



11-21-2022

How Occam's Razor Guides Human Inference

Eugenio Piasini

Shuze Liu

University of Pennsylvania, liushuze@sas.upenn.edu

Pratik Chaudhari

University of Pennsylvania, pratikac@seas.upenn.edu


Vijay Balasubramanian

University of Pennsylvania, vijay@physics.upenn.edu

Joshua I. Gold

University of Pennsylvania, jigold@penmedicine.upenn.edu

Follow this and additional works at: https://repository.upenn.edu/physics_papers

 Part of the [Cognition and Perception Commons](#), [Cognitive Science Commons](#), and the [Physics Commons](#)

Recommended Citation

Piasini, E., Liu, S., Chaudhari, P., Balasubramanian, V., & Gold, J. I. (2022). How Occam's Razor Guides Human Inference. Retrieved from https://repository.upenn.edu/physics_papers/662

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/physics_papers/662
For more information, please contact repository@pobox.upenn.edu.

How Occam's Razor Guides Human Inference

Abstract

Occam's razor is the principle stating that, all else being equal, simpler explanations for a set of observations are preferred over more complex ones. This idea is central to multiple formal theories of statistical model selection and is posited to play a role in human perception and decision-making, but a general, quantitative account of the specific nature and impact of complexity on human decision-making is still missing. Here we use preregistered experiments to show that, when faced with uncertain evidence, human subjects bias their decisions in favor of simpler explanations in a way that can be quantified precisely using the framework of Bayesian model selection. Specifically, these biases, which were also exhibited by artificial neural networks trained to optimize performance on comparable tasks, reflect an aversion to complex explanations (statistical models of data) that depends on specific geometrical features of those models, namely their dimensionality, boundaries, volume, and curvature. Moreover, the simplicity bias persists for human, but not artificial, subjects even for tasks for which the bias is maladaptive and can lower overall performance. 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.

Keywords

psychophysics, decision-making, probabilistic, Bayesian, normative modeling

Disciplines

Cognition and Perception | Cognitive Science | Physics

How Occam's razor guides human inference

Eugenio Piasini^{✉1,2*}, Shuze Liu^{2*}, Pratik Chaudhari², Vijay Balasubramanian^{2†}, Joshua I. Gold^{2†}

¹International School for Advanced Studies (SISSA), Trieste, Italy

²University of Pennsylvania, Philadelphia PA

* Contributed equally

† Contributed equally

✉ epiasini@sissa.it

Occam's razor is the principle stating that, all else being equal, simpler explanations for a set of observations are preferred over more complex ones¹. This idea is central to multiple formal theories of statistical model selection²⁻⁵ and is posited to play a role in human perception and decision-making⁶, but a general, quantitative account of the specific nature and impact of complexity on human decision-making is still missing. Here we use preregistered experiments to show that, when faced with uncertain evidence, human subjects bias their decisions in favor of simpler explanations in a way that can be quantified precisely using the framework of Bayesian model selection. Specifically, these biases, which were also exhibited by artificial neural networks trained to optimize performance on comparable tasks, reflect an aversion to complex explanations (statistical models of data) that depends on specific geometrical features of those models, namely their dimensionality, boundaries, volume, and curvature. Moreover, the simplicity bias persists for human, but not artificial, subjects even for tasks for which the bias is maladaptive and can lower overall performance. 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 and model selection

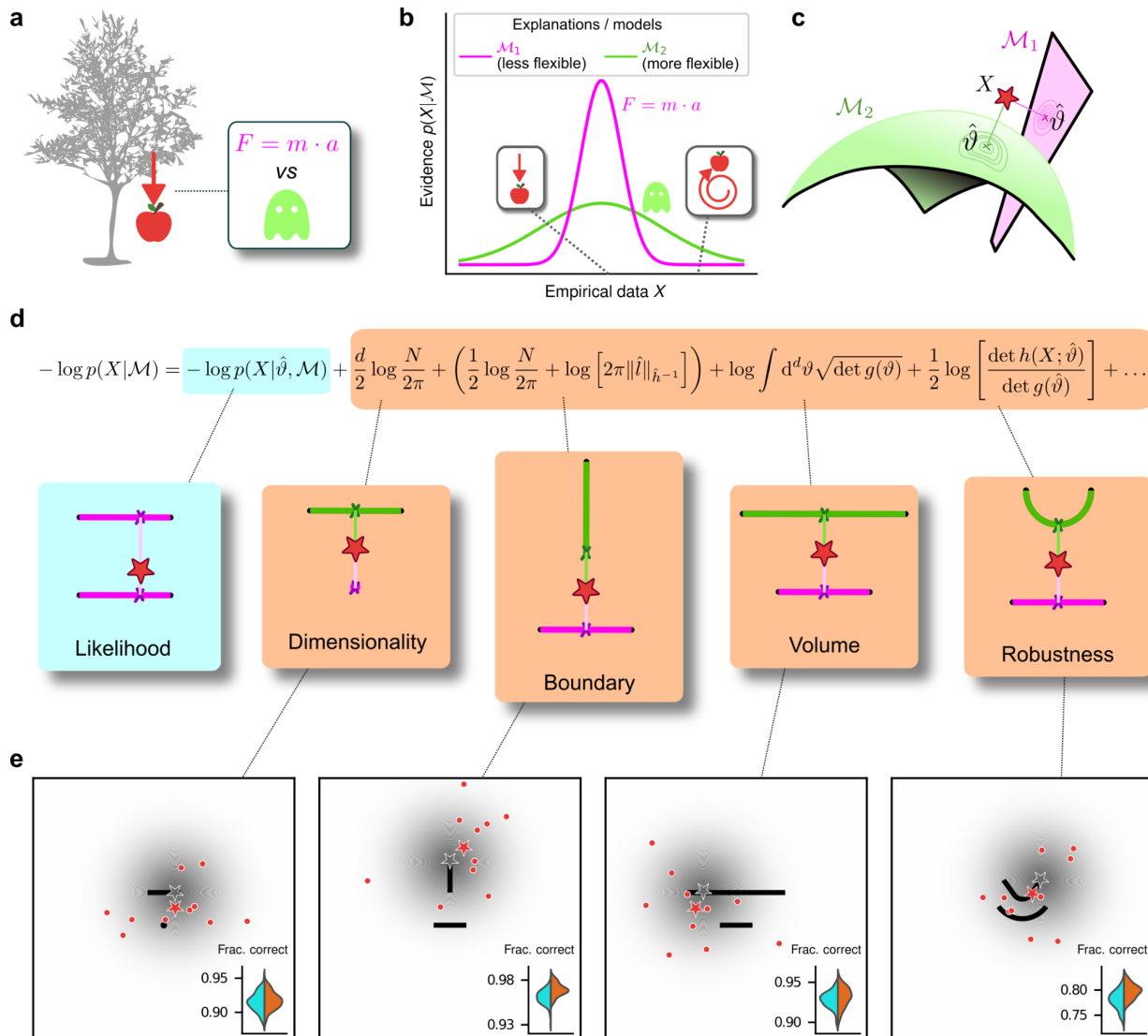


Figure 1: Occam's razor, Bayesian model selection, and testing their roles in machine and human inference.

a: Occam's razor prescribes a bias against complex 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 we 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 panel a. Classical mechanics (pink) can explain a narrower range of observations than the ghost (green), which is a valid explanation for just about any conceivable phenomenon (e.g., both a falling and spinning-upward trajectory, as in the insets). In the absence of further evidence, Occam's razor suggests that the simpler model (classical mechanics) is preferred, because its hypothesis space is more concentrated around the sparse, noisy evidence 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:** Large- N expansion of the log evidence of a model M . $\hat{\vartheta}$ is the

maximum-likelihood point on model M for data X , N is the number of observations, d is the number of parameters of the model, l is the likelihood gradient, h is the observed Fisher information, and g is the expected Fisher information. The ellipsis collects terms that get smaller as N grows. Each term of the expansion represents a distinct geometrical feature of the model: dimensionality penalizes models with many parameters; boundary (a novel contribution of this work) penalizes models for which $\hat{\vartheta}$ is on their boundary; volume counts the number of distinguishable probability distributions contained in M ; and robustness captures the shape (curvature) of M near $\hat{\vartheta}$. **e:** Psychophysical task with variants designed to probe each of the geometrical features in **d**. On each trial, a random location on one of the models is selected (gray star), and data (red dots) are sampled from a Gaussian centered around that point (gray shading). Subjects see the models (black lines) and the data (red dots) and have to choose which model is best for the data. Red star (not shown to the subjects): empirical centroid of the data, by analogy with **c**. In this task, the maximum-likelihood point can be found by projecting the empirical centroid onto one of the models. 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). For this task, the advantage conferred by the simplicity biases is very small.

Making a decision based on uncertain evidence often amounts to choosing between alternative, plausible explanations for noisy and sparse data. When evaluating such competing explanations, Occam's razor posits that we should consider not just how well they account for the actual observations, but also the extent to which they may be overly flexible in accounting for many diverse sets of potential observations (such as "a ghost did it!"; Figure 1a). In cognitive science, simplicity, or parsimony, has long been proposed as an organizing principle in sensory perception⁶, from the early concept of Prägnanz in Gestalt psychology⁷, to a number of minimum principles for vision⁸, to investigations based on information-theoretical approaches⁹. However, despite evidence that such a preference for simplicity exists in various forms of human decision-making¹⁰⁻¹⁴, we lack a principled understanding of what, exactly, defines the complexity of an alternative explanation that should be avoided, and by how much complexity should (and does) trade off against the ability to account for observations when we make decisions.

To provide this understanding, we turn to an approach based on Bayesian statistics, which allows us to measure the complexity of an explanation for data on a universal scale. Our process is formalized as a model-selection problem: given a set X of N observations and a set of possible statistical models $\{M_1, M_2, \dots\}$, we seek the model 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 against goodness-of-fit to maximize decision accuracy, typically because the increased flexibility provided by increased complexity tends to cause errors by overfitting noise in the observations^{2,3,15}.

Specifically, according to this framework models should be compared based on their evidence or marginal likelihood $p(X|M) = \int d\vartheta w(\vartheta) p(X|M, \vartheta)$, where ϑ represents model parameters

and $w(\vartheta)$ their associated prior (Figure 1b). Under mild regularity assumptions and for large N , the (log) evidence can be written as the sum of the maximum log likelihood of M and several penalty factors, which can be interpreted as a measure of model complexity^{15,16}. This approach, called the Fisher Information Approximation (FIA), generalizes the well-known Bayesian Information Criterion (BIC) for model selection^{17,18}. If the prior $w(\vartheta)$ is taken to be uninformative¹⁹, each penalty factor can be shown to capture a distinct geometric property of the model¹⁶, including dimensionality (number of parameters), boundary (a novel term, detailed below), volume, and shape (Figure 1c). Similar quantitative definitions of statistical model complexity or model selection prescriptions can be obtained with different theoretical approaches, such as the Minimum Description Length^{20–22}, Minimum Message Length²³, and Predictive Information²⁴ frameworks, testifying to the generality of this approach.

Measuring the simplicity bias in simple decisions

We designed a simple decision-making task to relate the FIA complexity terms to the biases exhibited by artificial and human decision-makers. On each trial, $N=10$ 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, and the location of the center of the Normal distribution within the selected shape was also sampled uniformly at random, in a way that did not depend on the model parametrization, by using Jeffrey’s prior¹⁹. Given the observations, the subjects decided which shape (model) was more likely to contain the center of the generative distribution. We designed four task variants, each conceived to probe primarily one of the distinct geometrical features that are penalized in Bayesian model selection (Figure 1d and e).

A key feature of the task is that the observations, and their empirical centroid, tended not to fall exactly on one of the two alternative models. Accordingly, the maximum-likelihood projection of the data often fell on the boundary of one of the models, even when the data were sampled from that model. These conditions are common in the real world (as implied by the “all models are wrong to some degree” mantra) but pose a challenge for the FIA and related model-selection approaches, which typically assume that the maximum likelihood solution is in the interior of the parameter space of a given model¹⁶. To overcome this challenge, we extended the FIA to deal with the simple case of a linear boundary in parameter space (see Appendix). When the maximum-likelihood solution is on such a boundary, an additional penalty term appears in the FIA, which we denote “boundary” (Figure 1d).

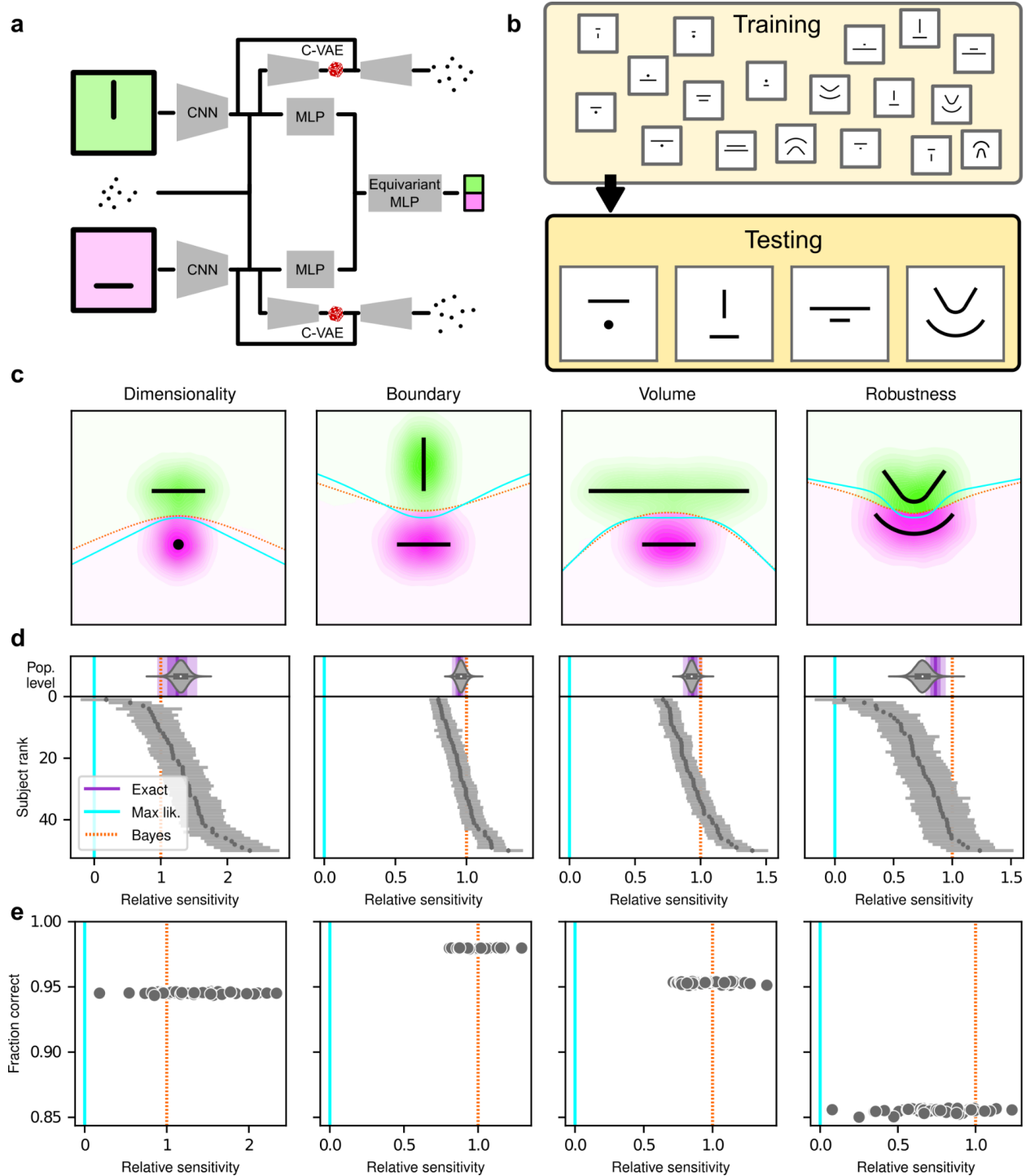


Figure 2: Simplicity biases in deep neural networks

a: A novel deep neural-network architecture for statistical model selection. The model takes as inputs two images, 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 model-selection task (Figure 1e), including systematically varying model length or curvature, then tested using the same configurations as for the human tests. **c:**

*Summary of model 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. **d**: Network behavior analyzed via hierarchical logistic regression, using the terms of the FIA (the features that determine model complexity) as predictors. Each fitted bias 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 only on the model that was the closest to the data) would have a relative sensitivity of zero (solid cyan lines). Top, gray: population-level estimates. 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., obtained from a simulation. Bottom: individual network-level sensitivity estimates for a population of 50 randomly initialized neural networks. **e**: Accuracy as a function of relative sensitivity of individual networks (points) for each term (columns and lines are as in c and d).*

To show that these complexity terms are not just abstractions of the FIA framework but instead are tangible, learnable quantities that impact performance, we designed a novel artificial neural network architecture that could perform statistical model selection, in a form applicable to the task described above (Figure 2a,b). On each trial, the network took as input two images representing the models to be compared, and a set of coordinates representing the data point. The output of the network was a decision between the two models, encoded as a softmax vector. We analyzed 50 instances of the deep network that differed only in the random initialization of their weights and in the examples seen during training.

After training, the networks' choices were consistent with having learned decision boundaries that were close to those of an ideal Bayesian observer (Figure 2c). These decision boundaries reflected tradeoffs between simplicity and goodness-of-fit that also were close to optimal, for each of the four complexity terms we tested (dimensionality, boundary, volume, and curvature; Figure 2d). These simplicity biases varied slightly in magnitude across the different networks, but this variability was not related systematically to any differences in the generally high accuracy rates for each condition (Figure 2e; posterior mean \pm st. dev. of Pearson correlation coefficient between accuracy and $|1-\beta|$, where β is the sensitivity: dimensionality, -0.11 ± 0.10 ; boundary, 0.07 ± 0.10 ; volume, -0.16 ± 0.11 ; robustness, -0.18 ± 0.12). This result implies that the networks were all trained comparably well, and some variability in bias magnitude occurred because of the stochastically generated stimuli used for training and testing. Overall, 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^{25–27}. 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 biases.

Humans are sensitive to the geometric complexity of statistical models

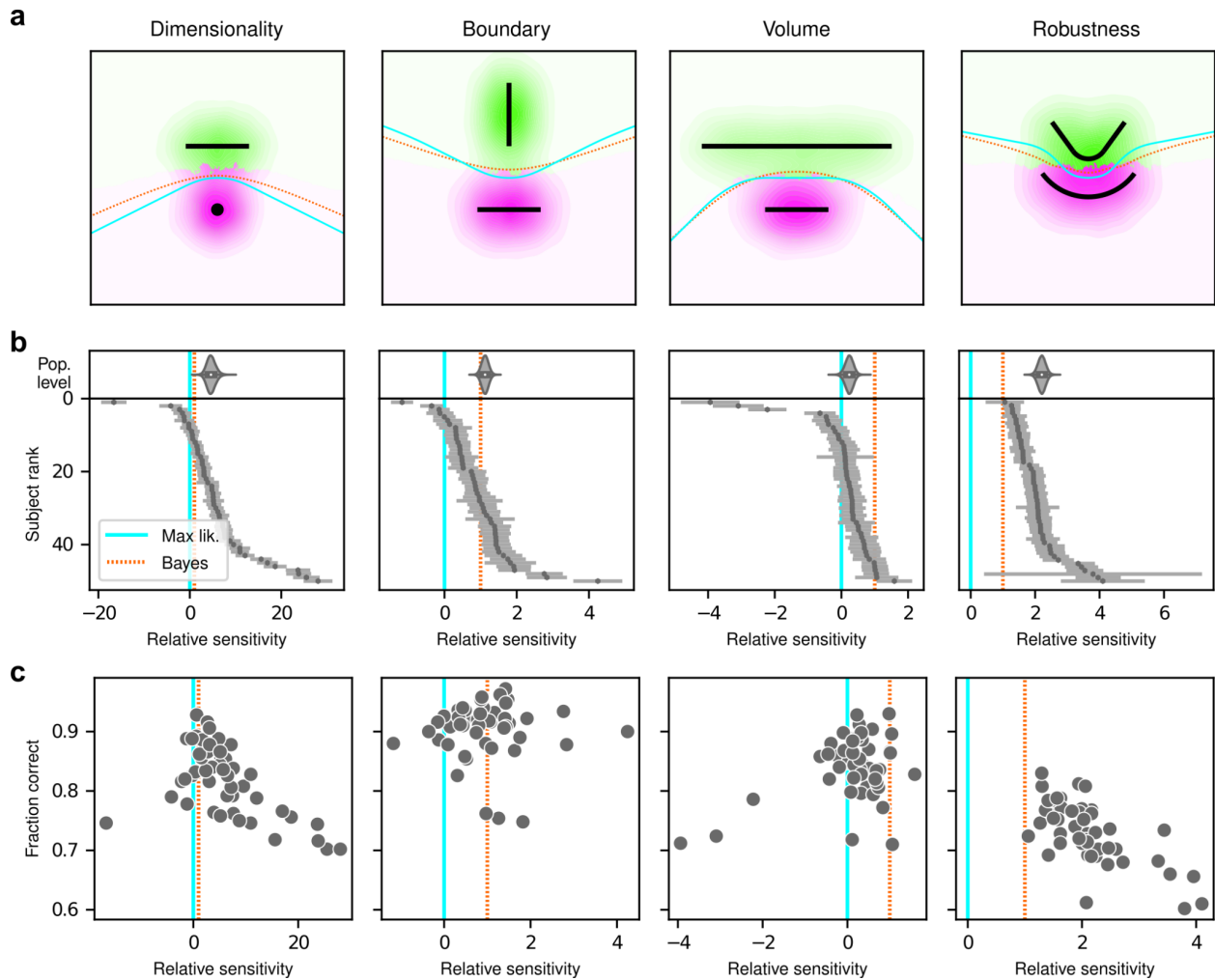


Figure 3: Simplicity biases in human behavior

a: Summary of human-subject behavior (plotted as in Figure 2c). 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 be biased towards the simpler model, particularly near the center of the screen. 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:** Estimated relative sensitivity to geometrical features characterizing model complexity (plotted as in Figure 2d), measured for subjects who performed the task variant illustrated directly above in panel 2a. Relative sensitivity is defined as sensitivity to a model feature divided by sensitivity to likelihood (see Figure 2). Top, gray: Population-level estimates. Bottom: subject-level estimates. Solid cyan lines: relative sensitivity of a maximum-likelihood observer. Orange dashed lines: relative sensitivity of an ideal Bayesian

observer. c: Accuracy as a function of relative sensitivity of individual networks (points) for each term (columns and lines are as in a and b).

To assess the relevance of these simplicity biases to human decision-making, we conducted an on-line study using the crowdsourcing platforms Prolific and Pavlovia. Following our preregistered approaches^{28,29}, 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). We used hierarchical Bayesian modeling to measure the sensitivity of each subject to model likelihood (distance from the data to a given model) and to each of the geometrical features characterizing model complexity. Just like for the artificial neural networks, this approach allowed us to define the relative sensitivity to each feature, by dividing the sensitivity to that feature by the subject's sensitivity to the likelihood.

The human subjects exhibited all four forms of simplicity bias, despite substantial individual variability that was greater than that found for the ANNs (note the different axis scales when comparing Figures 2d and 3b). Specifically, the estimated normalized population-level sensitivity for human subjects (posterior mean \pm st. dev.) 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. Formal model comparison (WAIC; see Appendix) confirmed that their behavior was better described by taking into account the geometric penalties described 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 broad range of individual variability also highlighted the importance of appropriately tuned (i.e., close to Bayesian) simplicity biases, because accuracy tended to decline for subjects with biases further away from the Bayesian predictions (Figure 3c; posterior mean \pm st. dev. of Pearson correlation coefficient between accuracy and $|1-\beta|$, where β is the sensitivity: dimensionality, -0.75 ± 0.02 ; boundary, -0.12 ± 0.10 ; volume, -0.42 ± 0.06 ; robustness, -0.58 ± 0.09). Overall, these results show that the intuitive aversion of human subjects for complex explanations for empirical data can be quantified precisely, in terms of the geometrical features identified in a Bayesian model-selection framework.

Robustness of simplicity biases to different instructions

Penalizing complex models by the appropriate amount is optimal in model selection, but in practice for our task the expected performance advantage from doing so is minimal. In simulations, the difference in performance between ideal observers that penalized model complexity according to the FIA and simulated observers that only used 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. Moreover, trial-by-trial feedback was not provided to the subjects. It is therefore unlikely that the task itself could provide sufficient incentive or information for the human subjects to learn to penalize complex models by adaptively optimizing their performance as they performed the task. We thus sought to determine if and how behavior depended on the specific form of our task instructions.

We compared performance for two different data sets collected using identical task conditions but different instruction sets. The data depicted in Figure 3 were collected using instructions that were formulated to mirror the Bayesian model-selection problem. Specifically, those instructions 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, respectively. In contrast, the instructions (and brief training block, with feedback) for the second data set asked 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.

Subject behavior was similar for both tasks, suggesting 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). To confirm that the difference between the two tasks was in principle learnable, we trained the deep neural networks on the maximum-likelihood task. In stark contrast to the human data, 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, 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 simplicity biases). Because the maximum-likelihood solutions are deterministic (they depend only on which model the data centroid is closest to, and thus the decision boundary is infinitely steep) and the generative solutions are not (they depend probabilistically on the likelihood and bias terms, and thus the decision boundary is not infinitely steep), 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. Taken together, these results imply that human decision-making has a natural tendency to follow principles of optimal model selection, even when those principles are not instructed or incentivized.

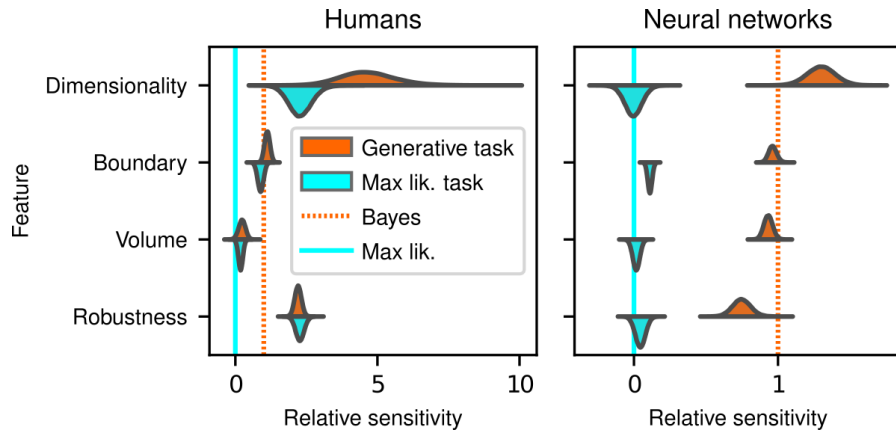


Figure 4: Simplicity biases in a task where penalizing complex models is suboptimal

a: Relative sensitivity of human subjects to the geometric complexity terms (population-level estimates, as in Figure 3b, 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 3b, 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 2d, top).

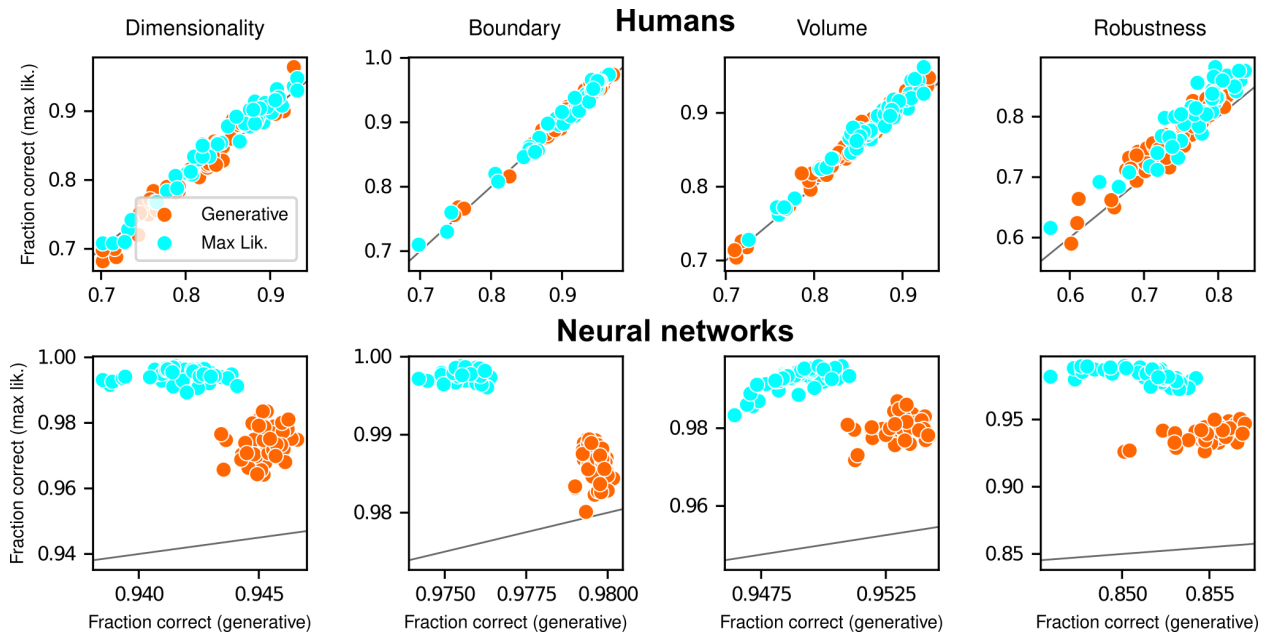


Figure 5: Summary of performance accuracy for ANNs and humans

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 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).

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 philosophy¹, psychology^{6,9}, statistics^{2,3,15,21,23}, and more recently machine learning^{25,26}. Accordingly, multiple studies have identified biases towards simplicity in human cognition^{10,12,13}, such as a tendency to prefer smoother (simpler) curves as the inferred, latent source of noisy observed data^{11,14}. However, the quantitative form and magnitude of such bias have never been identified. In this work, we showed that the bias is closely related to a specific mathematical formulation of Occam's razor, situated at the convergence of Bayesian model selection and information theory. 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 FIA in Bayesian model selection that accounts for the possibility that the best model is on the boundary of the model family. This boundary 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 data³⁰. 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 placing boundaries on such a set of possibilities, which tend to increase local complexity (because they tend to reduce the concentration of local hypotheses; see Figure 1b).

The second contribution of this work relates to ANNs: these networks can learn to use or ignore the simplicity biases in an optimal way (i.e., according to the magnitudes prescribed by the theory), depending on how they are trained. On the one hand, these results do not seem particularly surprising, because ANNs (and deep networks in particular) are powerful function approximators that perform well in practice on a vast range of inference tasks³¹. Accordingly, our ANNs trained with respect to the true generative solutions were able to make effective decisions, including simplicity biases, about the generative source of a given set of observations, whereas our ANNs trained with respect to maximum-likelihood solutions were able to make effective decisions, without simplicity biases, 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 for the presence or absence of these kinds of simplicity biases 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 biases when making decisions, and unlike ANNs these biases 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 full Bayesian solution (i.e., one that includes the simplicity biases) and a maximum-likelihood solution (i.e., one that does not include the simplicity biases) 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 biases as prescribed by the Bayesian solution. We hope 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 biases 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), and 3) different for distinct features and different subjects. 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 on 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^{32,33}. 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 common class of models and proposed instead the use of data-dependent priors^{34,35}. The two methods lead to different results, especially in the data-limited regime³⁶. 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.

Acknowledgements

We thank Kamesh Krishnamurthy for discussions. We acknowledge the financial support of R01 NS113241 (EP), R01 EB026945 (JIG and VB), and a hardware grant from the NVIDIA Corporation (EP). We acknowledge the HPC Collaboration Agreement between SISSA and CINECA for granting 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.

Bibliography

1. Baker, A. Simplicity. in *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research Lab, Stanford University, 2022).
2. Jeffreys, H. *Theory of probability*. (Clarendon Press, 1939).
3. Gull, S. F. Bayesian Inductive Inference and Maximum Entropy. in *Maximum-Entropy and Bayesian Methods in Science and Engineering: Foundations* (eds. Erickson, G. J. & Smith, C. R.) 53–74 (Springer Netherlands, 1988). doi:10.1007/978-94-009-3049-0_4.
4. Grünwald, P. Model Selection Based on Minimum Description Length. *J. Math. Psychol.* **44**, 133–152 (2000).
5. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer-Verlag, 2009).
6. Feldman, J. The simplicity principle in perception and cognition: The simplicity principle. *Wiley Interdiscip. Rev. Cogn. Sci.* **7**, 330–340 (2016).
7. Koffka, K. *Principles of Gestalt psychology*. (Mimesis international, 2014).
8. Hatfield, G. The status of the minimum principle in the theoretical analysis of visual perception. *Psychol. Bull.* **97**, 155 (1985).
9. Chater, N. & Vitányi, P. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* **7**, 19–22 (2003).
10. Pothos, E. M. & Chater, N. A simplicity principle in unsupervised human categorization. *Cogn. Sci.* **26**, 303–343 (2002).
11. Genewein, T. & Braun, D. A. Occam's Razor in sensorimotor learning. *Proc. R. Soc. B Biol. Sci.* **281**, 20132952 (2014).
12. Gershman, S. & Niv, Y. Perceptual estimation obeys Occam's razor. *Front. Psychol.* **4**,

- (2013).
13. Little, D. R. B. & Shiffrin, R. Simplicity Bias in the Estimation of Causal Functions. *Proc. Annu. Meet. Cogn. Sci. Soc.* **31**, (2009).
 14. Johnson, S., Jin, A. & Keil, F. Simplicity and Goodness-of-Fit in Explanation: The Case of Intuitive Curve-Fitting. in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36(36) (2014).
 15. MacKay, D. J. C. Bayesian Interpolation. *Neural Comput.* **4**, 415–447 (1992).
 16. Balasubramanian, V. Statistical Inference, Occam’s Razor, and Statistical Mechanics on the Space of Probability Distributions. *Neural Comput.* **9**, 349–368 (1997).
 17. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **6**, 461–464 (1978).
 18. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2012).
 19. Jaynes, E. T. *Probability Theory: The Logic of Science*. (Cambridge University Press, 2003).
 20. Rissanen, J. J. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **42**, 40–47 (1996).
 21. Grünwald, P. D. *The Minimum Description Length Principle*. (MIT press, 2007).
 22. Lanterman, A. D. Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection. (2000).
 23. Wallace, C. S. *Statistical and inductive inference by minimum message length*. (Springer, 2005).
 24. Bialek, W., Nemenman, I. & Tishby, N. Predictability, Complexity and Learning. *Neural Comput.* 2409–2463 (2001) doi:10.1162/089976601753195969.
 25. De Palma, G., Kiani, B. & Lloyd, S. Random deep neural networks are biased towards simple functions. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).

26. Valle-Perez, G., Camargo, C. Q. & Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. in *International Conference on Learning Representations* (2019).
27. Zhang, S., Reid, I., Pérez, G. V. & Louis, A. Why flatness does and does not correlate with generalization for deep neural networks. (2021).
28. Piasini, Eugenio. Effect of geometric complexity on intuitive model selection - experiment on Pavlovia. (2021) doi:10.17605/OSF.IO/2X9H6.
29. Piasini, Eugenio. Addendum #2 to osf.io/2x9h6 - Effect of geometric complexity on intuitive model selection - rounded task. (2022) doi:10.17605/OSF.IO/5HDQZ.
30. Transtrum, M. K. *et al.* Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.* **143**, 010901 (2015).
31. Bengio, Y., Lecun, Y. & Hinton, G. Deep learning for AI. *Commun. ACM* **64**, 58–65 (2021).
32. Tavoni, G., Balasubramanian, V. & Gold, J. I. What is optimal in optimal inference? *Curr. Opin. Behav. Sci.* **29**, 117–126 (2019).
33. Tavoni, G., Doi, T., Pizzica, C., Balasubramanian, V. & Gold, J. I. Human inference reflects a normative balance of complexity and accuracy. *Nat. Hum. Behav.* **6**, 1153–1168 (2022).
34. Mattingly, H. H., Transtrum, M. K., Abbott, M. C. & Machta, B. B. Maximizing the information learned from finite data selects a simple model. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1760–1765 (2018).
35. Quinn, K. N., Abbott, M. C., Transtrum, M. K., Machta, B. B. & Sethna, J. P. Information geometry for multiparameter models: New perspectives on the origin of simplicity. Preprint at <https://doi.org/10.48550/arXiv.2111.07176> (2022).
36. Abbott, M. C. & Machta, B. B. Far from Asymptopia. Preprint at <http://arxiv.org/abs/2205.03343> (2022).