submanifolds of the same, larger model — the 2-dimensional exponential family of 2D Gaussian distributions with known isotropic covariance and unknown center. We call this larger family the *ambient family* $\mathcal{S} \supset \mathcal{M}$, composed by all probability distributions whose density is of the form:

$$p(x|\mu) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{\|x-\mu\|^2}{2\sigma^2}\right] \tag{48}$$

We can reduce Equation 39 to the notation of Equation 42 by noting that

$$\begin{aligned}
\ln p(x|t) &= -\frac{\|x-\mu\|^2}{2\sigma^2} - \ln\left(2\pi\sigma^2\right) \\
&= -\frac{1}{2}\|x\|^2_{g_{ij}} + \mu^i(t)g_{ij}x^j - \frac{1}{2}\left[\|\mu(t)\|^2_{g_{ij}} + \ln\left((2\pi)^2\det g_{ij}\right)\right]
\end{aligned} \tag{49}$$

where we indicate by g_{ij} the Fisher information of the ambient family $\mathcal{S}$:

$$g_{ij} = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} \tag{50}$$

By comparing Equation 49 with Equation 42 we see that

$$C(x) = -\frac{1}{2}\|x\|^2_{g_{ij}} \tag{51}$$

$$F^i(x) = x^i \qquad (\text{and } F_i(x) = g_{ij}x^j = \frac{x^i}{\sigma^2}) \tag{52}$$

$$\psi(t) = \frac{1}{2}\left[\|\mu(t)\|^2_{g_{ij}} + \ln\left((2\pi)^2 \det g_{ij}\right)\right] \tag{53}$$

and that $\mu(t)$ plays the role that $\vartheta(u)$ played in Equation 42.

We can now compute the expected and observed Fisher information for our models by specializing Equation 44 and Equation 45:

$$g(t) = \dot{\mu}^i(t) g_{ij} \dot{\mu}^j(t) \tag{54}$$

$$h(t) = g(t) + \ddot{\mu}^i(t) g_{ij}\left[\mu^j(t) - \bar{x}\right] \tag{55}$$

Where $\bar{x}$ is the empirical centroid of the dataset X,

$$\bar{x} = \bar{x}(X) := \frac{1}{N}\sum_{i=1}^{N} x_i \tag{56}$$

and g and h have no indices, because they are scalar functions of t.

We note then that $g(t)$ is simply the squared Euclidean norm of the vector $\dot{\mu}(t)$ divided by σ^2. In other words, the geometry of $\mathcal{M}$ coincides, up to scaling by σ^2, with the Euclidean geometry of the plane curve $\mu(t)$. This very convenient fact is a consequence of the particularly simple noise model we have assumed (Gaussian with known isotropic covariance).

Model volume The volume of a model described by $\mu(\cdot)$ is

$$\int_0^1 \mathrm{d}t \sqrt{g(t)} = \int_0^1 \mathrm{d}t \sqrt{\dot{\mu}^i(t) g_{ij} \dot{\mu}^j(t)} = \frac{1}{\sigma}\int_0^1 \mathrm{d}t \|\dot{\mu}(t)\| \tag{57}$$

In other words, it is simply the length of the curve $\mu(\cdot)$ measured in units of σ.

Likelihood gradient and maximum-likelihood point In the following, we will indicate the log-likelihood function for a model by

$$l = l(x;t) = \ln p(x|t) \tag{58}$$

In order to find the maximum-likelihood point for our models, it is convenient to write a general expression for the score function (the derivative of the log likelihood with respect to the parameter). We start by noting that

$$\begin{aligned}\frac{\partial}{\partial t}\ln p(X|t) &= -\sum_i \frac{\partial}{\partial t}\frac{\|x_i - \mu(t)\|^2}{2\sigma^2} = \frac{N}{2\sigma^2}\left[\frac{2}{N}\sum_n x_n - \mu(t)\right]\cdot\dot{\mu}(t) \\ &= N\frac{\partial}{\partial t}\ln p(\bar{x}|t) = N\frac{\partial l(\bar{x};t)}{\partial t}\end{aligned}$$

Therefore, to find the maximum likelihood point $\hat{t}$ for a certain X we can simply solve the corresponding one-sample ($N = 1$) case for the centroid $\bar{x}$. We can also write the rescaled likelihood gradient (which appears in the FIA as I_μ) as

$$-\frac{1}{N}\frac{\partial l}{\partial t}(X;t) = -\frac{1}{N}\frac{\partial}{\partial t}\ln p(X|\mu(t)) = \dot{\mu}^i(t) g_{ij}\left[\mu^j(t) - \bar{x}^j\right] \tag{59}$$

If we interpret $\dot{\mu}(t)$ as the tangent vector to μ in t, we see that away from model boundaries this equation expresses the familiar condition that the maximum-likelihood point (where $\partial l/\partial t = 0$) is the (Euclidean) orthogonal projection of $\bar{x}$ onto the model manifold. Again, this convenient property is a consequence of assuming isotropic Gaussian noise.

A.3.3 Horizontal model

This model, used in the dimensionality, boundary, and volume task variants, is defined as

$$\mu(t) = \begin{bmatrix} T\left(t - \frac{1}{2}\right) \\ \tau \end{bmatrix} \tag{60}$$

It is immediately evident that this model has volume (length) T/σ. The "base" model corresponds to $T = 1$, $\tau = 0$, and the model type called "horizontal" is defined with $T = 3$, $\tau = 1$.

Because

$$\dot{\mu}(t) = T \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{61}$$

and following Equation 54 and Equation 55, the observed and expected Fisher information coincide and are given by

$$g = h = \frac{T^2}{\sigma^2} \tag{62}$$

Given a centroid $X = [X^1, X^2]^T$, the rescaled likelihood gradient is (from Equation 59)

$$\frac{1}{N}\frac{\partial l}{\partial t}(X;t) = \frac{T^2}{\sigma^2}\left[\frac{\bar{x}^1}{T} - \left(t - \frac{1}{2}\right)\right] \tag{63}$$

and the maximum-likelihood point $\hat{t}$ is

$$\hat{t}(X) = \begin{cases} 0 & \text{if } \bar{x}^1 < -T/2 \\ \frac{1}{2} + \frac{\bar{x}^1}{T} & \text{if } -T/2 < \bar{x}^1 < T/2 \\ 1 & \text{if } \bar{x}^1 > T/2 \end{cases} \tag{64}$$

All the FIA terms can be computed in closed form from these expressions:

$$L(X) = -\frac{N}{2\sigma^2}\left[(\bar{x}^1 - T(\hat{t}(X) - 1/2))^2 + (\bar{x}^2 - \tau)^2\right] - \frac{N}{2}\ln\left(2\pi\sigma^2\right) \tag{65}$$

$$D = \frac{1}{2}\ln\frac{N}{2\pi} \tag{66}$$

$$B = \frac{1}{2}\ln\frac{N}{2\pi} + \ln\left[2\pi\frac{T}{\sigma}\left|\frac{\bar{x}^1}{T} - \left(\hat{t}(X) - \frac{1}{2}\right)\right|\right] \quad \left(\text{if } |\bar{x}^1| > \frac{T}{2}\right) \tag{67}$$

$$V = \ln\frac{T}{\sigma} \tag{68}$$

$$R = 0 \tag{69}$$

A.3.4 Vertical model

This model, used the boundary task variant, is just a rotated and translated version of the horizontal model. It is defined as

$$\mu(t) = \begin{bmatrix} 0 \\ \tau + Tt \end{bmatrix} \tag{70}$$

where we keep T and τ as arbitrary parameters for notational clarity, although in practice they are both fixed to 1 in our study. From the definition, it follows that

$$\dot{\mu}(t) = T \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{71}$$

$$g = h = \frac{T^2}{\sigma^2} \tag{72}$$

$$\frac{1}{N}\frac{\partial l}{\partial t}(X;t) = \frac{T^2}{\sigma^2}\left[\frac{X^2 - \tau}{T} - t\right] \tag{73}$$

and

$$\hat{t}(X) = \begin{cases} 0 & \text{if } \bar{x}^2 < \tau \\ \frac{\bar{x}^2 - \tau}{T} & \text{if } \tau < \bar{x}^2 < \tau + T \\ 1 & \text{if } \bar{x}^2 > \tau + T \end{cases} \tag{74}$$

so that the FIA terms can be written as

$$L(X) = -\frac{N}{2\sigma^2}\left[(\bar{x}^1)^2 + (\bar{x}^2 - \tau - T\hat{t}(X))^2\right] - \frac{N}{2}\ln\left(2\pi\sigma^2\right) \tag{75}$$

$$D = \frac{1}{2}\ln\frac{N}{2\pi} \tag{76}$$

$$B = \frac{1}{2}\ln\frac{N}{2\pi} + \ln\left[2\pi\frac{T}{\sigma}\left|\frac{\bar{x}^2 - \tau}{T} - \hat{t}(X)\right|\right] \quad \left(\text{if } \bar{x}^2 < \tau \vee \bar{x}^2 > \tau + T\right) \tag{77}$$

$$V = \ln\frac{T}{\sigma} \tag{78}$$

$$R = 0 \tag{79}$$

A.3.5 Circular-arc model

This model, used in the robustness task variant, is constituted by an arc of a circle, and is defined as

$$\mu(t) = \begin{bmatrix} \frac{T}{\gamma}\sin(\alpha) \\ \tau + \frac{T}{\gamma}(1 - \cos(\alpha)) \end{bmatrix} \tag{80}$$

where

$$\alpha = \gamma\left(t - \frac{1}{2}\right) \tag{81}$$

and γ is a positive constant. Concretely, in the experiments we fixed $\gamma = (3/5)\pi$, and T to the value determined below for the rounded model type (Equation 99). We note that the radius of the circle is $r = T/\gamma$, and the y-coordinate of the center is $\tau + r$. The tangent vector $\dot{\mu}$ and the acceleration vector $\ddot{\mu}$ in t are

$$\dot{\mu}(t) = T\begin{bmatrix} \cos(\alpha) \\ \sin(\alpha) \end{bmatrix} \tag{82}$$

$$\ddot{\mu}(t) = T\gamma\begin{bmatrix} -\sin(\alpha) \\ \cos(\alpha) \end{bmatrix} \tag{83}$$

so that, by substitution in Equation 54,

$$g = \dot{\mu}^i g_{ij}\dot{\mu}^j = \frac{T^2}{\sigma^2}(\cos^2(\alpha) + \sin^2(\alpha)) = \frac{T^2}{\sigma^2} \tag{84}$$

and by substitution in Equation 55

$$\begin{aligned} h &= g + \frac{T^2}{\sigma^2}\left[-\sin^2(\alpha) + \frac{\gamma\bar{x}^1}{T}\sin(\alpha) + \frac{\gamma\tau}{T}\cos(\alpha) + \cos(\alpha) - \cos^2(\alpha) - \frac{\gamma\bar{x}^2}{T}\cos(\alpha)\right] \\ &= g\left[\frac{\gamma\bar{x}^1}{T}\sin(\alpha) + \frac{\gamma\tau}{T}\cos(\alpha) + \cos(\alpha) - \frac{\gamma\bar{x}^2}{T}\cos(\alpha)\right] \\ &= \frac{g}{r}\left[\sin(\alpha)\bar{x}^1 + \cos(\alpha)\left(\tau + r - \bar{x}^2\right)\right] \end{aligned} \tag{85}$$

The rescaled likelihood gradient is (from Equation 59)

$$\begin{aligned} -\frac{1}{N}\frac{\partial}{\partial t}\ln(p(X|t)) &= \frac{T^2}{\gamma\sigma^2}\Big[\cos(\alpha)\sin(\alpha) - \frac{\gamma\bar{x}^1}{T}\cos(\alpha) \\ &\quad + \frac{\gamma\tau}{T}\sin(\alpha) + \sin(\alpha) - \sin(\alpha)\cos(\alpha) - \frac{\gamma\bar{x}^2}{T}\sin(\alpha)\Big] \\ &= \frac{T}{\sigma^2}\left[-\bar{x}^1\cos(\alpha) + \tau\sin(\alpha) + r\sin(\alpha) - \bar{x}^2\sin(\alpha)\right] \\ &= \frac{T}{\sigma^2}\left[-\cos(\alpha)\bar{x}^1 + \sin(\alpha)\left(\tau + r - \bar{x}^2\right)\right] \end{aligned} \tag{86}$$

(note that h can also be obtained by differentiating this last expression).

To compute the FIA, we need the maximum-likelihood projection. As for the other models, this projection is defined piecewise due to the presence of model boundaries. To properly partition the plane, we need to define first the equation for the line intersecting the model perpendicularly at t:

$$\rho(x;t) = \tau + r - \cot(\alpha(t))x \tag{87}$$

With this definition, the maximum-likelihood point is

$$\hat{t}(X) = \begin{cases} 0 & \text{if } \bar{x}^1 < 0 \wedge \bar{x}^2 > \rho(\bar{x}^1;0) \\ 1 & \text{if } \bar{x}^1 > 0 \wedge \bar{x}^2 > \rho(\bar{x}^1;1) \\ \frac{1}{2} + \frac{1}{\gamma}\arctan\frac{\bar{x}^1}{\tau+r-\bar{x}^2} & \text{otherwise} \end{cases} \tag{88}$$

and therefore the FIA terms are:

$$L(X) = -\frac{N}{2\sigma^2}\|\bar{x} - \mu(\hat{t}(X))\|^2 - \frac{N}{2}\ln\left(2\pi\sigma^2\right) \tag{89}$$

$$D = \frac{1}{2}\ln\frac{N}{2\pi} \tag{90}$$

$$B = \frac{1}{2}\ln\frac{N}{2\pi} + \ln\left[2\pi\sqrt{\frac{r}{\sigma^2}\frac{\left[-\cos\left(\alpha(\hat{t})\right)\bar{x}^1 + \sin\left(\alpha(\hat{t})\right)(\tau + r - \bar{x}^2)\right]^2}{\sin\left(\alpha(\hat{t})\right)\bar{x}^1 + \cos\left(\alpha(\hat{t})\right)(\tau + r - \bar{x}^2)}}\right] \tag{91}$$

$$V = \ln\frac{T}{\sigma} \tag{92}$$

$$R = \sin\left(\alpha(\hat{t})\right)\frac{\bar{x}^1}{r} + \cos\left(\alpha(\hat{t})\right)\frac{\tau + r - \bar{x}^2}{r} \tag{93}$$

where the value given for B is relevant only when $\hat{t}$ is either 0 or 1. Note that, due to the shape of the model and the presence of the boundary, there are regions of the data space such that the log-likelihood function at the maximum likelihood point will not be concave. These regions represent a complete breakdown of the FIA, but they are not a problem in practice because the approximation holds in the region of data space that is relevant for the experiments (see Figure S1).

A.3.6 Rounded model

This model, also used in the robustness task variant, is a circular arc (like the "circular" model described above) with two straight arms attached on either side. The ratio of the length of the circular section of the model over its total length is defined as a parameter f. The model definition is

$$\mu(t) = \begin{cases} \begin{bmatrix} -T\left[\frac{f}{\gamma}\sin(\gamma/2) - \left(t - \frac{(1-f)}{2}\right)\cos(\gamma/2)\right] \\ \tau + T\left[\frac{f}{\gamma}(1-\cos(\gamma/2)) - \left(t - \frac{1-f}{2}\right)\sin(\gamma/2)\right] \end{bmatrix} & \text{if } t < \frac{1-f}{2} \\ \mu_c\left(\frac{t-(1-f)/2}{f};\tau,\gamma,fT\right) & \text{if } \frac{1-f}{2} \le t \le \frac{1+f}{2} \\ \begin{bmatrix} T\left[\frac{f}{\gamma}\sin(\gamma/2) + \left(t - \frac{(1+f)}{2}\right)\cos(\gamma/2)\right] \\ \tau + T\left[\frac{f}{\gamma}(1-\cos(\gamma/2)) + \left(t - \frac{1+f}{2}\right)\sin(\gamma/2)\right] \end{bmatrix} & \text{if } t > \frac{1+f}{2} \end{cases} \tag{94}$$

where μ_c is the μ mapping defined for the circular model, Equation 80.

For the experiment, the values of the parameters were chosen to guarantee that the circular section of this model would have the same center as a circular model (described above) with $\gamma = (3/5)\pi$ and $\tau = 0$,

and that a relatively large fraction of the two models is in close proximity. The values are

$$f = 1/3 \tag{95}$$
$$\gamma = (3/5)\pi \tag{96}$$
$$\tau = 3/5 \tag{97}$$
$$T = \frac{\tau\gamma}{1-f} \tag{98}$$

Closed-form expressions for all FIA terms can be derived for this model by a straightforward, if somewhat laborious, extension of those presented above for the circular-arc model. We do not report them here in the interest of brevity.

A.3.7 Point model

This model, used in the dimensionality task variant, has no associated latent parameters (it is zero-dimensional). To cast it in the same language as the others, we can define it as

$$\mu(t) = \mu = \begin{bmatrix} 0 \\ \tau \end{bmatrix} \tag{99}$$

For the point model, the FIA (which is an approximation to a model's log evidence) is replaced by the exact evidence, which simply coincides with the log likelihood. For notational consistency, we adopt the following values for the FIA terms:

$$L(X) = \frac{N}{2\sigma^2}\left[(\bar{x}^1)^2 + (\bar{x}^2 - \tau)^2\right] - \frac{N}{2}\ln\left(2\pi\sigma^2\right) \tag{100}$$
$$D = 0 \tag{101}$$
$$B = 0 \tag{102}$$
$$V = 0 \tag{103}$$
$$R = 0 \tag{104}$$

A.4 Numerical experiments with artificial neural networks

A.4.1 Inputs

On each trial, our artificial neural network (henceforth ANN) takes in two images, each depicting one candidate model's location in the data space. It also takes in a length-20 vector, containing the horizontal and vertical coordinates of the $N = 10$ data points. Each image is provided as one RGB matrix of size (3*256*256). In data space units (used for the model definitions in subsection A.3), each image extends from $x = -4$ to $x = 4$ and from $y = -3.5$ to $y = 4.5$, so that the center of the image (located in $(0, 0.5)$) is equidistant from the models in each model pair.

A.4.2 Training dataset

The training dataset consisted of 5000 model pairs. Each model pair was used for generating 50 trials. This approach led to a total of 250000 trials in the entire dataset.

The random generation of model pairs was as follows (see subsection A.3 for the detailed mathematical definitions of each model and the precise meaning of the parameters controlling its shape). Each model pair could be of one of the four variants described in subsection A.2, chosen randomly with equal probabilities. Each model pair could be flipped vertically with probability 0.5. For the robustness variant, the separation of the model pair was 0.6 data space units; for all other model pairs, the separation was 1 data space unit. For the dimensionality variant, the length T (in data space units) of the one-dimensional model was sampled uniformly from $\mathcal{U}(0.5, 5)$. For the boundary variant, the length of both model families were kept identical and sampled from $\mathcal{U}(0.5, 3)$. For the volume variant, the lengths of both models were sampled independently

from $\mathcal{U}(0.5, 5)$; if their length difference was no greater than the task's noise level $\sigma = 1$, then the length of one model was resampled from $\mathcal{U}(0.5, 5)$ until the length difference was greater than 1. For the robustness variant, the length of both model families was kept constant at $(27/50) \cdot \pi$. The length proportion of the rounded model that was perfectly circular was $f = 1/3$, and both model families share the same curvature parameter γ sampled from $\mathcal{U}(1.5, 3)$. The model pairs were centered around the center of each input image.

Given a model pair, each trial was randomly generated as follows. Select one model randomly with equal probability. Sample a location along this model uniformly. Using this location as the center of a 2D isotropic Gaussian and standard deviation of $\sigma = 1$ data space units, sample $N = 10$ data points that were observable to the subject.

The training dataset was pre-shuffled randomly for training purposes. The input batch size was always 50 trials.

A.4.3 Test dataset

The test dataset consisted of 8 model pairs, each generating 15000 trials. Thus, there was a total of 120000 trials in the dataset.

The model pairs were as follows. For the point variant, the one-dimensional model had length (in data space units) 1. For the boundary variant, both model families had length 1. For the volume variant, one model had length 1 while the other had 3. For the robustness variant, both model families had length $T = (27/50) \cdot \pi$, $f = 1/3$, and curvature parameter $\gamma = (3/5)\pi$. Each model pair was presented in the "upright" position (as per the definitions in subsection A.3) and in the vertically flipped position, for a total of 8 cases. The separation between model families and the generation of trials was identical to as in the training dataset.

A.4.4 Artificial neural network architecture

Our ANN had the following architecture (see Figure 3. Each of the two model input images was passed through the pretrained convolutional neural network VGG16, which had its parameters frozen during training. We replaced the fully connected layers at the end of VGG16 with our own structure of Linear-ReLU-BatchNorm1D layers and allowed the updating of weights in these and all subsequent layers. For each image input, the output of this image-processing module was a length-50 vector (model image representation).

In parallel, the length-20 vector of raw data point coordinates was fed through a permutation-invariant layer. This layer featured shared weights such that its outputs were not affected by the sequence of the N=10 data points in the length-20 vector input. This layer also outputted a length-20 vector, which was concatenated to the end of each of the length-50 vectors (the model image representations) along the preexisting dimension, producing two length-70 vectors.

Each length-70 vector was fed through Linear-ReLU-BatchNorm1D layers (identical weights used to process each vector). The resultant two length-50 output vectors were then concatenated together along the preexisting dimension, with the first input image's representation in front.

The resultant length-100 vector was then fed through EquiLinear-ReLU-BatchNorm1D layers. The EquiLinear layers were permutation-equivariant layers of our design, again achieved by weight sharing. They ensure that if we concatenated the two length-50 output vectors in the opposite sequence, then their output, a length-2 vector, also had the same values but in opposite sequence. This length-2 vector was passed through a log softmax layer to produce the ANN's final output, which was also a length-2 vector.

We also introduced a conditional variational encoder (CVAE) structure and used its output as part of the loss function (discussed later), to encourage model representations to preserve information about the data generation process. The details are described below.

We concatenate the length-20 raw data points vector (before passing input the permutation-invariant layers) to the end of each length-50 model image representation vector. The resultant two length-70 vectors (each corresponding to one model) were used as inputs for our CVAE (identical weights used to process each vector). The CVAE took each length-70 vector through its encoder structure to produce 10-dimensional vectors, which were used as parameters $(\mu_{CVAE}, \sigma_{CVAE})$ for the Gaussian random generation of another 10-dimensional vector. The latter vector was again concatenated to the end of the length-50 model image

representation vector responsible for its own generation, before being fed to the CVAE decoder, which mapped back to a 20-dimensional output vector reminiscent of data points. Hence, there were two 20-dimensional output vectors generated, each originating from one model.

A.4.5 Loss function

The loss function for each trial consisted of 2 parts: 1) the final output loss, and 2) the CVAE output loss. For the final output loss, we used Pytorch's negative log likelihood loss function `NLLLoss()`, which computed the loss between the ANN's length-2 output vector and the target label. For each trial's CVAE output loss, we considered only the CVAE output associated with the correct model image/target label (hence one out of the two CVAE output vectors). The CVAE output loss was the sum of a MSE reconstruction loss (between the length-20 CVAE output vector and the length-20 raw data points vector) and a KL Divergence Loss (considering $(\mu_{CVAE}, \sigma_{CVAE})$ used in the CVAE data generation process, using sum reduction). The total loss was the sum of the final output loss and the CVAE output loss.

A.4.6 Update rule

We used Pytorch's Adam optimizer with learning rate 0.005, keeping all other arguments to their default values.

A.4.7 ANN predictions

To evaluate ANN task performance in a way that is comparable to human performance, we need to specify how the ANN output, a length-2 log softmax vector, maps onto a chosen candidate model. The mapping is as follows: we compare the two entries in the output vector and assume that the ANN chooses the candidate model associated with the larger entry.

A.5 Experimental data analysis

For both human and artificial neural network (ANN) experiments, we modeled behavior assuming that each subject samples from a posterior over models determined by a modified version of the Fisher Information Approximation (FIA), where each term of the approximation is multiplied by a free parameter to be inferred, representing the sensitivity of the subject to that term.

Specifically, in our experimental scenario the theory of Bayesian model selection applies directly. Given two models $\mathcal{M}_1$ and $\mathcal{M}_2$, assuming a flat prior over models $p(\mathcal{M}_1) = p(\mathcal{M}_2) = 1/2$ and an uninformative prior (Jeffrey's prior, see Balasubramanian (*11*) and Jaynes (*15*)) over the parameters of each model, when N is sufficiently large we can use the asymptotic expansion in Figure 1 and Equation 33 to write the log posterior ratio for $\mathcal{M}_1$ over $\mathcal{M}_2$ as

$$\begin{aligned}\log \frac{p(\mathcal{M}_1|X)}{p(\mathcal{M}_2|X)} &= \log \frac{p(\mathcal{M}_1|X)}{1-p(\mathcal{M}_1|X)} \\ &\simeq (L_2 - L_1) + (D_2 - D_1) + (B_2 - B_1) + (V_2 - V_1) + (R_2 - R_1)\end{aligned} \tag{105}$$

where L_i, D_i, etc represent the FIA terms for model i:

$$\begin{aligned}L_i &= -\log p(X|\hat{\vartheta}, \mathcal{M}_i) \quad \text{(Likelihood)} \\ D_i &= \frac{d}{2} \log \frac{N}{2\pi} \quad \text{(Dimensionality)} \\ B_i &= \frac{1}{2} \log \frac{N}{2\pi} + \log\left[2\pi \left\|\hat{l}\right\|_{\hat{h}^{-1}}\right] \quad \text{(Boundary)} \\ V_i &= \log \int \mathrm{d}^d\vartheta \sqrt{\det g(\vartheta)} \quad \text{(Volume)} \\ R_i &= \frac{1}{2} \log \left[\frac{\det h(X;\hat{\vartheta})}{\det g(\hat{\vartheta})}\right] \quad \text{(Robustness)}\end{aligned}$$

This expression suggests a simple normative model for subject behavior. Equation 105 determines the probability of reporting $\mathcal{M}_1$ for an ideal Bayesian observer performing probability matching. We can then compare subject behavior to the normative prescription by allowing subjects to have distinct sensitivities to the various terms of the FIA:

$$\begin{aligned}\log \frac{p(\text{report } \mathcal{M}_1|X)}{p(\text{report } \mathcal{M}_2|X)} =& \alpha + \beta_L(L_2 - L_1) + \beta_D(D_2 - D_1) + \\ &+ \beta_B(B_2 - B_1) + \beta_V(V_2 - V_1) + \beta_R(R_2 - R_1)\end{aligned} \tag{106}$$

where α and β were free parameters: α captures any fixed bias, β_L the sensitivity to differences in maximum likelihood, β_D the sensitivity to differences in dimensionality, and so on.

We fitted the model expressed by Equation 106 to subject behavior using a hierarchical, Bayesian logistic regression scheme:

$$\nu_\alpha, \nu_L, \ldots, \nu_R \sim 1 + \text{Exponential}(29) \tag{107}$$

$$\mu_\alpha, \mu_L, \ldots, \mu_R \sim \text{Normal}(0, 3) \tag{108}$$

$$\sigma_\alpha, \sigma_L, \ldots, \sigma_R \sim \text{Exponential}(3) \tag{109}$$

$$\alpha_i \sim \text{StudentT}(\nu_\alpha, \mu_\alpha, \sigma_\alpha) \tag{110}$$

$$\beta_{L,i} \sim \text{StudentT}(\nu_L, \mu_L, \sigma_L) \tag{111}$$

$$\vdots \tag{112}$$

$$\beta_{R,i} \sim \text{StudentT}(\nu_R, \mu_R, \sigma_R) \tag{113}$$

$$C_{i,t} \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\text{lpr}\left(\alpha_i, \beta_{L,i}, \beta_{D,i}, \beta_{B,i}, \beta_{V,i}, \beta_{R,i}, X_{i,t}\right)\right)\right) \tag{114}$$

where $C_{i,t}$ is the choice made by subject i on trial t, $X_{i,t}$ is the sensory stimulus on that same trial, lpr is the log posterior ratio defined by Equation 106, α_i is the bias for subject i, $\beta_{L,i}$ is the likelihood sensitivity of that same subject, and so on for the other sensitivity parameters. The bias and sensitivity parameters describing each subject are modeled as independent samples from a population-level Student-T probability distribution characterized by a certain shape (ν), location (μ) and scale (σ). The priors assumed over these population-level parameters are standard weakly informative priors (*28*, *29*), and broader or flat priors lead to similar results to those presented below. The model was implemented in PyMC (*49*), and inference was performed by sampling from the posterior for the parameters given the experimental data $\{C_{i,t}, X_{i,t}\}$ using the No-U-Turn Sampler algorithm (*50*, *51*). Further technical details on the inference procedure can be found below, in subsubsection A.5.2.

Definition of relative sensitivity and presentation of sensitivity estimates. Relative sensitivity for a certain feature was defined as the sensitivity for that feature divided by the relevant posterior mean for the likelihood sensitivity. For instance, for dimensionality:

$$\tilde{\beta}_D = \frac{\beta_D}{\langle \beta_L \rangle_{p(\beta_L|\text{data})}} \tag{115}$$

This formulation applies both at the subject level and at the population level.

We note that, because each human subject performed only one task variant, not all sensitivities could be estimated for all subjects. For instance, β_D only entered the behavioral model (and therefore could be estimated) for the subjects that performed the *point* task variant, where the alternative models had different dimensionality. The same holds with β_V and the *horizontal* task variant, and β_R and the *rounded* task variant. The boundary term entered the behavioral model for all task variants, although by design it took on a much broader range of values for the *vertical* task. For consistency, for each sensitivity parameter, we reported its estimate only for those subjects that performed the task variant designed to test it.

A.5.1 Lapse-rate modeling

We designed a variant of the behavioral model that accounts for lapses in subjects' responses (i.e., errors on easy trials). Specifically, we modified Equation 106 as follows:

$$\begin{aligned}\log \frac{p(\text{report } \mathcal{M}_1|X)}{p(\text{report } \mathcal{M}_2|X)} = &\frac{\epsilon}{2} + (1-\epsilon)\big[\alpha + \beta_L(L_2 - L_1) + \beta_D(D_2 - D_1) + \\ &+ \beta_B(B_2 - B_1) + \beta_V(V_2 - V_1) + \beta_R(R_2 - R_1)\big]\end{aligned} \tag{116}$$

where $\epsilon \in [0, 1]$ is the lapse rate, representing the probability that a given response is completely random. For $\epsilon = 1$ the responses are random on every trial, whereas for $\epsilon = 0$ this model is equivalent to the original one in Equation 106.

To estimate λ from our experimental data jointly with all other parameters, we kept the same structure as in Equations 107- Equation 114 and exteded it by modeling the population level distribution of ϵ as a Beta distribution, parameterized by count parameters a and b. Following the recommendations in (*29*, section 5.3) and (*52*, section 24.2), we specify hyperpriors in terms of the mean of the distribution $\phi = a/(a+b)$ and the total count $\lambda = a + b$:

$$\phi \sim \text{Uniform}(0, 1) \tag{117}$$

$$\lambda \sim \text{Pareto}(0.1, 1.5) \quad \left(p(\lambda) \propto x^{-2.5}\right) \tag{118}$$

$$a = \lambda \cdot \phi \tag{119}$$

$$b = \lambda \cdot (1 - \phi) \tag{120}$$

$$\epsilon_i \sim \text{Beta}(a, b) \tag{121}$$

where ϵ_i is the lapse rate for subject i.

A.5.2 Technical details of the inference procedure

Posterior sampling was performed with PyMC (*49*) version 4.2.0, using the NUTS Hamiltonian Monte Carlo algorithm (*50*). Target acceptance probability was set to 0.9 for the human data (both generative and maximum-likelihood task), to 0.8 for the generative task with neural networks and for 0.99 for the maximum-likelihood task for neural networks. The posterior distributions were built by sampling 12 independent Markov chains for 10000 draws each. No divergence occurred in any of the chains. Effective sample size and $\hat{R}$ diagnostics for some of the key parameters are given in Table S5.

A.5.3 Reporting of posterior distributions for inferred parameters

The posterior distributions reported in all figures are Kernel Density Estimates with bandwidth chosen according to Scott's rule (*53*).

B Supplementary Figures and Tables

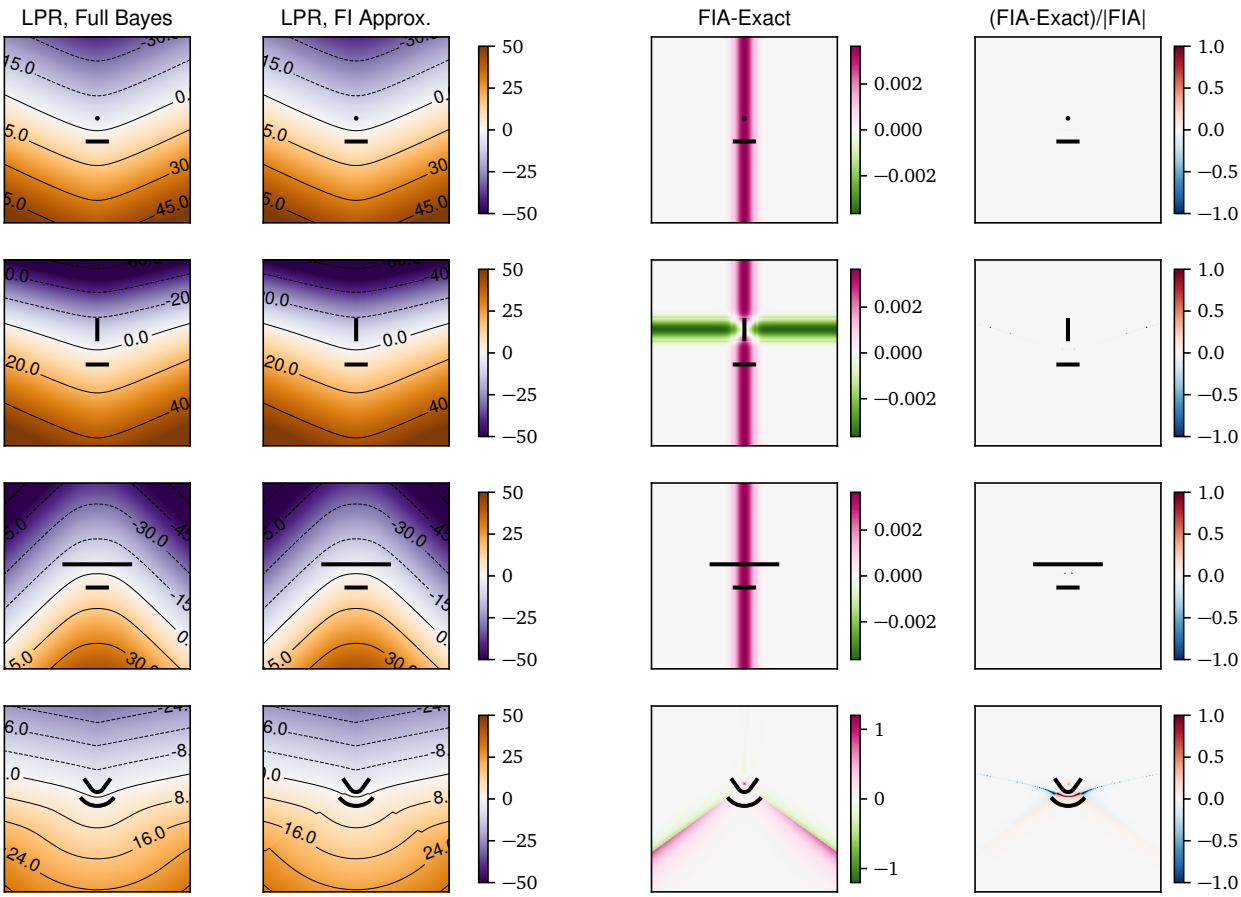

Figure S1: Comparison of the full Bayesian and Fisher Information Approximation computation of the log posterior ratio (LPR) for the model pairs used in our psychophysics tasks ($N = 10$). Each row corresponds to one task variants (from top to bottom, "dimensionality", "boundary", "volume", "robustness"). First column from the left: full Bayesian LPR, computed by numerical integration. Second column: LPR computed with the Fisher Information Approximation. Third column: difference between FIA and exact LPR. Fourth column: relative difference (difference divided by the absolute value of the FIA LPR). Adapted from (*46*).

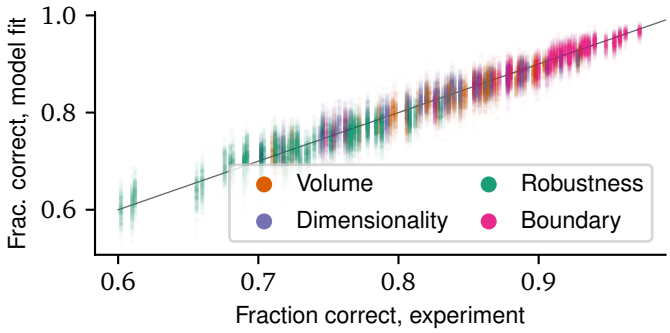

Figure S2: Posterior predictive check for the human subjects in the generative task, looking at subject performance. We sampled 240 samples from the posterior over model parameters by thinning the MCMC chains used for model inference. For each of these samples, we ran a simulation of the experiment using the actual stimuli shown to the subjects, and we recorded the resulting performance of all 202 simulated subjects. This procedure yielded 240 samples of the joint posterior-predictive distribution of task performance over all experimental subjects. To visualize this distribution, for each subject we plotted a cloud of 240 dots where the y coordinate of each dot is the simulated performance of that subject in one of the simulations, and the x coordinate is the true performance of that subject in the experiment plus a small random jitter (for ease of visualization). The gray line is the identity, showing that our inference procedure captures well the behavioral patterns in the experimental data. In the figure, all task types are pooled together, but subjects that performed different task types are distinguished by the color of the dots.

	Orig. task		Max lik. task	
Parameter	Humans	ANNs	Humans	ANNs
μ_α (up/down bias)	0.107 ± 0.023	-0.242 ± 0.151	0.056 ± 0.024	0.010 ± 0.07
μ_L (likelihood)	0.461 ± 0.012	6.529 ± 0.188	0.561 ± 0.018	7.966 ± 0.401
μ_D (dimensionality)	2.150 ± 0.445	8.484 ± 0.653	1.285 ± 0.231	-0.030 ± 0.486
μ_B (boundary)	0.518 ± 0.045	6.286 ± 0.186	0.499 ± 0.058	0.883 ± 0.121
μ_V (volume)	0.108 ± 0.057	6.089 ± 0.204	0.105 ± 0.044	0.128 ± 0.196
μ_R (robustness)	1.018 ± 0.056	4.882 ± 0.417	1.276 ± 0.085	0.356 ± 0.281

Table S1: Posterior mean±standard deviation for population-level parameters. See Equation 106 to Equation 114 for the precise definition of each parameter and its role in the hierarchical model of behavior.

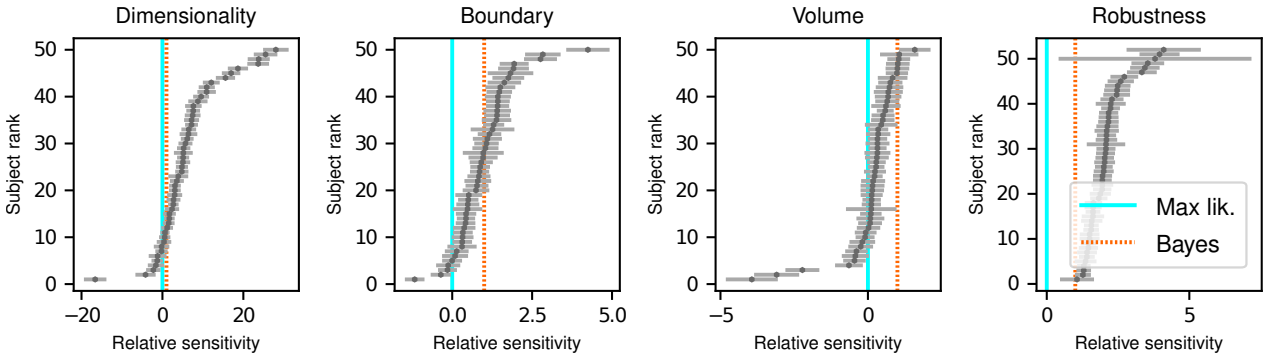

Figure S3: Subject-level relative sensitivities to the geometric features that determine model complexity. Dots with error bars: posterior mean ± standard deviation of the relative sensitivity (the dots are the same as in Figure 2c). For ease of visualization, subjects are ranked based on their posterior mean.

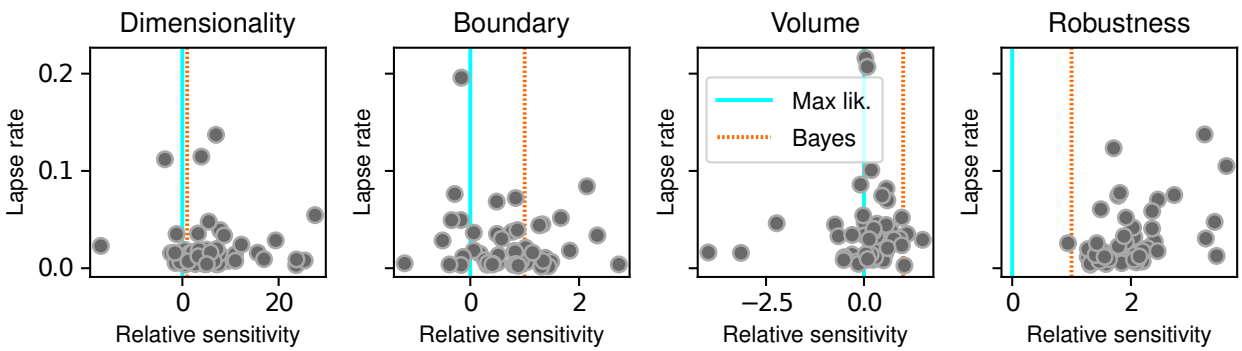

Figure S4: Lapse rate versus relative sensitivity to complexity across subjects. Each dot gives the posterior mean estimate of the relative sensitivity to one of the features that determine model complexity (abscissa) and the posterior mean estimate of the lapse rate, as defined in Section A.5.1.

Model	Rank	WAIC	pWAIC	dWAIC	SE	dSE
Full	0	-34824.6	641.963	0	183.981	0
Likelihood only	1	-37524.9	370.340	2700.4	183.923	69.3817

Table S2: WAIC comparison of the full model and the likelihood-only model for the human subjects in the generative task, reported in the standard format used by (*27*, section 6.4.2). Briefly, WAIC is the value of the criterion (log-score scale — higher is better); pWAIC is the estimated effective number of parameters; dWAIC is the difference between the WAIC of the given model and the highest-ranked one; SE is the standard error of the WAIC estimate; and dSE is the standard error of the difference in WAIC. These estimates were produced with the `compare` function provided by ArviZ (*54*), using 12 MCMC chains with 10000 samples each for each model (in total, 120000 samples for each model).

Model	Rank	WAIC	pWAIC	dWAIC	SE	dSE
Full	0	-31022.8	638.926	0	184.912	0
Likelihood only	1	-33155.1	374.023	2132.28	186.667	63.1851

Table S3: Same as Table S2, for the maximum-likelihood task, where subjects were asked to report the model that was closest to the data.

Parameter	ROPE	95% HDI (generative)	PD (generative)	95% HDI (max lik.)	PD (max lik.)
Likelihood	$\lvert\beta_L\rvert < 0.0076$	$[0.012, 0.437]$	1.00	$[0.526, 0.597]$	1.00
Dimensionality	$\lvert\beta_D\rvert < 0.43$	$[1.299, 3.048]$	1.00	$[0.835, 1.745]$	1.00
Boundary	$\lvert\beta_B\rvert < 0.06$	$[0.43, 0.604]$	1.00	$[0.386, 0.612]$	1.00
Volume	$\lvert\beta_V\rvert < 0.091$	$[-0.005, 0.218]$	0.97	$[0.019, 0.193]$	0.99
Robustness	$\lvert\beta_R\rvert < 0.11$	$[0.908, 1.126]$	1.00	$[1.11, 1.446]$	1.00

Table S4: HDI vs ROPE comparison and Probability of Direction (PD) for the population-level parameters in the human experiments. See Supplementary Information section C.7.2 and (*28*) for an explanation of the ROPE-HDI comparison, and (*55*, *56*) for more details on the probability of direction metric. Note that the ROPE and HDI definitions were preregistered (*24–26*).

Parameter	ESS	$\hat{R}$
μ_α	52148	1.000
μ_L	19254	1.000
μ_D	27384	1.000
μ_B	32434	1.000
μ_V	64614	1.000
μ_R	112118	1.000

Table S5: $\hat{R}$ statistic and effective sample size (ESS) for 12 Markov Chain traces run as described in the text, for the fit to human data in the generative task. See (*29*, sections 11.4–11.5) and Vehtari *et al.* (*57*) for in-depth discussion of chain quality diagnostics. Briefly, $\hat{R}$ depends on the relationship between the variance of the draws estimated within and between contiguous draw sequences. $\hat{R}$ is close to 1 when the chains have successfully converged. The effective sample size estimates how many independent samples one would need to extract the same amount of information as that contained in the (correlated) MCMC draws.

C Supplementary Text

C.1 Numerical comparison of the extended FIA vs exact Bayes

Figure S1 shows that the FIA computed with the expressions given in this document provides a very good approximation to the exact Bayesian log posterior ratio (LPR) for the model pairs used in the psychophysics experiments, and for the chosen sample size ($N = 10$). As highlighted in the panels in the rightmost column, the discrepancies between the exact and the approximated LPR are generally small in relative terms, and therefore are not very important for the purpose of model fitting and interpretation. Note that here, as well as for the results in the main text, the B term in the FIA is computed using Equation 34 rather than Equation 38 in order to avoid infinities (that for finite N can arise when the likelihood gradient is very small) and discontinuities (that for finite N can arise on the interior of the manifold, in proximity to the boundary, where the value of B goes from zero when $\hat{\vartheta}$ is in the interior to log(2) when $\hat{\vartheta}$ is exactly on the boundary).

Even though overall the agreement between the approximation is good, it is interesting to look more closely at where the approximation is poorest. The task type for which the discrepancies are the largest (both in absolute and relative terms) is the "robustness" type (fourth row in Figure S1). This discrepancy arises because the FIA hypotheses are not fully satisfied everywhere for one of the models. More specifically, the models in that task variant are a circular arc (the bottom model in Figure S1, third row) and a smaller circular arc, concentric with the first, with a straight segment attached to either side (the top model). The log-likelihood function for this second model is smooth only to first order, but its second derivative (and therefore its Fisher Information and its observed Fisher Information) is not continuous at the points where the circular arc is joined with the straight segments, locally breaking hypothesis number 3 in subsubsection A.1.1. Geometrically, this discontinuity is analogous to saying that the curvature of the manifold changes abruptly at the joints. It is likely that the FIA for a model with a smoother transition between the circular arc and the straight arms would have been even closer to the exact value for all points on the 2D plane (the data space). More generally, this line of reasoning suggests that it would be interesting to investigate the features of a model that affect the quality of the Fisher Information Approximation.

C.2 Subject demographics

For the generative task, after excluding subjects based on pre-registered performance thresholds as detailed above, we were left with data from 202 subjects. The age range was 18–68 (mean: 28.0, std: 9.5). Sex was reported as male for 129 subjects, female for 69, other for 4. Ethnicity was as follows: White 174, Asian 12, Hispanic or Latino 5, Other 4, Black or African American 3, American Indian or Alaskan Native 2, do not wish to respond 2. For the maximum-likelihood task, the total number of subjects after exclusions was 201, and the with the following demographics. Age: 18–87 (mean: 35.9, std: 12.8). Sex: male 124, female 75, other 1, prefer not to say 1. Ethnicity: White 157, Asian 25, Black or African American 8, Other 6, Hispanic or Latino 3, American Indian or Alaskan Native 1, do not wish to respond 1.

C.3 Posterior predictive checks

We performed a simple posterior predictive check (*28*) to ensure that the Bayesian hierarchical model described in the text captures the main pattern of behavior across our subjects. In Figure S2, the behavioral performance of the subjects is compared with its posterior predictive distribution under the model, for the case of the human subjects in the generative task. As can be seen from the figure, the performance of each subject is correctly captured by the model, across systematic differences between task types (with subjects performing better in the boundary task variant than the robustness task variant, for instance) as well as individual differences between subjects that performed the same task variant.

C.4 Details on raw estimated sensitivities

Table S1 reports the posterior mean and standard deviation of the population-level parameters entering the regression (Equation 106). Note that these are the raw parameters, not their normalized counterparts relative to the likelihood sensitivity as reported in the rest of the paper.

C.5 Uncertainty in subject-level sensitivities

Figure S3 illustrates the uncertainty in the estimate for the relative sensitivity of each subject. This uncertainty is typically small compared to between-subject variability of the sensitivity, which is therefore not a trivial consequence of the noise in the sensitivity estimation for individual subjects.

C.6 Lapse-rate analysis

By applying the model variant described in Section A.5.1, we were able to estimate a lapse rate for each subject simultaneously with the sensitivity parameters. The results are summarized in Figure S4, showing that although there is a substantial spread of lapse rates in the range 0–0.2, there is no clear relationship between lapse rates and sensitivity. The sensitivity parameters estimated with this extended model were qualitatively compatible with those presented everywhere else in the text.

C.7 Outcome of significance tests specified in the preregistration documents

C.7.1 Formal comparison between ideal observers

We compared the Bayesian hierarchical model described in section A.5 to a simpler model, where subjects were assumed to only be sensitive to likelihood differences, or in other words to choose $\mathcal{M}_1$ over $\mathcal{M}_2$ based only on which model was on average closer to the dot cloud constituting the stimulus on a given trial. Mathematically, this "likelihood-only" model was equivalent to fixing all β parameters to zero except for β_L in the model described in section A.5. All other details of the model were the same, and in particular the model still had a hierarchical structure with adaptive shrinkage (the subject-level parameters α and β_L were modeled as coming from Student T distributions controlled by population-level parameters). We compared the full model and the likelihood-only model on our human behavior data using the Widely Applicable Information Criterion (*29*). This comparison indicates strong evidence in favor of the full model not only in the generative task (Table S2), but also in the maximum-likelihood task (Table S3).

C.7.2 Other statistical tests

As described in the preregistration documents (*24–26*), in this work we have emphasized parameter estimation and information criteria-based model comparison over null hypothesis significance testing (see for instance (*27*), and (*28*) for a discussion and comparison of these ideas). However, for completeness, we report in Table S4 (1) the comparison between the Regions of Practical Equivalence (ROPE, (*28*)) and the 95% highest-density interval (HDI) for each population-level parameter, and (2) the "probability of direction" (*55*, *56*) for the same parameters (see below for more details on these methods). The ROPE-HDI tests highlight that the null value of zero sensitivity is not credible (rejected) for L, D and R, and neither rejected not accepted for V. The probability of direction is high for all parameters, including V, which has $PD = 0.97$ for the generative task and $PD = 0.99$ for the maximum-likelihood task. Overall, these analyses point to a significant sensitivity for all terms of the FIA in both experiments (generative and maximum-likelihood), with V having a more moderate effect size than the other terms.

Technical details on the ROPE-HDI comparison and on the Probability of Direction for sensitivity parameters Briefly, the ROPE for a parameter is the range around a null value for that parameter such that variations within this range would imply only a "negligible change" in the behavior of the model, if all other parameters were held at their null values. The HDI is the smallest interval that contains a certain probability mass for the posterior of that parameter. The ROPE-HDI comparison is based on the idea that if the bulk of the posterior distribution for that parameter (represented by the HDI) falls outside the ROPE, then the null value for that parameter can be considered not credible (rejected). On the other hand, if the bulk of the posterior for the parameter falls within the ROPE, the null value can be considered credible (accepted). Finally, if the posterior distribution has a partial overlap with the ROPE (neither mostly contained within it, nor mostly falling outside of it), then the test is inconclusive. Note that, just like frequentist null hypothesis significance testing procedures and unlike the information criterion approach used above, this

method depends on some arbitrary assumptions, namely the definition of the ROPE and the probability to use in computing the HDI.

In practice, as explained in more details in our preregistration documents (*24–26*), here we define, conventionally, the HDI as the smallest interval that contains 95% of the posterior. The ROPE is computed as follows. We start by defining a "negligible change" over the probability of the choice variable over the "main range" $[\mu_x - 2\sigma_x, \mu_x + 2\sigma_x]$ of one of the predictors in our model (L, D, B, V or R). In other words pick an interval of probabilities $[\pi_0 - \delta, \pi_0 + \delta]$ such that if the probability stays within $[\pi_0 - \delta, \pi_0 + \delta]$ when x varies over its typical range, then the probability is not meaningfully affected by x. Mathematically, if the probability of choosing one of the alternatives in the task is π and the log-odds is $\text{logit}(\pi) = \log(\pi/(1-\pi)) =$, then in a logistic regression setting

$$\text{logit}(\pi) = \alpha + \beta x \quad . \tag{122}$$

If $\pi_0 = \text{logit}^{-1}(\alpha)$, then the ROPE for β is defined as

$$\text{logit}(\pi_0 + \delta) = \alpha + \beta_+(\mu_x + 2\sigma_x) \tag{123}$$

$$\text{logit}(\pi_0 - \delta) = \alpha + \beta_-(\mu_x - 2\sigma_x) \tag{124}$$

so that

$$\beta_+ = \frac{\text{logit}(\pi_0 + \delta) - \alpha}{\mu_x + 2\sigma_x} \tag{125}$$

$$\beta_- = \frac{\text{logit}(\pi_0 - \delta) - \alpha}{\mu_x - 2\sigma_x} \tag{126}$$

In our case, assuming a negligible influence of the up/down bias (α in Equation 106), $\pi_0 = 0.5$ and therefore we can assume $\alpha = 0$. The definition of the ROPE further depends on the arbitrary choice of δ, and on the values of μ_x and σ_x. We choose $\delta = 0.025$, and we estimate μ_x and σ_x by generating 25000 experimental trials per task type (Dimensionality, Boundary, Volume, Robustness) and computing the empirical average and standard deviation of the predictors over that trial set. These numbers were all fixed at preregistration time (*24–26*).

References

11. V. Balasubramanian, Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions. *Neural Computation* **9**, 349–368 (1997).

15. E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Apr. 1, 2003), 753 pp., ISBN: 0-521-59271-2.

24. E. Piasini, V. Balasubramanian, J. I. Gold, *Preregistration Document* (`https://doi.org/10.17605/OSF.IO/2X9H6`).

25. E. Piasini, V. Balasubramanian, J. I. Gold, *Preregistration Document Addendum* (`https://doi.org/10.17605/OSF.IO/5HDQZ`).

26. E. Piasini, S. Liu, V. Balasubramanian, J. I. Gold, *Preregistration Document Addendum* (`https://doi.org/10.17605/OSF.IO/826JV`).

27. R. McElreath, *Statistical Rethinking* (CRC Press, 2016), ISBN: 978-1-4822-5344-3.

28. J. K. Kruschke, *Doing Bayesian Data Analysis* (Academic Press, ed. 2, 2015), ISBN: 978-0-12-405888-0.

29. A. Gelman *et al.*, *Bayesian Data Analysis* (CRC Press, ed. 3, 2014), ISBN: 978-1-4398-4095-5.

46. E. Piasini, V. Balasubramanian, J. I. Gold, presented at the The First International Symposium on AI and Neuroscience - ACAIN 2021.

47. M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1972), ISBN: 0-486-61272-4.

48. S.-i. Amari, H. Nagaoka, *Methods of Information Geometry*, trans. by D. Harada (American Mathematical Society, 2000), 206 pp., ISBN: 0-8218-4302-8.

49. J. Salvatier, T. V. Wiecki, C. Fonnesbeck, Probabilistic Programming in Python Using PyMC3. *PeerJ Computer Science* **2**, e55 (Apr. 2016).

50. M. D. Hoffman, A. Gelman, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623, (`http://jmlr.org/papers/v15/hoffman14a.html`) (2014).

51. M. Betancourt, *A Conceptual Introduction to Hamiltonian Monte Carlo*

52. S. development team, *Stan Modeling Language Users Guide, Version 2.31*.

53. F. Pedregosa *et al.*, Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830, ISSN: 1533-7928, (2023; `http://jmlr.org/papers/v12/pedregosa11a.html`) (2011).

54. R. Kumar, C. Carroll, A. Hartikainen, O. Martin, ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in Python. *Journal of Open Source Software* **4**, 1143 (2019).

55. D. Makowski, M. S. Ben-Shachar, D. Lüdecke, bayestestR: Describing Effects and Their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software* **4**, 1541, ISSN: 2475-9066, (2022; `https://joss.theoj.org/papers/10.21105/joss.01541`) (Aug. 13, 2019).

56. D. Makowski, M. S. Ben-Shachar, S. H. A. Chen, D. Lüdecke, Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology* **10**, ISSN: 1664-1078, (2022; `https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02767`) (2019).

57. A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner, Rank-Normalization, Folding, and Localization: An Improved $\hatR$ for Assessing Convergence of MCMC. *Bayesian Analysis* (2020).