Published in final edited form as:

Nat Hum Behav. 2022 August ; 6(8): 1153–1168. doi:10.1038/s41562-022-01357-z.

Human inference reflects a normative balance of complexity and accuracy

Gaia Tavoni[1,*], **Takahiro Doi**[2], **Chris Pizzica**[3], **Vijay Balasubramanian**[3,4,**], **Joshua I. Gold**[3,**]

[1]Department of Neuroscience, Washington University in St. Louis, MO 63110, USA

[2]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

We must often infer latent properties of the world from noisy and changing observations. Complex, probabilistic approaches to this challenge such as Bayesian inference are accurate but cognitively demanding, relying on extensive working memory and adaptive processing. Simple heuristics are easy to implement but may be less accurate. What is the appropriate balance between complexity and accuracy? Here, we model a hierarchy of strategies of variable complexity and find a power law of diminishing returns: increasing complexity gives progressively smaller gains in accuracy. The rate of diminishing returns depends systematically on the statistical uncertainty in the world, such that complex strategies do not provide substantial benefits over simple ones when uncertainty is either too high or too low. In between, there is a complexity dividend. In two psychophysical experiments, we confirm specific model predictions about how working memory and adaptivity should be modulated by uncertainty.

Keywords

adaptivity; working memory; learning; on-line inference; change-point processes; complexity; diminishing returns

[*]Corresponding author (gaia.tavoni@wustl.edu).

Author Contributions

G.T., V.B., J.G. developed the ideas and wrote the paper; all the authors designed the psychophysics experiments; T.D., C.P. performed the experiments; G.T. developed the theory and analysed the data.

[**]These authors jointly supervised this work

Code Availability Statement

Data were collected using:

Psiturk (v2.3.0) [66]: http://psiturk.org/, https://github.com/NYUCCL/psiTurk

jsPsych (v6.0) [67]: https://www.jspsych.org/, https://github.com/jspsych/jsPsych/

chart.js (v2.8.0): https://www.chartjs.org/, https://github.com/chartjs/Chart.js

papaparse.js (v5.0): https://www.papaparse.com/, https://github.com/mholt/PapaParse

Custom code that supports the findings of this study is available from the corresponding author upon request.

Competing Interests Statement

The authors declare no competing interests.

Introduction

We often use sequences of sensory observations to arrive at judgements about current and future states of the world. This kind of sequential inference can be modeled in ways that differ widely in form, accuracy, and complexity [1–8]. This diversity leaves open basic questions about the relevance of these models to cognition. The goal of the present study was to identify fundamental principles governing when particular cognitive operations are important to perform inferences that are both effective and efficient; that is, sufficiently accurate but also consistent with computational and information-gathering constraints that lead to bounded rationality [6, 9, 10]. We reasoned that computational complexity in models of inference can represent a cognitive cost (e.g., in terms of the amount of working memory and the degree of adaptivity) that under some conditions might outweigh the benefits of potential gains in accuracy.

To test this idea, we constructed a hierarchical class of inference models that can be rated in terms of both accuracy and computational complexity. At the top of the hierarchy is Bayesian inference, which uses a probabilistic framework to combine both noise and volatility into a strategy that makes the most accurate inferences about current and future states of the world based on all previous observations [3, 11–13]. This model provides a maximum-accuracy benchmark for our analyses, but it also can require virtually unlimited computational resources and thus provides a maximum-complexity benchmark as well. We then extended previous work showing that exact Bayesian inference can be approximated using weighted combinations of simpler computational units [7], by constructing two nested families of models that generated increasingly simple approximations to Bayesian inference (Fig. 1). These models have progressively lower adaptivity and memory requirements, along with lower accuracy and complexity, along the hierarchy.

We examined the performance of these nested models on a general class of tasks that require inference in the presence of noise (stochastic fluctuations in the observations) and volatility (fundamental changes in the structure of the environment) [14–16]. We used several complementary approaches to identify two fundamental principles. The first principle is a law of diminishing returns, whereby gains in accuracy become progressively smaller with increasing complexity, regardless of the amount of uncertainty in the environment. The second principle is a non-monotonic relationship between uncertainty and the complexity of the most efficient model: simple models are the most efficient when uncertainty is very high or low, whereas more complex models are useful at intermediate levels of uncertainty, when cues are both identifiable and helpful. We then used two behavioral experiments to show that these principles apply to human behavior. Overall, these results provide new insights into the cognitive processes that may be differentially engaged to perform efficient and effective inference under different conditions.

Results

A hierarchy of inference strategies

Numerous models have been proposed to solve inference problems in dynamic, noisy environments. These models range from Bayesian probabilistic strategies to simpler, heuristic update processes like non-adaptive delta rules that are often described as implementing a "model-free" form of learning [7, 12, 14, 16, 17]. Here we show that many of these models can be described parsimoniously in terms of two partially overlapping nested families that represent systematic simplifications of the Bayesian ideal observer (Fig. 1). The two families each include a progression from adaptive, memory-dependent learning to inflexible processes with no memory or learning.

In general, a system for online inference aims to identify the current source of observations (estimation) or to predict the next source (prediction), in the presence of noise and unsignalled change-points, given all past and present data $x_{1:t} = \{x_1, \ldots, x_t\}$. We consider a standard problem in which change-points in the source occur according to a Bernoulli process with a fixed probability h (volatility), and the source, characterized by a single number μ_t at a time t, generates observations with Gaussian variability (Fig. 2) [7, 14]. Noise in this generative process is measured by the ratio R between the standard deviation of the observations with respect to their source and the standard deviation of the sources across many change-points.

In this setting, the Bayesian ideal observer estimates the full distribution of the source in terms of two quantities: (1) the conditional probability $p(r_t|x_{1:t})$ of the run-length r_t, which is the number of time steps elapsed at time t since the last inferred change-point in the source; and (2) the probability $p(\mu_t|r_t)$ that the source is μ_t given data observed over just the run-length r_t. By multiplying these probabilities and summing over possible run-lengths, we can compute the probability that the source is μ_t given all the data:

$$p(\mu_t \mid x_{1:t}) = \sum_{r_t = 1}^{t} p(\mu_t \mid r_t) p(r_t \mid x_{1:t}) \,. \tag{1}$$

The Bayesian model computes $p(\mu_t|r_t)$ and $p(r_t|x_{1:t})$ exactly [3, 12, 13]. The optimal Bayesian estimate of the source, $\hat{\mu}_t$, is then simply the expected value of μ_t in the conditional distribution (Eq. 1). To optimally predict the next source, $\hat{\mu}_{t+1}$, given this estimate, we must include the expected rate of change-points so that

$$\hat{\mu}_{t+1} = h\bar{\mu} + (1 - h)\hat{\mu}_t \tag{2}$$

where $\bar{\mu}$ is the asymptotic average value of the source (Fig. 2).

Thus, the Bayesian model balances prior belief against evidence integrated over temporal windows of all possible lengths, with each window weighted by the likelihood that the latent variable has been stable over that duration. This model minimizes mean squared error but is computationally expensive: the time needed to make an estimate or a prediction grows linearly with t, because the model requires a sum over possible run-lengths (Eq. 1; [7]). In cognitive terms, exact Bayesian inference requires working memory to increase with time. This computation is systematically simplified in a hierarchically organized set of models depicted in Fig. 1 (see Methods for model details). The Mixture of Sliding Windows truncates the Bayesian model to a finite number of windows of fixed durations.

The Delta Rules instead weigh past observations exponentially (examples of window and exponential integration kernels are depicted as grey areas in Fig. 1). The inferences from different Sliding Windows or Delta Rules are weighted optimally in the Mixture models. The Memoryless model simply combines current evidence with the prior and maintains no working memory (Dirac delta kernel). The Prior model sticks to the prior belief regardless of evidence. The Evidence model follows the current evidence and ignores both prior beliefs and past evidence.

Each model estimates the source of observations and uses it to make predictions by computing a function that depends on: (1) observations ($x_{1:t}$), (2) fixed parameters of the environment (the average source $\bar{\mu}$, the volatility h, and the noise level R), and (3) model-dependent "meta-parameters" (the run-lengths and the learning rates). The simplifications giving rise to the two families of strategies from the full Bayesian model can be interpreted in terms of progressive reductions of cognitive demand (right arrow in Fig. 1).

Adaptive models reduce to calibrated simpler strategies

When probability distributions are inferred from limited samples, complex models can generalize worse to new data than simple models [18, 19]. How to identify the model that best trades off fitting accuracy and generalization performance is the subject of a vast literature on model selection. Here, we address a different problem. Even if a complex model has a lower prediction error, is the increase in accuracy relative to a simple model "worth the effort"? When and how can complex models be reduced to simpler ones?

To gain insights into these questions, we characterized the structure of our hierarchically organized models in terms of their Redundancy and Alignment. Redundancy is defined for each model as the ratio between the maximum and minimum eigenvalues of the Hessian matrix (H) evaluated at the minimum of the model's error E. The eigenvalues of H indicate how much the error E increases when moving away from the minimum error in parameter space in the direction of the eigenvectors of H [20]. Thus, a high Redundancy indicates that the full model is reducible to a simpler one through a procedure that removes the least-relevant degree of freedom (e.g., the "manifold boundary approximation" method of [21]). Alignment is defined for each model as the (normalized) angle between the least-relevant eigenvector and the direction of the parameter change that would produce the next-simpler model in the hierarchy with optimal parameters (Fig. 3A). A high Alignment, associated with high Redundancy, is interesting because it indicates that removing the irrelevant degree of freedom from the more-complex model generates an optimal (or nearly optimal) model that has lower complexity and automatically minimizes prediction error, without need for parameter recalibration. A high Alignment is not a trivial or necessary consequence of a high model Redundancy.

Using this formalism, we found that the adaptive Mixture models tend to be redundant in two opposite conditions: (1) when noise and volatility are low so that inference is easy and complex strategies are unnecessary; and (2) when noise or volatility are high so that inference is difficult, making complex strategies ineffective (Fig. 3B). Furthermore, when the models are redundant, their Alignment is also high (Fig. 3CD). This relationship, which is particularly strong for Delta-Rule models, implies that the brain could, in principle, use local estimates of performance gradients [22] to identify and eliminate a currently irrelevant parameter and automatically generate a calibrated, simpler strategy.

A power law of diminishing returns

The results above suggest that complex solutions to on-line inference problems may not always be worth the effort. To investigate the exact scaling of model performance with computational costs, we evaluated the algorithmic complexity of the models in Fig. 1 in terms of the average number of computational and memory operations required to implement them (Fig. 4A, Supplementary Table 1). For the non-parametric Bayesian model, this cost grows linearly with the number of observations because the entire past provides a probabilistic context for each prediction or estimate, and thus the algorithmic complexity diverges to infinity. The other models are parametric and have constant complexity that is partly related to the number of free parameters. The values we computed are qualitatively consistent with other notions of complexity from Bayesian model selection, information geometry, and data compression, for which the leading-order term of model complexity grows with the number of parameters, and lower-order terms depend on the model's functional form [18, 19, 23]. However, unlike those notions of complexity, algorithmic complexity can be applied readily to the kinds of deterministic models considered here.

We related the algorithmic complexity of each model to its performance relative to the Bayesian benchmark and found a power-law of diminishing returns: increasing the complexity of a model gives progressively smaller improvements in prediction (Fig. 4B–F and Supplementary Fig. 2B–F). Prediction accuracy is maximized by using the most complex model. However, at both high and low noise (large and small R), low complexity models are already within 10% of the Bayesian optimum (light blue and dark blue lines in Fig. 4E). Likewise, when volatility is large, low complexity strategies perform almost as well as the full Bayesian model (red and brown lines in Fig. 4F). These results suggest that sophisticated inference procedures are only useful in a narrow range of conditions with an intermediate amount of noise and low underlying volatility. These conclusions are robust across a very wide range of thresholds for "good enough" performance (shifting black threshold lines in Fig. 4EF) and to very different model implementations (we recapitulated the results from Fig. 4 using neural-network implementations [24]; Supplementary Figs. 2 and 3D–F).

Simple is usually best

The scaling laws identified above suggest that complex models are necessary only for a narrow range of conditions, and otherwise simpler models can be good enough. We tested this idea by identifying, for a broad range of noise and volatility conditions, the simplest model in our hierarchy that achieved performance within 10% of the Bayesian optimum for prediction and estimation tasks (qualitatively similar results were obtained using alternative metrics and tolerance levels; Supplementary Fig. 4 and Supplementary Methods). Also note that these results depend only on the accuracy and the complexity ranking of the models along the hierarchy of Fig. 1 and are therefore independent of the specific measure of complexity that is being used, as long as the ranking is maintained.

For prediction problems (Fig. 5A), relatively complex strategies using adaptivity and working memory are necessary to maximize accuracy over a relatively small range of conditions at low volatility and intermediate noise. In contrast, extremely simple strategies (e.g., Evidence, Prior, and Memoryless models) reach nearly peak accuracy over a wide range of conditions. For example, when volatility is high, the simplest models do nearly as well as the Bayesian predictor, because the world is so variable that past observations do not provide much useful information. When volatility is low, the underlying latent variables are persistent over time, so past observations become more useful for predicting the future. However, this usefulness depends on noise, which obviates the benefits of complex inference when it is too high (and all models perform equally poorly) or too low (and even simple models perform well). Thus, when volatility is low, there is a non-monotonic ("inverted-U") pattern such that simple models are sufficient at low and high noise but complex strategies are needed at intermediate noise; when volatility is high, simple strategies are always good enough.

For estimation problems, slightly different patterns emerge (Fig. 5B). For high noise and volatility, and for low noise across volatilities, certain simple strategies are nearly as effective as complex ones, like for prediction problems. As noise increases from zero at fixed volatility, complex models become useful to balance the current noisy evidence against past observations and the prior. But as noise becomes high (and observations unreliable), increasingly simple models are sufficient again to achieve near-optimal estimation performance. Thus, like for the prediction problem, when volatility is low, there is an inverted-U relationship between the complexity required for good estimation and noise. However, over much of the noise-volatility landscape, estimation problems benefit more than prediction problems from the use of complex inference schemes.

Optimizing cognitive engagement

Above, we selected the simplest model whose performance exceeded a hard threshold as compared to the optimal Bayesian strategy. It might be more realistic to imagine a smooth reward function that decreases with increasing inaccuracy. This reward function can have a characteristic scale that sets the range of inaccuracies over which the animal receives a substantial reward. As a simple example, we can take the reward or performance level to be a Gaussian function of inaccuracy with standard deviation σ_r, so that substantial rewards are obtained when inaccuracy is $O(\sigma_r)$ or smaller.

We can then use this function to derive an expression for expected performance per unit complexity for each noise/volatility pair:

$$\frac{\mathscr{P}(h,R)}{\mathscr{C}(h,R)} = \frac{1}{\sigma_r\sqrt{2\pi}\mathscr{C}(h,R)} e^{-\frac{a(h,R)^2\mathscr{C}(h,R)^{-2b(h,R)}}{2\sigma_r^2}} \tag{3}$$

where $a(h, R)$ and $b(h, R)$ are the parameters of the power-law fits of inaccuracy versus complexity in Fig. 4. Because increased complexity in the inference strategy requires greater cognitive engagment, the ratio in Eq. 3 represents a trade-off between reward and cognitive cost per prediction or estimation. Because algorithmic complexity (Eq. 30) can also be thought of as a qualitative estimate of the time required to make an inference, the ratio (Eq. 3) can also be interpreted as an estimate of the reward one can obtain per unit time [25–28], which is a meaningful fitness function from an evolutionary and behavioural perspective.

The performance per unit cost can be optimized by maximizing the expression on the right hand side of Eq. 3 with respect to the complexity $\mathscr{C}$. This gives

$$\mathscr{C}_{opt}(h, R) = \left(\frac{a(h, R)\sqrt{b(h, R)}}{\sigma_r}\right)^{1/b(h, R)} \tag{4}$$

for the complexity, or cognitive cost, of the optimal inference strategy. Fig. 6 represents log C_{opt} for prediction and estimation tasks across a range of volatilities and noise levels. The results confirm features seen in Fig. 5. For example, high complexity or cognitive engagement is needed only in a small subset of conditions and follows an inverted-U trend with noise at low volatility.

Decreasing the width σ_r of the reward function decreases the tolerance for large inaccuracies and thus broadens the domain in which complex strategies are necessary but otherwise leaves the inverse-U trend unaffected. Likewise, changes in the reflexive component of complexity $\mathscr{C}_{\text{reflex}}$ (Eq. 30) leave the optimal reflective cost, and thus the optimal strategy, unaffected (Supplementary Methods). These results also generalize to other forms of reward functions, including exponential and linear instead of Gaussian, and fitness functions, including a linear Mixture of performance and cost instead of their ratio (Supplementary Fig. 5 and Supplementary Methods).

Subjects' inferences are consistent with the theory

We performed two distinct psychophysical experiments to relate our theory to human inference. In the first experiment, subjects were shown sequences of random numbers sampled from the kind of stochastic processes described in Fig. 2 and, on each trial, were asked to estimate the generative mean of the most recently observed number. Noise and volatility were held constant in blocks of trials and changed from block to block. One group of subjects was tested in three conditions of fixed (low) volatility and variable noise (circles in Fig. 7A). A separate group was tested in three conditions of fixed noise and variable volatility (diamonds in Fig. 7A). These six conditions substantially extend the range of volatility and noise probed in previous experiments, which focused on low-noise, low-volatility environments ([7, 14, 29], small markers in Fig. 7A) that require complex, adaptive inference according to our theory. After training, subject behavior was on average sensitive to the different values of noise and volatility that we tested (Supplementary Fig. 6).

The subjects tended to adjust both adaptivity (their use of flexible time scales for linear integration of past observations) and working memory (their maximum integration time scale) across changes in noise or volatility in a manner that reflected key features of our theory (Fig. 7B–E). In our theory, adaptivity is most useful for estimation tasks with intermediate noise and low volatility. Accordingly, subjects tended to use higher adaptivity for the intermediate versus low (one-tailed t-test, $p < 10^{-4}$, Cohen's $d = 1.1365$) or high ($p = 0.0038$, $d = 0.4225$) noise, and for the low versus intermediate or high volatility conditions

($p < 10^{-4}$ for both comparisons, $d = 0.7647$ and 2.0974, respectively). Furthermore, in our theory working memory is most useful for estimation tasks with intermediate or higher noise and low volatility. Accordingly, our subjects tended to have smaller working-memory loads at low versus intermediate and high noise conditions ($p < 10^{-4}$ for both comparisons, $d = 1.1599$ and 1.1546, respectively) and as a function of increasing volatility ($p < 10^{-4}$ for all the comparisons, $d = 1.2461$, 2.0255, and 0.6337 for the low-intermediate, low-high and intermediate-high comparisons, respectively). These trends across conditions tended to be more pronounced for the theoretical than for the subject values, which likely reflected the use of mixed strategies, different tolerances to errors, and other sources of variability. Nevertheless, even with these additional sources of behavioural variability, these results show that our theoretical framework can be used to identify the task conditions in which different cognitive functions are most likely to be used by human subjects to solve inference problems.

We emphasize that the purpose of this analysis is to show that subjects make inferences that use working memory to different degrees and in more or less flexible (adaptive) ways, regardless of the specific model underlying such inferences. The nested nature of the models in Fig. 1 and the results of Fig. 3 imply that, for each model, the use of adaptivity and working memory changes across conditions corresponding to when that model is reduced to its simpler forms or expanded to its more complex forms.

We complemented this analysis with a standard Bayesian model-selection approach to identify which of the eight models of Fig. 1 best explained the subjects' behavior in each noise and volatility condition (see Supplementary Fig. 8 for confusion analyses). The pattern of selected models matched key predictions of our theory (including when using different inaccuracy tolerances; Supplementary Fig. 4G–I). For low noise and volatility, the probabilities were split between two extreme models for the human data (Fig. 7F, left), which is roughly consistent with the theoretical transition point between the most- and least-complex models in that regime. For intermediate noise (Fig. 7F, middle), the more-complex models had, on average, higher probabilities than simpler models. For high noise (Fig. 7F, right), the single Sliding-Window model had the highest probability, again coinciding with the most-efficient model in the theoretical map. For low volatility (Fig. 7G, left), adaptive models had the highest probabilities, in agreement with the theoretical map. For intermediate (Fig. 7G, middle) and high volatility (Fig. 7G, right), the highest-probability models corresponded to those with intermediate/low complexity and no adaptivity and with the lowest complexity, respectively. In these two conditions, the theoretically most-efficient models of Fig. 7A were also non-adaptive and only slightly more complex. In general, increasing volatility progressively shifted the bulk of the probability distribution from highly complex to highly simple strategies (Fig. 7G, from left to right), as expected from the theory.

To show the generalizability of our results, we extended the study in two ways. First, we repeated the analyses of the behavioral data after including two additional models, the Kalman filter (KF) and the Hierarchical Gaussian Filter (HGF). In contrast to the models of Fig. 1, which were specifically developed for change-point inference, the KF and HGF were developed to track latent states that drift in time according to diffusion processes. Nevertheless, they can also work well in the presence of abrupt change-points [30, 31].

Both models are adaptive and make use of working memory but have lower algorithmic complexity than the Mixture models and the Bayesian model of Fig. 1 (Methods). In Supplementary Fig. 9, we show that all our conclusions remained valid after adding KF and HGF to the model set. From a theoretical standpoint, the models of Fig. 1 were more effective than KF and HGF in practically all volatility/noise conditions and across different tolerances to errors. From an experimental standpoint, the models that best explained the behavioral data in each condition remained those shown in Fig. 7FG.

Second, we performed an additional experiment using a Bernoulli prediction task. Observations were sampled from two possible sources. Each source generated binary observations with constant and complementary generative probabilities, and the source switched to the alternative at random times with constant probability h (Methods). On each trial, subjects were asked to predict the next observation. Thus, two important differences between this experiment and the previous one are the discrete as opposed to continuous nature of the latent process, and the prediction as opposed to estimation nature of the task. Noise R (a function in the [0, 1] range of the generative probabilities of the sources) and volatility h were held constant in blocks of trials and changed from block to block. Subjects were tested in three conditions of fixed (low) volatility and variable noise (circles in Fig. 8A) and in two conditions of fixed noise and variable volatility (diamonds in Fig. 8A, one coinciding with the low-noise condition).

The behavioral results from this task also closely matched theoretical predictions. Here we considered three models that solve this task with decreasing accuracy and complexity [32]: (1) the maximally accurate Bayesian model, (2) a model that approximates the prior expectation with a volatility-dependent constant, and (3) a Leaky-Accumulator model in which the prior expectation is a fraction of the previous belief. Strategies (2) and (3) are approximations of the Bayesian model in regimes of low and high belief uncertainty, respectively. For low and intermediate noise (Fig. 8B, left and middle), and for the two volatility conditions (Fig. 8C), the behavioral responses of almost every subject were best explained by the model predicted by the theoretical map (Fig. 8A). For high noise, behavior was best explained by the Bayesian and Leaky-Accumulator models (Fig. 8B, right); this tested condition is close to the transition between these two models in the theoretical map (Fig. 8A). Similar results were obtained using Bayesian model selection (Supplementary Fig. 10). In summary, this different task supported both theoretically and experimentally the generality of the inverse-U trend of model complexity with noise and the decreasing trend of model complexity with volatility.

Discussion

We used a family of nested models and their mappings to particular cognitive functions to identify principles that govern the trade-off between the accuracy and simplicity of inference in noisy and changing environments. To support the broad applicability of our findings, we chose models that span a wide range of inference strategies, some of which have been studied extensively, including the complex Bayesian observer, its close approximations (the Mixture of Sliding Windows and Mixture of Delta Rules), further approximations that perform simpler forms of integration and “model-free” learning (Sliding Window and Delta

Rule), and even simpler heuristics based on using prior knowledge (Prior), the most-recent observation (Evidence), or both (Memoryless). By deriving these models in a novel, nested framework, we were able to highlight their interrelatedness while defining precisely how their particular, distinguishing features contribute to this trade-off. We also focused on classical inference problems, for which we probed noise and volatility conditions that extended the ranges considered in previous studies [7, 14, 29] and more generally are relevant to tasks that include both forms of uncertainty, such as commonly used reversal-learning tasks [16, 33, 34]. Below we discuss our results relative to other computational, behavioral, and neural findings, focusing on the new insights provided by our work and potential future directions.

Each of our models is characterized by both its complexity and its inaccuracy compared to the exact Bayesian model. By analyzing two nested families of models, we identified a power-law scaling of inaccuracy with complexity: $\mathcal{I} \propto 1/\mathcal{C}^b$. This scaling, with an exponent that depends on noise and volatility in the environment, implies a law of diminishing returns such that increasing the complexity of the inference strategy gives progressively smaller returns in prediction or estimation accuracy. This law is reminiscent of a similar result in rate-distortion theory: the minimum achievable distortion $\mathcal{D}$ of a transmitted signal is a continuous, monotonically decreasing, convex function of the information transmission rate $\mathcal{R}$ [35]. This universal property of rate-distortion functions implies that, independently of the source of information, increasing the communication rate confers diminishing returns in reconstruction accuracy at the receiver.

This relationship to rate-distortion theory implies that constraints on the inference algorithm, imposed by bounded rationality [9], create a sort of information bottleneck [36]. Specifically, in simple contexts, the rate is measured in bits as the mutual information, $I(X; \widehat{X})$, between the input X and output $\widehat{X}$ of an information channel [35]. In this formulation, the distortion for a Gaussian channel (similar to our Gaussian source) scales as $\log D \sim -\mathcal{R}$, for distortions smaller than the variance of the samples. We similarly showed that inaccuracy scales as $\log \mathcal{I} \sim -\log \mathcal{C}$, which suggests an interpretation of the log algorithmic complexity of our models as an effective transmission rate of information about the environment to a decision-making “receiver”, who gathers this information to make inferences about the world. The connection with information theory may provide new practical tools to help understand the diversity of strategies used across tasks and individuals to solve inference problems [11]. Such tools also have the potential to deepen our understanding of the diversity of deep neural networks, for which a power-law scaling between accuracy and computational complexity reminiscent of our findings was recently identified [37].

We also showed that complex strategies that use adaptive processes and/or working-memory are necessary only for a restricted range of conditions, characterized by low volatility and moderate noise, with working memory being useful across a slightly wider set of uncertainty levels, particularly towards higher noise. These dependencies give rise to an inverted-U relationship between cognitive demands and task difficulty: simple strategies are good enough when inference is easy, such as when the current evidence from the environment is highly reliable and thus historical information is not needed, and when inference is hard, such as when incoming information is so noisy or volatile that there is little to gain from complex reasoning. This relationship is reminiscent of a known feature of combinatorial optimization problems: NP-complete problems such as K-satisfiability, graph coloring, the traveling salesman, and the Hamiltonian path problem all have characteristic easy-hard-easy patterns in the computational complexity required to find a solution. Hard problems are typically clustered around a critical intermediate value of an order parameter, which marks a phase transition from solvability to unsolvability [38–41]. In a broad sense, this order parameter plays a role similar to task difficulty or environmental uncertainty in our inference task.

Our behavioral findings build on previous studies that showed that humans tend to use relatively complex inference strategies in conditions of low volatility and moderate noise (Fig. 7A). For example, in [7, 42], subjects performed a Gaussian changepoint task similar to the one considered here, with volatility ~ 0.1 and noise ~ 0.1. Their predictions were consistent with a Mixture of Delta-Rules strategy with two computational units, which provided a better fit than models with either one or three units. This result matches our theoretical findings, for which the (h ~ 0.1, R ~ 0.1) point in the volatility-noise plane falls in the small region where the 2-Delta-Rule model is the most efficient strategy for prediction tasks (Fig. 5). Likewise, in [14], subjects performed a similar Gaussian change-point task with volatility ~ 0.04 and noise between ~ 0.06 and ~ 0.4. Their behavior was better fit by a Delta-Rule model with an adaptive versus a fixed learning rate, which is in agreement with the adaptive domain in our map of efficient models for both prediction and estimation tasks (Fig. 7A). In [29], subjects performed a probabilistic object-reversal task with object-reversal probabilities (analogous to volatility) that ranged between ~ 0.008 and ~ 0.08 and the fraction of trials in which the statistically best option did not receive the top reward (analogous to noise) was 0.2. Again, here adaptive learning was found in a regime of low volatility and intermediate noise, compatible with our theory (Fig. 7A).

Our theoretical framework also makes other predictions that we hope will be tested in future experiments. First, do people solve estimation and prediction problems according to the differences prescribed by our theory? For estimation problems, we showed that memory is not necessary when noise is low, regardless of volatility, whereas for prediction problems, memory is not necessary when both noise and volatility are low (as volatility increases, current evidence carries increasingly little information about the future, and it instead becomes more useful to retain a long-term memory of the average source position). Moreover, complex strategies are useful over a wider region of the volatility/noise landscape for estimation problems than for prediction problems. Second, do subjects learn from recent evidence in conditions of high volatility and variable noise? Our theory predicts a transition between a domain where only prior information about the average source position is useful and a domain where that prior knowledge should be updated based on new evidence (Fig. 5). These two domains are separated by a roughly power-law curve in the volatility-noise plane, so that decreasing volatility increases the noise level beyond which learning from new evidence is useless. Answering these questions will further help to establish if and how trade-offs between accuracy and complexity govern the cognitive operations used to perform inference in the brain.

Our models have plausible neural implementations. We considered exponentially decaying, sliding-window, and instantaneous (Dirac-delta function) integration kernels (implemented in the Delta-Rule, Sliding-Window and Memoryless models, respectively). The exponentially decaying kernels correspond to "α-synapses" used widely in biophysical models of neuron spiking dynamics [43]. Implemented as Delta Rules, they are also closely related to reward-prediction errors that are thought to be encoded by dopaminergic neurons and drive learning in the striatum and possibly elsewhere [16, 44]. This implementation, compared to exponentially decaying integration, has advantages in terms of working memory, because it effectively produces Markovian estimates of the source: each estimate depends only on the current observation and on the immediately previous estimate. The Sliding-Window kernels are more memory intensive, requiring representations of each sample used in the given window. Such memory signals could, in principle, be based on persistent activity that maintains representations of a sequence of observations, such as those found in the prefrontal cortex network [45]. The Dirac-delta kernels can be implemented trivially without any working memory.

Adaptivity is achieved in our models using a bank of different integration timescales, consistent with multiple reports describing different integration timescales in the brain [46–50]. In our formulation, the estimates obtained from these different integration timescales are weighted optimally and combined to produce a single output [7, 42]. Consistent with this idea, learning rates with more relevance to an ongoing estimate of choice values have been shown to explain more variance in fMRI signals [51]. This weighting process is thought to require the noradrenergic, cholinergic, and dopaminergic neuromodulatory systems, each of which has been linked to adaptive inference via pupillometry and other measures [52, 53]. The noradrenergic and dopaminergic systems are also thought to be responsible for an inverted-U relationship between learning and arousal, via their effects on neural activity in the prefrontal cortex and perhaps elsewhere in the brain [53–56]. It is tempting to think that the statistical difficulty of a task might modulate activity in these brain areas similarly to arousal states, to engage or disengage mental resources in a way that best meets task demands.

An alternative hypothesis on how adaptive Bayesian inference might be approximated by the brain is based on particle filters and importance sampling [3, 4, 26, 57]. In these approaches, a limited number of samples (particles) is used to represent the posterior distribution of the hidden state given the observations. Unlike in our models, in these approaches the hypothesis space for the hidden state varies in time, as new hypotheses are continuously sampled from the Bayesian posterior distribution given the observations. By contrast, in our Mixture models the hypothesis space (set of run-lengths or integration timescales) is constrained and fixed in a given environment and adaptivity is achieved by weighting the different hypotheses by their time-dependent posterior probability. It would be useful for future work to compare the computational complexity of these different kinds of approaches to adaptive inference, which could help advance our understanding of if and when they could be used in the brain.

Overall, our study provides a unified view of several plausible models of on-line statistical inference, showing that they can be regarded as special cases of a single formalism. This novel interpretation suggests a hierarchical (nested) organization of cognitive processes and a natural, efficient way in which the brain could engage or disengage them. This organization implies that the brain could meet the demands of a wide range of different environments and tasks, by adjusting the parameters of a single, flexible inference process.

Methods

Statement on Ethics Approval

The study protocol for human subjects research was reviewed by the University of Pennsylvania Institutional Review Board and determined to be exempt as authorized by 45 CFR 46.104, category #2. All participants provided informed consent prior to participating.

Change-point tasks

We considered two different change-point tasks: a continuous Gaussian change-point task and a discrete Bernoulli task.

All analyses in Figs. 1–7 and Supplementary Figs. 1–9 refer to the Gaussian change-point task (Fig. 2) [7, 14]. Observations x_t take continuous values and are Gaussian distributed ($p(x_t) = \mathcal{N}(x_t|\mu_t, \sigma^2)$) around a source located at an unknown mean position μ_t. The mean position changes at random times, with probability h (the volatility parameter). At these change-points, the source is resampled from another Gaussian distribution $\left(p(\mu_t) = \mathcal{N}\left(\mu_t \mid \bar{\mu}, \sigma_0^2\right)\right)$. The goal of an observer is to infer the current position of the source μ_t from the history of observations up to time t (estimation problem), or to predict the position of the source at the next time step μ_{t+1} (prediction problem). For each simulation and task condition, the parameters $\bar{\mu}$, σ_0, and σ are constant. The ratio ($R = \sigma/\sigma_0$) is the noise parameter of the process $(R = 1/\sqrt{SNR})$. The volatility and noise parameters determine the statistical difficulty of the inference problem.

The analyses in Fig. 8 and Supplementary Fig. 10 refer to the discrete Bernoulli task. In this case, observations are generated from two possible sources: $\mu_t = 0$ and $\mu_t = 1$. The first one generates observations $x_t = 0$ with probability $0.5 < p < 1$ and $x_t = 1$ with probability $1 - p$; vice versa, the second source generates observations $x_t = 1$ and $x_t = 0$ with probabilities p and $1 - p$, respectively. At any time, the source can switch to the alternative with hazard rate h, according to a Bernoulli process. The goal of an observer is to predict the next source μ_{t+1} from the history of observations $x_{1:t}$. For each simulation and task condition, the parameters h (volatility) and p are constant. In this task, noise can be quantified as $R = 2(1 - p)$, which is a number between 0 and 1: $R = 0$ implies that observations are always consistent with their source, whereas $R = 1$ implies that observations are independent of the source.

Exact Bayesian inference

For the Gaussian change-point task, we derive expressions for $\mu_t^{r_t}$ and $p(r_t|x_{1:t})$ to obtain the optimal Bayesian estimate of the current source position and the optimal Bayesian prediction of the next source position (text around Eq. 2) [12].

For Gaussian processes, the posterior probability of the source μ_t given run-length r_t is

$$p(\mu_t \mid r_t) = p\left(\mu_t \mid x_{t-r_t+1:t}\right) \propto \mathcal{N}\left(\mu_t \mid \frac{\chi_p}{\nu_p}, \frac{\sigma^2}{\nu_p}\right) \prod_{i=t-r_t+1}^{t} \mathcal{N}\left(\mu_t \mid x_i, \sigma^2\right) \tag{5}$$

where we have used the Bayes rule $p\left(\mu_t \mid x_{t-r_t+1:t}\right) \propto p(x_t \mid \mu_t)p\left(\mu_t \mid x_{t-r_t+1:t-1}\right)$ recursively. Note that $\mathcal{N}\left(\mu_t \mid \frac{\chi_p}{\nu_p}, \frac{\sigma^2}{\nu_p}\right)$ is the Gaussian prior distribution over μ_t with mean $\bar{\mu} = \frac{\chi_p}{\nu_p}$ and variance $\sigma_0^2 = \frac{\sigma^2}{\nu_p}$. Using the relation $\mathcal{N}\left(\mu \mid \mu_1, \sigma_1^2\right)\mathcal{N}\left(\mu \mid \mu_2, \sigma_2^2\right) \propto \mathcal{N}\left(\mu \mid \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$ we obtain:

$$p(\mu_t \mid r_t) = \mathcal{N}\left(\mu_t \mid \mu_t^{r_t}, \frac{\sigma^2}{\nu_t^{r_t}}\right) \tag{6}$$

with

$$\mu_t^{r_t} = \frac{\chi_t^{r_t}}{\nu_t^{r_t}} \quad ; \quad \chi_t^{r_t} = \chi_p + \sum_{i=t-r_t+1}^{t} x_i \quad ; \quad \nu_t^{r_t} = \nu_p + r_t \tag{7}$$

As expected for a Gaussian prior and a Gaussian likelihood, the posterior distribution (Eq. 6) is also Gaussian.

The posterior probability of the run-length r_t given observations $x_{1:t}$ can be computed recursively:

$$\begin{aligned} p(r_t \mid x_{1:t}) &= \frac{p(r_t, x_{1:t})}{p(x_{1:t})} \\ &= \frac{1}{p(x_{1:t})} \sum_{r_{t-1}=1}^{t-1} p(r_t \mid r_{t-1}, x_{1:t})p(r_{t-1}, x_{1:t}) \\ &= \frac{1}{p(x_{1:t})} \sum_{r_{t-1}=1}^{t-1} p(r_t \mid r_{t-1}, x_t)p(x_t \mid r_{t-1})p(r_{t-1}, x_{1:t-1}) \end{aligned} \tag{8}$$

Because $r_t = 1$ if there is a change-point (“cp” below) at time t, $r_t = r_{t-1} + 1$ if there is no change-point, and change-points occur with constant probability h, we can rewrite $p(r_t|r_{t-1}, x_t)$ as:

$$\begin{aligned}
& p(r_t = r_{t-1} + 1 \mid r_{t-1}, x_t) = p(\text{no cp} \mid r_{t-1}, x_t) \\
& = \frac{p(x_t \mid \text{no cp}, r_{t-1})\, p(\text{no cp})}{p(x_t \mid r_{t-1})} \\
& = \frac{1-h}{p(x_t \mid r_{t-1})} \int_{-\infty}^{\infty} d\mu_{t-1}\, p(x_t \mid \text{no cp}, \mu_{t-1})\, p(\mu_{t-1} \mid r_{t-1}) \\
& = \frac{1-h}{p(x_t \mid r_{t-1})} \int_{-\infty}^{\infty} d\mu_{t-1}\, \mathcal{N}\left(x_t \mid \mu_{t-1}, \sigma^2\right) \mathcal{N}\left(\mu_{t-1} \mid \mu_{t-1}^{r_{t-1}}, \frac{\sigma^2}{\nu_{t-1}^{r_{t-1}}}\right) \\
& = \frac{1-h}{p(x_t \mid r_{t-1})} \mathcal{N}\left(x_t \mid \mu_{t-1}^{r_{t-1}}, \sigma^2\left(1 + \frac{1}{\nu_{t-1}^{r_{t-1}}}\right)\right)
\end{aligned} \tag{9a}$$

$$\begin{aligned}
& p(r_t = 1 \mid r_{t-1}, x_t) = p(\text{cp} \mid r_{t-1}, x_t) \\
& = \frac{p(x_t \mid \text{cp})\, p(\text{cp})}{p(x_t \mid r_{t-1})} \\
& = \frac{h}{p(x_t \mid r_{t-1})} \int_{-\infty}^{\infty} d\mu_t\, p(x_t \mid \mu_t)\, p(\mu_t \mid \text{cp}) \\
& = \frac{h}{p(x_t \mid r_{t-1})} \int_{-\infty}^{\infty} d\mu_t\, \mathcal{N}\left(x_t \mid \mu_t, \sigma^2\right) \mathcal{N}\left(\mu_t \mid \bar{\mu}, \sigma_0^2\right) \\
& = \frac{h}{p(x_t \mid r_{t-1})} \mathcal{N}\left(x_t \mid \bar{\mu}, \sigma^2 + \sigma_0^2\right)
\end{aligned} \tag{9b}$$

$$p(r_t \mid r_{t-1}, x_t) = 0 \quad \forall\, r_t \neq r_{t-1} + 1\,;\; r_t \neq 1 \tag{9c}$$

Substituting Eqs. 9 into Eq. 8 we obtain:

$$p(r_t \mid x_{1:t}) = \frac{1}{C} \mathcal{N}\left(x_t \mid \mu_{t-1}^{r_{t-1}}, \sigma^2\left(1 + \frac{1}{\nu_{t-1}^{r_{t-1}}}\right)\right) \sum_{r_{t-1}=1}^{t-1} p(r_t \mid r_{t-1})\, p(r_{t-1} \mid x_{1:t-1} \tag{10}$$

$$)$$

with C being a normalization constant, $\mu_t^0 = \bar{\mu}$, $\nu_t^0 = \nu_p$ (for any t) and

$$\begin{cases}
p(r_t \mid r_{t-1}) = 1 - h & \text{if} \quad r_t = r_{t-1} + 1 \\
p(r_t \mid r_{t-1}) = h & \text{if} \quad r_t = 1 \\
p(r_t \mid r_{t-1}) = 0 & \text{otherwise}
\end{cases} \tag{11}$$

Eq. 10 simplifies to:

$$\begin{cases}
p(r_t \mid x_{1:t}) = \frac{1}{C}(1-h) \mathcal{N}\left(x_t \mid \mu_{t-1}^{r_{t-1}}, \sigma^2\left(1 + \frac{1}{\nu_{t-1}^{r_{t-1}}}\right)\right) p(r_{t-1} = r_t - 1 \mid x_{1:t-1}) & \text{if} \;\; r_t \neq 1 \\
p(r_t \mid x_{1:t}) = \frac{1}{C} h \mathcal{N}\left(x_t \mid \mu_{t-1}^{r_{t-1}}, \sigma^2\left(1 + \frac{1}{\nu_{t-1}^{r_{t-1}}}\right)\right) & \text{if} \;\; r_t = 1
\end{cases} \tag{12}$$

In conclusion, we can compute:

$$\begin{aligned} p(\mu_t \mid x_{1:t}) &= \sum_{r_t=1}^{t} p(\mu_t \mid r_t) p(r_t \mid x_{1:t}) \\ &= \sum_{r_t=1}^{t} p(r_t \mid x_{1:t}) \mathcal{N}\left(\mu_t \mid \mu_t^{r_t}, \frac{\sigma^2}{\nu_t^{r_t}}\right) \end{aligned} \tag{13}$$

and the optimal (mean-squared-error minimizing) estimate of the source μ_t given the history of observations $x_{1:t}$ is

$$\hat{\mu}_t = \langle \mu_t \rangle_{p(\mu_t \mid x_{1:t})} = \sum_{r_t=1}^{t} p(r_t \mid x_{1:t}) \mu_t^{r_t} \tag{14}$$

From $p(\mu_t|x_{1:t})$ it is straightforward to derive the posterior probability distribution for the position of the source at the next time step:

$$\begin{aligned} p(\mu_{t+1} \mid x_{1:t}) &= \int_{-\infty}^{\infty} d\mu_t p(\mu_{t+1} \mid \mu_t) p(\mu_t \mid x_{1:t}) \\ &= \int_{-\infty}^{\infty} d\mu_t (p(\mu_{t+1} \mid \mu_t, \mathrm{cp}) p(\mathrm{cp}) + p(\mu_{t+1} \mid \mu_t, \text{ no cp}) p(\text{no cp})) p(\mu_t \mid x_{1:t}) \\ &= h \int_{-\infty}^{\infty} d\mu_t p(\mu_{t+1} \mid \mathrm{cp}) p(\mu_t \mid x_{1:t}) + (1-h) \int_{-\infty}^{\infty} d\mu_t \, \delta(\mu_{t+1} - \mu_t) p(\mu_t \mid x_{1:t}) \\ &= h \, \mathcal{N}\left(\mu_{t+1} \mid \bar{\mu}, \sigma_0^2\right) + (1-h) \int_{-\infty}^{\infty} d\mu_t \delta(\mu_{t+1} - \mu_t) \sum_{r_t=1}^{t} p(r_t \mid x_{1:t}) \mathcal{N}\left(\mu_t \mid \mu_t^{r_t}, \frac{\sigma^2}{\nu_t^{r_t}}\right) \\ &= h \, \mathcal{N}\left(\mu_{t+1} \mid \bar{\mu}, \sigma_0^2\right) + (1-h) \sum_{r_t=1}^{t} p(r_t \mid x_{1:t}) \mathcal{N}\left(\mu_{t+1} \mid \mu_t^{r_t}, \frac{\sigma^2}{\nu_t^{r_t}}\right). \end{aligned} \tag{15}$$

It follows that the optimal Bayesian prediction of μ_{t+1} given the history of observations up to time t is

$$\hat{\mu}_{t+1} = \langle \mu_{t+1} \rangle_{p(\mu_{t+1} \mid x_{1:t})} = h\bar{\mu} + (1-h)\hat{\mu}_t \, . \tag{16}$$

Building a hierarchy of approximations to Bayesian inference

The Bayesian computation is simplified by considering only a fixed set of run-lengths $\{r_1, \ldots, r_N\}$ chosen to minimize the mean squared error in the estimator. This reduction approximates the full Bayesian model with N computational units, each charged with generating an estimate of μ_t based on a sliding-window integration of past observations over the duration r_i, combined with prior information on the average value of the source $\bar{\mu}$:

$$\mu_t^{r_i} = \frac{\chi_t^{r_i}}{\nu_t^{r_i}} = \frac{1}{\nu_p + r_i}\left(\nu_p\bar{\mu} + dx_{t-\lfloor r_i\rfloor} + \sum_{k=t-\lfloor r_i\rfloor+1}^{t} x_k\right) \tag{17}$$

We will think of the model run-length r_i as being allowed to take non-integer values in the mathematical expression to allow greater flexibility, and $d = r_i - \lfloor r_i \rfloor$ is the decimal part of r_i. The relative weight of the prior mean with respect to each observation is $\nu_p = \sigma^2/\sigma_0^2 = R^2$: the larger the noise, the more the model relies on the prior mean as opposed to the empirical mean computed from the observations.

Note that Eq. 17 corresponds to Eq. 7 in the Bayesian model, with $r_t = r_i$. Eq. 17 can also be expressed recursively as:

$$\mu_t^{r_i} = \mu_{t-1}^{r_i} + \alpha_i\left(x_t - (1-d)x_{t-\lfloor r_i\rfloor} - dx_{t-\lfloor r_i\rfloor-1}\right) \quad ; \quad \alpha_i \equiv \frac{1}{\nu_p + r_i} \tag{18}$$

with effective learning rate α_i and initial condition $\mu_{\lfloor r_i\rfloor+1}^{r_i} = \frac{1}{\nu_p + r_i}\left(\nu_p\bar{\mu} + dx_1 + \sum_{k=2}^{\lfloor r_i\rfloor+1} x_k\right)$. Estimates computed by each unit are summed with a relative weight set adaptively by the posterior probabilities $p(r_i|x_{1:t})$ (see next section). As such, low/high volatility in the world will lead to preferential use of long/short Sliding Windows [16, 58]. This Mixture of Sliding-Windows model is simpler than the full Bayesian procedure, but implementing it in the brain would still require extensive working memory, up to the longest run-length, and circuitry to compute the adaptive weights given to different run-lengths.

The working-memory load can be reduced by replacing the Sliding Windows with units that weigh past observations according to exponentially decaying kernels

$$\mu_t^{\alpha_i} = (1-\alpha_i)^t\bar{\mu} + \sum_{k=1}^{t} \alpha_i e^{-\frac{t-k}{\tau}} x_k \tag{19}$$

Expressing the time constant as $\tau = -1/\ln(1 - \alpha_i)$ ($\sim 1/\alpha_i$ for $\alpha_i \ll 1$), these units are equivalent to the delta rules of reinforcement learning:

$$\mu_t^{\alpha_i} = \mu_{t-1}^{\alpha_i} + \alpha_i\left(x_t - \mu_{t-1}^{\alpha_i}\right) \tag{20}$$

with learning rates α_i in the range [0, 1] and initial condition $\mu_0^{\alpha_i} = \bar{\mu}$. The Delta-Rule units are approximations of the Sliding-Window units, in which the weighted average of the two observations occurring ~ r_i time steps back in the past $(1-d)x_{t-\lfloor r_i\rfloor} + dx_{t-\lfloor r_i\rfloor-1}$ (Eq. 18) is replaced by the unit estimate $\mu_{t-1}^{\alpha_i}$ of the mean at time t−1: this approximation reduces the working-memory demand to the previous time step only.

The demand for computational resources in both Sliding-Window and Delta-Rule model families is reduced dramatically by making them non-adaptive. This reduction amounts to considering a single Sliding Window or a single Delta Rule, each with an optimal fixed timescale for integrating evidence. These models do not need to estimate the adaptive weights, but still require working memory to carry out the integration.

An even simpler inference strategy estimates the source μ_t as a weighted average between the present observation x_t and the average source $\bar{\mu}$ (Dirac-delta kernel):

$$\hat{\mu}_t = (1 - \alpha)\bar{\mu} + \alpha x_t \tag{21}$$

This Memoryless model is nested in the Sliding-Window model, from which it is derived by choosing an integration window of just one observation. The Memoryless model is the minimal model that learns and updates prior biases, or knowledge of stable features of the environment ($\bar{\mu}$), based on current evidence from a rapidly changing variable (x_t). We stress that the name “Memoryless” is used to indicate that this model does not perform any integration of evidence over time, thus it does not require any working memory of past observations or past inferences; however, the model maintains a long-term memory of prior information.

Both the Memoryless and Delta-Rule models can be further reduced to the simple Prior model ($\hat{\mu}_t = \bar{\mu}$) by setting the learning rate to zero. This Prior model represents knowledge acquired, after a sufficiently long exposure to a given environment, about the constant or slow (stable across many change-points) features of the process generating the observations. Inferring and storing the slowly varying structure of the environment presumably still requires some cognitive effort and long-term memory resources.

Removing this last cognitive demand leads to a strategy that simply returns the current observation x_t as both an estimate of μ_t and a prediction of μ_{t+1}. This strategy, which we call Evidence, can also be seen as the simplest possible model nested in both the Memoryless and the Delta-Rule models, because it is obtained from them by setting the learning rate to one.

The hierachical relationships between the models are summarized, in terms of parameter reductions, in Supplementary Fig. 1.

The models were optimized by finding the parameters r_i and α_i that minimized mean-squared error of the model estimates (or predictions). Specifically, for each combination of volatility h and noise $\nu_p = R^2$, we used the Matlab “interior-point” algorithm to find the parameter values minimizing mean-squared error over a 5000 time-long instance of the Gaussian change-point process. This duration guaranteed that any instance contained a large number of change points (on average 100 at the minimum tested volatility $h = 0.02$ and larger numbers at higher h). To reduce dependency on the specific instance, we averaged the output of this optimization over 10 different random instances (of 5000 time-steps each). The resulting optimal parameters of the Sliding-Window and Delta-Rule models vary with noise and volatility, whereas the optimal parameter of the Memoryless model is $\alpha = \frac{1}{v_p + 1}$ and is independent of volatility.

Posterior probabilities in the Mixture models

In the Mixture models, the posterior probabilities of the run-lengths $\{r_i\}$, $i = 1, \ldots, N$ are obtained as an approximation of the Bayesian posterior $p(r_t|x_{1:t})$ (compare with Eq. 10 above)

$$p\left(r_i \mid x_{1:t}\right) = \frac{1}{C}\mathcal{N}\left(x_t \mid \mu_t^{r_i}, \sigma^2\left(1 + \frac{1}{v_t^{r_i}}\right)\right)\sum_{j=1}^{N} p\left(r_i \mid r_j\right)p\left(r_j \mid x_{1:t-1}\right), \tag{22}$$

where the transition probabilities $p(r_i|r_j)$ approximate $p(r_t|r_{t-1})$ of the exact Bayesian model [7, 42]:

$$p\left(r_i \mid r_j\right) = hp\left(r_i \mid r_j, \mathrm{cp}\right) + (1-h)p\left(r_i \mid r_j, \text{ no cp}\right) \tag{23}$$

We sort the N model run-lengths in ascending order: $r_1 < r_2 < \cdots < r_N$. When there is a change-point, the Bayesian run-length drops to 1. This condition is approximated by resetting the model run-length to the smallest possible value r_1:

$$p\left(r_i \mid r_j, \mathrm{cp}\right) = \begin{cases} 1 & \text{if} \quad i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

When there is not a change-point, the Bayesian run-length increases by 1. Given the finite number of run-lengths in the Mixture models, the distance between any r_j and r_{j+1} is in general different from 1. To approximate the Bayesian transition, two cases are considered: (1) when $r_{j+1} \geq r_j + 1$, the model run-length increases from r_j to r_{j+1} with a probability inversely proportional to the distance $r_{j+1} - r_j$ and it remains constant with the complementary probability, so that the increase in model run-length is equal to 1 on average; (2) when $r_{j+1} < r_j +1$, transition always occurs. More formally:

For all $j < N$:

If $r_{j+1} \geq r_j + 1$ then:

$$p\left(r_i \mid r_j, \text{ no cp }\right) = \begin{cases} \frac{1}{r_{j+1} - r_j} & \text{if } i = j+1 \\ 1 - \frac{1}{r_{j+1} - r_j} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Else if $r_{j+1} < r_j + 1$ then:

$$p\left(r_i \mid r_j,\ \text{no cp}\right) = \begin{cases} 1 & \text{if} \quad i = j + 1 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

For $j = N$:

$$p(r_i \mid r_N,\ \text{no cp}) = \begin{cases} 1 & \text{if } i = N \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

We have used Mixture models with $N = 2$ units.

Redundancy and Alignment

In a given environment characterized by parameters $\{e_l\}$, we can quantify a model reducibility, or the relative importance of the most vs. least relevant degrees of freedom at the optimal parameter values $\{\hat{\alpha}_k\}$, as:

$$\text{Redundancy} = \log_{10}\left(\frac{\lambda_{max}(\{e_l, \hat{\alpha}_k, \})}{\lambda_{min}(\{e_l, \hat{\alpha}_k\})}\right) \tag{28}$$

where λ_{min} and λ_{max} are the eigenvalues of the Hessian of the model error function. In this study, the environmental parameters are the volatility h and the noise R; the error function E is the mean-squared prediction error over 5000 time steps of the process for each choice of h and R; both adaptive Mixture models have two parameters $\{\alpha_1, \alpha_2\}$ describing effective learning rates, which are related to (1) the window length of evidence integration in the Sliding Windows (Eqs. 17, 18), and (2) the timescale of the exponential evidence-integration kernel in the Delta Rules (Eqs. 19, 20). For both model families, the Hessian is defined as $H = \frac{\partial^2 E}{\partial \alpha_1 \partial \alpha_2}$.

The Alignment is evaluated as the (normalized) angle between the most irrelevant eigenvector of H and the direction in parameter space connecting the optimal Mixture model to the optimal non-adaptive nested model (Fig. 3A). Let $\boldsymbol{\delta\alpha} = \boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha}^{(2)}$, where $\boldsymbol{\alpha}^{(2)} = (\hat{\alpha}_1, \hat{\alpha}_2)$ is the two-component vector of the optimal parameters of the Mixture model, and $\boldsymbol{\alpha}^{(1)} = (\hat{\alpha}, \hat{\alpha})$ is the vector with both components equal to the optimal parameter of the non-adaptive nested model. The vector $\boldsymbol{\delta\alpha}$ is directed along the parameter transformation collapsing the best Mixture model into the best nested single-unit model. Thus, we can define

$$\text{Alignment} = \frac{\left|\pi/2 - \theta(h, R)\right|}{\pi/2} \tag{29}$$

where $0 \leq \theta \leq \boldsymbol{\pi}$ is the angle between the irrelevant eigenvector and the direction of $\boldsymbol{\delta\alpha}$ (Fig. 3A). By definition, $0 \leq$ Alignment ≤ 1 and is a function of volatility h and noise R. To reduce numerical noise, in Fig. 3, Redundancy and Alignment were averaged over 10 instances of the Gaussian change-point process for each volatility and noise value.

Algorithmic complexity

The algorithmic complexity (Fig. 4 and Supplementary Fig. 2) of an inference strategy is defined as the average number of computational and memory operations required to implement it:

$$\mathscr{C}(h, R) = C_{\text{reflex}} + \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \left(C_t^A + C_t^W + C_t^R + C_t^S(h, R) \right) \quad (30)$$

Here T is the total number of observations, C_t^A denotes the cost of arithmetic operations and C_t^W, C_t^R, and C_t^S denote the costs of memory-related operations (writing, reading, and storing, respectively) required to make an inference (estimation or prediction) at time t. We interpret the sum of these terms as an estimate of the "reflective" cost of making a decision, whereas C_{reflex} can be interpreted as a purely reflexive component that represents the irreducible cost of an action. This reflexive cost is not known, but because it is an equal constant for all models, its value does not influence the conclusions of this study (see Supplementary Information).

We computed the reflective costs in two different and complementary analyses.

SIMPLE ALGORITHMIC IMPLEMENTATION.—In the analysis of Fig. 4, for simplicity, we assigned cost = 1 to each arithmetic and memory operation. Thus, in this case, the reflective cost reduces to the total mean number of operations per inference: $\langle N^A \rangle + \langle N^W \rangle + \langle N^R \rangle + \langle N^S \rangle$, where we use the notation $\langle N^i \rangle$ to indicate $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} N_t^i$, with N_t^i number of operations of type $i \in \{A, W, R, S\}$ required in the t-th iteration (returning one inference) of the algorithm implemented by each model. More precisely, for memory operations, we define N_t^W as the number of variables that have to be written into memory (at iteration t), N_t^R as the number of times each variable has to be read from memory (at t), summed over all variables, and N_t^S as the number of iterations (starting at t) during which each variable has to be kept in memory to make future inferences, summed over all stored variables.

Supplementary Table 1 lists $\langle N^A \rangle$, $\langle N^W \rangle$, $\langle N^R \rangle$, and $\langle N^S \rangle$ for the estimation problem, for each of the seven models derived from the exact Bayesian strategy (Fig. 1). Below we explain how we determined these values, and how they can be readily converted into the corresponding values for the prediction problem. We will only indicate operations that are performed by the models in every inference, because $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} N_t^i = 0$ for one-off operations.

The Evidence model returns each instantaneous piece of evidence $\hat{\mu}_t = x_t$, and does not require any computation or memory operation.

The Prior model stores the prior mean $\bar{\mu}$ for one iteration at every t ($\langle N^S \rangle = 1$) and reads it from memory ($\langle N^R \rangle = 1$).

The Memoryless model estimates the source position as $\hat{\mu}_t = \overline{\mu} + \alpha(x_t - \overline{\mu})$, which requires $\langle N^A \rangle = 3$ (1 sum, 1 subtraction, 1 multiplication), $\langle N^S \rangle = 2$(to store $\overline{\mu}$ and α), and $\langle N^R \rangle = 3$(to read $\overline{\mu}$, twice, and α, once).

The Delta Rule computes $\hat{\mu}_t = \hat{\mu}_{t-1} + \alpha(x_t - \hat{\mu}_{t-1})$, which involves the same number of algorithmic and memory operations as the Memoryless model estimate, with the addition of one writing operation per iteration ($\langle N^W \rangle = 1$), because the computation is recursive, requiring to write $\hat{\mu}_t$ into memory at every t to compute $\hat{\mu}_{t+1}$. This one-time-step dependence allows the Delta Rule to integrate the evidence over time, unlike the Memoryless model.

The Sliding Window computes $\hat{\mu}_t = \hat{\mu}_{t-1} + \alpha\left(x_t - (1-d)x_{t-\lfloor r \rfloor} - dx_{t-\lfloor r \rfloor - 1}\right)$ (with $\alpha = \frac{1}{v_p + r}$), which requires $\langle N^A \rangle = 7$ arithmetic operations (1 sum, 3 differences, 3 products); $\langle N^W \rangle = 2$ operations to write, at every t, $\hat{\mu}_t$ (necessary to compute the estimate at $t + 1$) and x_t (necessary to compute the estimates at $t + \lfloor r \rfloor$ and $t + \lfloor r \rfloor + 1$); $\langle N^R \rangle = 6$ operations to read, at every t, $\hat{\mu}_{t-1}$, α, d (twice), $x_{t-\lfloor r \rfloor}$ and $x_{t-\lfloor r \rfloor - 1}$); finally $\langle N^S \rangle = \lfloor r \rfloor + 4$ operations to store, at every t, α, d, $\hat{\mu}_t$ (for one iteration), and x_t (for a duration of $\lfloor r \rfloor + 1$ iterations). Because of the dependence on $r(h, R)$ (the time scale of the Sliding-Window integration of past evidence that minimizes mean squared error), this model and its Mixture have complexity that depends on the environmental noise and volatility; for example, complexity increases with increasing noise to integrate observations over longer time scales, which allows more accurate estimates of the source. All the other models have complexity that is independent of noise and volatility, because they retain either no memory of past evidence (Evidence, Prior and Memoryless models), or only a memory of the previous estimate (Delta-Rule models), regardless of environmental statistics.

In the Mixture models, each of the N units performs the same computations as the corresponding single-unit models. Thus, the contribution to the complexity of the Mixture models coming from the computations taking place in the single units reduces to the complexity of the single Delta Rule and single Sliding Window, respectively, when $N = 1$ (Supplementary Table 1, first line of the respective slots). However, the largest contribution to the complexity of the Mixture models comes from the computations that combine the estimates provided by the N units into a single inference of the source (Supplementary Table 1, second line of the respective slots, where the Heaviside function $H_2 = H[N - 2]$ vanishes for $N = 1$). These computations are necessary to obtain the adaptive probabilities $p(r_i|x_{1:t})$ of the N run-lengths at each iteration t of the algorithm, which are then used to weigh the estimates of the single units. In particular, for both Mixtures, the leading-order term of N^A ($7N^2$) comes from 2 summations, over N terms each, required to compute each of the N adaptive $p(r_i|x_{1:t})$ (Eq. 22): (1) the summation over N run-lengths j appearing at the numerator of Eq. 22 (which involves $6N$−1 algorithmic operations), and (2) the summation necessary to compute the normalization constant (which involves N−1 algorithmic operations). The leading-order term of $\langle N^R \rangle$ ($6N^2$) comes from reading the terms in the same summations. $\langle N^W \rangle$ scales as ~ N (not as ~ N^2) because only the N probabilities $p(r_i|x_{1:t})$ are carried forward to the next iteration of the algorithm to compute the new $p(r_i|x_{1:t+1})$, whereas the individual addends of the summations mentioned above do not need to be memorized. Finally, the leading-order term of $\langle N^S \rangle$ $(2N^2)$ arises because computation of the adaptive $p(r_i|x_{1:t})$ requires maintaining in memory, at every iteration, the $N \times N$ matrices of the transition probabilities $p(r_i|r_j$, cp) and $p(r_i|r_j$, no cp) (Eqs. 24 through 27).

The differences between the complexities of the Mixture of Delta Rules and the Mixture of Sliding Windows only involve terms of order $\mathcal{O}(N)$ and $\mathcal{O}(1)$, and come entirely from the computations taking place in the single units (note that the second line in the slots of Supplementary Table 1 corresponding to the two Mixture models are identical).

The complexities in the prediction problem can be readily obtained from the complexities in the estimation problem, as follows. For the Evidence and Prior models, predictions coincide with estimations, thus their complexity is the same as in Supplementary Table 1. For all the other models, predictions are computed from estimations as $\hat{\mu}_{t+1} = (1-h)\hat{\mu}_t + h\bar{\mu}$. Thus, each prediction requires 4 more algorithmic operations than each estimation, 3 more reading operations (to retrieve from memory h, twice, and $\bar{\mu}$, once), and either 2 more storing operations for the Delta Rule and Sliding Window (to store both h and $\bar{\mu}$), or just 1 more storing operation for the Memoryless model (to store h, as this model already requires to store $\bar{\mu}$ to obtain the estimate $\hat{\mu}_t$) and for the Mixture models (to store $\bar{\mu}$, as h is already stored to estimate $\hat{\mu}_t$).

Following exactly the same approach described above, we also computed the algorithmic complexity of the two additional models in Supplementary Fig. 9. The Kalman Filter (algorithm of [59, 60]) requires $\langle N^A \rangle = 9$, $\langle N^W \rangle = 4$, $\langle N^R \rangle = 12$, and $\langle N^S \rangle = 2$ mean operations per inference. A Hierarchical Gaussian Filter with n_l levels (algorithm of [30, 61, 62]) requires $\langle N^A \rangle = 29n_l - 22$, $\langle N^W \rangle = 6n_l - 3$, $\langle N^R \rangle = 31n_l - 22$, and $\langle N^S \rangle = 6n_l - 1$. We used $n_l = 2$ in our analyses.

NEURAL-SPIKING PARALLEL IMPLEMENTATION.—In Supplementary Fig. 2, we considered plausible neural-network implementations of the 4 arithmetic operations (addition, subtraction, multiplication, division). These neural circuits, called Neural-Spiking Parallel Systems, and their operating principles are presented in [24].

We obtained the neural cost of each arithmentic operation by computing the number of spikes that the circuit consumes to perform the operation. If x, y are the inputs to the circuits (e.g., the addends for the addition operation), these neural costs are:

$$\text{Addition:} \quad 4(x+y)+3 \quad \text{spikes} \tag{31a}$$

$$\text{Subtraction:} \quad 4(x+y)+7 \quad \text{spikes} \tag{31b}$$

$$\text{Multiplication:} \quad 2(x+y)+8xy+9x+9 \quad \text{spikes} \tag{31c}$$

$$\text{Division:} \quad 4x + 16y + 11(y+1) * \lfloor x/y \rfloor + 9 \quad \text{spikes} \tag{31d}$$

We obtained the neural cost of the square root operation ($\sqrt{1+\alpha}$) from the costs of the other 4 operations using the Taylor expansion of the square root up to second order. We assigned a cost of 2 spikes to exponential operations, because in neural-spiking P systems any output is encoded in 2 spikes, and an exponential can be implemented through the integration kernel of a single neuron. The same encoding cost was assigned to each memory operation.

The number of spikes for each operation depends on the inputs (i.e., the specific numbers being summed, multiplied etc. (Eqs. 31). Thus, to compute the average cost of each operation, we sampled a large number (10^{10}) of random inputs in the typical range in which the algorithmic inputs vary and computed the average number of spikes for each operation over those inputs.

Finally, to readily compare results with those of Fig. 4, we considered addition as the unit of cost (i.e., we normalized all costs by the average number of spikes consumed by the addition operation).

We emphasize that this analysis is not meant to be a calculation of the actual brain computational costs (that are unknown), but a proof of concept that these algorithms could be implemented in the brain and that, if the costs are computed based on plausible neural-spiking costs, the scaling of inaccuracy vs. complexity would still be power-law with exponent changing with noise and volatility in the same way as in Fig. 4.

Measures of model performance

We used different measures of model performance. In Figs. 4, 5, 7A, 8A, and Supplementary Figs. 2, 4, 9A–C we used the “inaccuracy” of the models. Inaccuracy is defined as the difference in mean squared error (computed over ten 5000-time-long instances of the Gaussian or Bernoulli change point process) between the predictions of the model and those of the Bayesian ideal observer, normalized by the Bayesian benchmark, for each combination of volatility h and noise R:

$$\mathcal{I}(h, R) = \frac{E(h, R) - E_{Bayes}(h, R)}{E_{Bayes}(h, R)} \tag{32}$$

A vanishing inaccuracy implies that the inference strategy performs as well as the Bayesian model. In Fig. 4 and Supplementary Fig. 2, across task conditions, we fit the log-inaccuracy of the parametric models of Fig. 1 (with optimally chosen parameters) versus their log-complexity and found a power-law relationship between the two quantities

$$\mathcal{I}(h, R) = a(h, R)\mathcal{C}(h, R)^{-b(h, R)} \tag{33}$$

with an exponent that depends on volatility (h) and noise (R) in the underlying change-point process.

To better visualize this effect, we also defined the Accuracy as the ratio

$$\mathscr{A}(h, R) = \frac{E_{Bayes}(h, R)}{E(h, R)} \tag{34}$$

By definition, $0 \leq \mathscr{A}(h, R) \leq 1$.

In Fig. 6 and Supplementary Fig. 5, we used the “performance” of the models, a smooth decreasing function of inaccuracy. For Fig. 6:

$$\mathscr{P}(h, R) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma_r}e^{-\frac{\mathscr{I}(h, R)^2}{2\sigma_r^2}} \tag{35}$$

See Supplementary Information for other examples of performance functions and associated results. Combining Eqs. 33 and 35, then dividing by the complexity associated with a given level of inaccuracy from the power-law fits, yields Eq. 3.

Psychophysics experiments

We performed two psychophysics experiments: in the first one, we tested subjects’ estimates in the continuous Gaussian change point-task; in the second one, we tested subjects’ predictions in the discrete Bernoulli change-point task. All participants gave informed consent and the protocol was approved by the University of Pennsylvania. Participants were compensated at a rate of $10 per hour.

In the first experiment, we recruited 169 subjects using the Amazon Mechanical Turk crowdsourcing website. The subjects performed an estimation task (which from pilot studies was less confusing than a Gaussian prediction task). The task was presented as a card game. On each trial, the subject was shown a card number (corresponding to an observation x_t in Fig. 2) drawn from a card deck with Gaussian noise centered around the deck number. The deck number (corresponding to the mean μ_t in Fig. 2) was hidden to the subjects and changed at random times with a constant rate h. At the change-points, the deck number was resampled from a normal distribution with constant mean at 2500 ($\bar{\mu}$ in Fig. 2). Non-integer values for card and deck numbers were approximated to the nearest integer. Subjects were asked to guess, on each trial, from which deck the card was being picked (i.e., to estimate the generative mean). The ratio between the standard deviation of the card numbers around their deck number and the standard deviation of the deck numbers around their mean represents the noise parameter R.

Before starting the game, subjects were given written instructions about the statistics of card and deck numbers. First, they were informed that deck numbers were picked in a 0–5000 range, that decks in the middle of the range (around 2500) were most likely, and decks at the extremes (around 0 and 5000) were least likely. Second, they were informed that each deck had cards with numbers that were near but not always the same as the deck number. Third, they were informed that: (1) the randomness of the numbers in each deck, and (2) how often the decks switched without notice were both held constant within each block of trials but could change in the different blocks. The values of the randomness (noise) and the switching rate (volatility) were both explicitly indicated to the subjects via thermometer screen icons during task performance.

All 169 subjects were exposed to 40 training trials, in which we gave them the correct answer to familiarize them with the task. Training was followed by 360 test trials, in which we did not show the correct answer. To increase the subjects' motivation to perform as well as possible over many trials we provided feedback, after each trial, about the subject's error relative to six different competitors (the Bayesian model, the two Mixture models, the Sliding-Window, Delta-Rule, and Memoryless models) in a game-like setting. The same type of feedback was given for all noise/volatility conditions. Each subject performed 3 blocks of 400 trials in total (training + test). The blocks differed in terms of their values of noise or volatility (Fig. 7A): 85 subjects performed 3 blocks at constant volatility ($h = 0.1$) and variable noise ($R = 0.01$, $R = 0.8$, $R = 4$), and 84 subjects performed 3 blocks at constant noise ($R = 1$) and variable volatility ($h = 0.08$, $h = 0.38$, $h = 0.8$). Trials in which subjects input a number outside the 0–5000 range prescribed for the card decks, or responded in less than 100 ms, were excluded from further analyses.

In the second experiment, we recruited 53 subjects using the Amazon Mechanical Turk crowdsourcing website. The subjects performed the Bernoulli prediction task, in the following game setting. Two jars (the two sources) contained beads of different colors: the first jar contained white/black beads in the $p/(1 - p)$ proportion (with $0.5 < p < 1$), and the second jar contained black/white beads in the same proportion. On each trial, a bead (observation) was drawn from an unknown jar, and the jar changed from the previous trial with probability h. Subjects had to guess the color of the next bead. Before starting the game, they were exposed to examples of the task as part of the instructions and were told that, to predict the color of the next bead, they had to infer and track the jar that was most likely going to produce the next bead. Each subject performed 4 blocks of 300 trials each, corresponding to the following 4 conditions: (1) $h = 0.05$ and $p = 0.98$; (2) $h = 0.05$ and $p = 0.86$; (3) $h = 0.05$ and $p = 0.57$; (4) $h = 0.4$ and $p = 0.86$ (conditions (1), (2), (3) had fixed volatility and increasing noise; conditions (2), (4) had fixed noise and increasing volatility). Trials in which subjects responded in < 100 ms were excluded from further analyses. No statistical methods were used to pre-determine sample sizes but our sample sizes are substantially larger than those reported in previous publications [3, 5, 63] and large enough to ensure validity of the t-tests used in Fig. 7.

Data analysis

GAUSSIAN CHANGE-POINT TASK.—Values of adaptivity and working-memory load for subjects and models (Fig. 7B–E) were obtained with the following analyses:

Integration kernels.: For each subject and noise/volatility condition, we computed an integration kernel (linear weighting function) for each set of trials having the same lag from a change-point. Thus, we considered the set of subject responses $\left\{R_{t_q}\right\}$ from all trials $t_q = t_q^{cp} + \Delta t$, with Δt a fixed lag, t_q^{cp} the q-th change-point trial, and q running from 1 to M (number of change-points in a given block occurring at $t_q^{cp} > n - \Delta t$, see below). From the response set $\left\{R_{t_q}\right\}$ we estimated the subject integration kernel for the lag Δt, by finding the weights $\{K_p, K_0, \ldots, K_n\}$ of the multiple linear regression model

$$R_{t_q} = K_p \bar{\mu} + \sum_{\tau=0}^{n} K_\tau x_{t_q - \tau} + \epsilon_{t_q}\,;\ q = 1, \ldots, M \tag{36}$$

K_p (the weight given to the prior $\bar{\mu}$) and $K_0, \ldots, K_n$ (the weights given to the $n + 1$ most recent observations) were obtained using the Matlab *lsqlin* function, which minimizes the sum of squared residuals $\sum_q \epsilon_{t_q}^2$ for the system of linear equations, with constraints $K_p + \sum_{\tau=0}^{n} K_\tau = 1$ and $0 \le K_i \le 1$, $i = \{p, 0, \ldots, n\}$. We used $n + 1 = 15$ predictors (in addition to the prior) for the results in Fig. 7. Our conclusions were robust against changes in the number of predictors (Supplementary Fig. 7). We estimated one integration kernel (set of weights) for each subject, block, and lag Δt. We excluded only a few cases in which the linear regression model had fewer equations than predictors (because of an insufficient number of change-points at $t_q^{cp} > n - \Delta t$), yielding underconstrained weights.

Integration time scales.: For each integration kernel, we computed the normalized cumulative weight of the most recent τ observations

$$C(\tau) = \frac{\sum_{i=0}^{\tau} K_i}{\sum_{i=0}^{n} K_i} \tag{37}$$

and the time scale at which this normalized cumulative weight reaches a fixed threshold θ

$$\tau_\theta = \min\{\tau \mid C(\tau) > \theta\} \tag{38}$$

τ_θ represents an integration time scale. For the results of Fig. 7, we used $\theta = 0.8$, so that τ_θ is the time scale that explains 80% of the subject's integration over recently observed data. Conclusions were in general robust against changes in the threshold θ (Supplementary Fig. 7). We computed one value of τ_θ for each integration kernel; i.e., for each subject, block, and lag Δt. We estimated the standard error on each τ_θ by bootstrapping the regression model, Eqs. 36. We used 200 bootstrap samples of the form $\mathbf{b} = \left\{\left(R_{t_{i_1}}, \mathbf{x}_{i_1}\right), \ldots, \left(R_{t_{i_m}}, \mathbf{x}_{i_m}\right)\right\}$, with $\mathbf{x}_{i_k} = \left(\bar{\mu}, x_{t_{i_k}}, \ldots, x_{t_{i_k} - n}\right)$, $i_1, \ldots, i_m$ a random sample (with replacement) of the integers 1 through M (see [64] for more details about the bootstrap procedure).

Adaptivity.: For each subject and each block of trials, we computed adaptivity as the variance of τ_θ across all lags $0 \le \Delta t \le n$ (because the integration time scale τ_θ can not be larger than the time scale n in the linear regression model). This metric of adaptivity quantifies how much the integration time scale changes as more and more card numbers are observed from the same card deck, in a given block of trials. To capture changes in individual adaptivity values across noise/volatility levels, we normalized each subject's adaptivity in any given condition by the maximum adaptivity for the same subject across the three conditions, then we averaged the normalized values across subjects. The few subjects

(3 for the variable noise conditions, 1 for the variable volatility conditions) for whom adaptivity was zero in all the three conditions were excluded from the analyses, because the normalized adaptivity was undefined. Standard errors on the average normalized adaptivity were computed as $\mathrm{SEM} = \sqrt{\sum_i \left(\frac{\partial\langle \text{norm. adaptivity}\rangle}{\partial \tau_{\theta, i}}\right)^2 \left(SE\left(\tau_{\theta, i}\right)\right)^2}$ with the sum running over all the τ_θ obtained for different subjects and lags Δt.

The theoretical estimates of adaptivity were obtained by considering the most-efficient model in each tested condition (Fig. 7A) and using the model simulated outputs, under the same sequences of observations shown to the subjects, in place of R_{t_q} in Eq. 36; τ_θ and normalized-adaptivity values were then obtained in the same way as the subject's values. For the main bars in Fig. 7B, the most-efficient model was defined as the simplest model in our hierarchy with inaccuracy $\mathscr{I} < 0.1$ (as in Fig. 5B); theoretical predictions were qualitatively conserved across a relatively wide range of tolerance levels (e.g., between 0.02 and 0.2, see Fig. 7B, dashed gray lines).

Working-memory load.: For each subject or most-efficient model and each block of trials, we computed working-memory load as the maximum τ_θ across all possible lags $0 \leq \Delta t \leq n$; i.e., the maximum integration time scale (relative to the threshold θ) used for any given condition of noise and volatility. We then normalized this value by the maximum working-memory load across conditions for each subject and averaged the normalized values across subjects. Error bars were computed as for adaptivity.

The probabilities that each of the models of Fig. 1 generated the data of a randomly selected subjects (Fig. 7FG) were obtained with the following method:

Model fitting.: We assumed that a given subject in a given condition made inferences $\mu_t^{(s)}$ using one of the models of Fig. 1 with Gaussian decision noise: $\mu_t^{(s)} = \mu_t^{(m)} + \eta_t$, where $\eta_t = \mathcal{N};(0, \sigma_{\text{noise}})$ and $\mu_t^{(m)}$ are the model's deterministic inferences. For each subject, model (M), and condition, we identified the values of the model parameters θ maximizing the log-likelihood of the subjects' responses under the model:

$$\log p\left(\mu_{1:T}^{(s)} \mid M, \theta\right) = -\sum_{t=1}^{T} \frac{\left(\mu_t^{(s)} - \mu_t^{(m)}\right)^2}{2\sigma_{\text{noise}}^2} - T \log \sigma_{\text{noise}} - \frac{T}{2} \log 2\pi \tag{39}$$

Parameters are: σ_{noise} for the Evidence, Prior and Bayesian models; σ_{noise} and one learning rate for the single Delta-Rule, single Sliding-Window, and Memoryless models; σ_{noise} and two learning rates for the Mixture models. We used an upper bound on decision noise of 500 (1/10 of the maximum response range and approximately the mean plus one standard deviation of the models' residuals in the training phase). Higher upper bounds would reduce overall model distinguishability.

To generate Supplementary Fig. 9, we included the Kalman Filter (KF) and a 2-level Hierarchical Gaussian Filter (HGF) in the analysis. Parameters for the KF are: σ_{noise} and the process noise (i.e., the variance of the diffusion process) [59, 60]. Parameters for the HGF are: σ_{noise}, the prior mean and variance for each level, two parameters controlling the process noise of the first level, and one parameter controlling the process noise of the second level [30, 61, 62].

Model log-evidence values.: The maximum log-likelihood values were used to compute the BIC approximation of the log-evidence for each model and subject:

$$\log\, p\left(\mu_{1:T}^{(s)} \mid M\right) = 2\,\log\, p\left(\mu_{1:T}^{(s)} \mid M, \theta\right) - k\,\log T \tag{40}$$

with k the number of parameters.

Model probabilities.: We aggregated the model log-evidence values across subjects using the method developed in [65], which uses a variational Bayesian approach to estimate the parameters of a Dirichlet posterior distribution over model probabilities. Fig. 7FG shows the expected model probabilities under the estimated posterior distribution. Unlike other methods (such as the Group Bayes Factor), this approach does not assume that all the subjects' data are generated by the same model and is robust to the presence of outliers.

Confusion analysis.: The confusion matrices of Supplementary Fig. 8 were generated as follows. (1) We simulated each of the eight models of Fig. 1 using parameter values (including decision noise) fit to the subjects' data. For each model, we run a number of simulations equal to the number of subjects, with each simulation composed of the same number of trials as in the experiment. (2) We fit each of the eight models to data from each simulation. (3) We computed the probabilities $p(x|y)$ that a given model x best fit the data of a randomly chosen simulation generated with model y, applying the same Bayesian model-selection method for group-level analyses that we used in Fig. 7FG.

BERNOULLI CHANGE-POINT TASK.—We identified the best-fitting model for each subject in two ways. In Fig. 8, we defined "best-fit" as the model that minimized the sum of squared residuals for a given subject. This definition is justified by the non-parametric nature of the three models considered for this task (Fig. 8A); p(model) in Fig. 8BC is the fraction of subjects that were best-fit by each of the models. In Supplementary Fig. 10, we used the same Bayesian model-selection criterion described for the Gaussian change-point task, after introducing decision-noise (parametrized by σ_{noise}) in model inferences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Alex Filipowicz for sharing the codes to run the psycophysical experiment for the Bernoulli task. We also thank Kamesh Krishnamurthy and Eugenio Piasini for interesting discussions, Andrea Cavagna and Alessandro Ingrosso for pointing out a possible connection between one of our results and spin-glass systems. GT was supported by the Swartz Foundation (Award 575556) and the Computational Neuroscience Initiative of the University of Pennsylvania, and is currently supported by Washington University in St. Louis. VB and JG are supported in part by NIH BRAIN Initiative grant R01EB026945. JG is also supported by R01 MH115557 and

NSF-NCS 1533623. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon request.

References

[1]. Rao RPN, Bayesian computation in recurrent neural circuits, Neural Computation 16 (1) (2004) 1–38. [PubMed: 15006021]

[2]. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD, The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks, Psychological Review 113 (4) (2006) 700. [PubMed: 17014301]

[3]. Fearnhead P, Liu Z, On-line inference for multiple changepoint problems, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (4) (2007) 589–605.

[4]. Shi L, Griffiths TL, Neural implementation of hierarchical Bayesian inference by importance sampling, in: Advances in Neural Information Processing Systems, 2009, pp. 1669–1677.

[5]. Brown SD, Steyvers M, Detecting and predicting changes, Cognitive Psychology 58 (1) (2009) 49–67. [PubMed: 18976988]

[6]. Gigerenzer G, Gaissmaier W, Heuristic decision making, Annual Review of Psychology 62 (2011) 451–482.

[7]. Wilson R, Nassar M, Gold J, A mixture of delta-rules approximation to Bayesian inference in change-point problems, PLoS Computational Biology 9 (7) (2013) e1003150. [PubMed: 23935472]

[8]. Legenstein R, Maass W, Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment, PLoS Computational Biology 10 (10) (2014) e1003859. [PubMed: 25340749]

[9]. Gershman SJ, Horvitz EJ, Tenenbaum JB, Computational rationality: A converging paradigm for intelligence in brains, minds, and machines, Science 349 (6245) (2015) 273–278. [PubMed: 26185246]

[10]. Ortega PA, Braun DA, Thermodynamics as a theory of decision-making with information-processing costs, Proc. R. Soc. A 469 (2153) (2013) 20120683.

[11]. Glaze CM, Filipowicz ALS, Kable JW, Balasubramanian V, Gold JI, A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment, Nature Human Behaviour 2 (3) (2018) 213.

[12]. Adams R, MacKay D, Bayesian online changepoint detection, arXiv preprint arXiv:0710.3742 (2007).

[13]. Wilson RC, Nassar MR, Gold JI, Bayesian online learning of the hazard rate in change-point problems, Neural Computation 22 (9) (2010) 2452–2476. [PubMed: 20569174]

[14]. Nassar MR, Wilson RC, Heasly B, Gold JI, An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment, Journal of Neuroscience 30 (37) (2010) 12366–12378. [PubMed: 20844132]

[15]. Heilbron M, Meyniel F, Subjective confidence reveals the hierarchical nature of learning under uncertainty, bioRxiv (2018) 256016.

[16]. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS, Learning the value of information in an uncertain world, Nature Neuroscience 10 (9) (2007) 1214. [PubMed: 17676057]

[17]. Sutton RS, Barto AG, Reinforcement Learning: An Introduction, MIT press, 1998.

[18]. Balasubramanian V, Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions, Neural Computation 9 (2) (1997) 349–368.

[19]. Barron A, Rissanen J, Yu B, The minimum description length principle in coding and modeling, IEEE Transactions on Information Theory 44 (6) (1998) 2743–2760.

[20]. Gutenkunst R, Waterfall J, Casey F, Brown K, Myers C, Sethna J, Universally sloppy parameter sensitivities in systems biology models, PLoS Computational Biology 3 (10) (2007) e189.

[21]. Transtrum MK, Qiu P, Model reduction by manifold boundaries, Physical Review Letters 113 (9) (2014) 098701. [PubMed: 25216014]

[22]. Fan Y, Gold JI, Ding L, Ongoing, rational calibration of reward-driven perceptual biases, Elife 7 (2018) e36018. [PubMed: 30303484]

[23]. Schwarz G, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461–464.

[24]. Zeng X, Song T, Zhang X, Pan L, Performing four basic arithmetic operations with spiking neural P systems, IEEE Transactions on Nanobioscience 11 (4) (2012) 366–374. [PubMed: 22893452]

[25]. Shenhav A, Musslick S, Lieder F, Kool W, Griffiths TL, Cohen JD, Botvinick MM, Toward a rational and mechanistic account of mental effort, Annual Review of Neuroscience 40 (2017) 99–124.

[26]. Vul E, Goodman N, Griffiths TL, Tenenbaum JB, One and done? Optimal decisions from very few samples, Cognitive Science 38 (4) (2014) 599–637. [PubMed: 24467492]

[27]. Schmidhuber J, Formal theory of creativity, fun, and intrinsic motivation (1990–2010), IEEE Transactions on Autonomous Mental Development 2 (3) (2010) 230–247.

[28]. Gold JI, Shadlen MN, Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward, Neuron 36 (2) (2002) 299–308. [PubMed: 12383783]

[29]. Krugel LK, Biele G, Mohr PNC, Li SC, Heekeren HR, Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions, Proceedings of the National Academy of Sciences 106 (42) (2009) 17951–17956.

[30]. Mathys C, Weber L, Hierarchical Gaussian filtering of sufficient statistic time series for active inference, in: International Workshop on Active Inference, Springer, 2020, pp. 52–58.

[31]. Lee S, Gold JI, Kable JW, The human as delta-rule learner, Decision 7 (1) (2020) 55.

[32]. Glaze CM, Kable JW, Gold JI, Normative evidence accumulation in unpredictable environments, Elife 4 (2015) e08825.

[33]. Walton ME, Behrens TEJ, Buckley MJ, Rudebeck PH, Rushworth MFS, Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning, Neuron 65 (6) (2010) 927–939. [PubMed: 20346766]

[34]. Sul JH, Jo S, Lee D, Jung MW, Role of rodent secondary motor cortex in value-based action selection, Nature Neuroscience 14 (9) (2011) 1202. [PubMed: 21841777]

[35]. Cover TM, Thomas JA, Elements of Information Theory, John Wiley & Sons, 2012.

[36]. Tishby N, Pereira FC, Bialek W, The information bottleneck method, arXiv preprint physics/0004057 (2000).

[37]. Canziani A, Paszke A, Culurciello E, An analysis of deep neural network models for practical applications, arXiv preprint arXiv:1605.07678 (2016).

[38]. Cheeseman PC, Kanefsky B, Taylor WM, Where the really hard problems are., in: IJCAI, Vol. 91, 1991, pp. 331–340.

[39]. Biroli G, Cocco S, Monasson R, Phase transitions and complexity in computer science: an overview of the statistical physics approach to the random satisfiability problem, Physica A: Statistical Mechanics and its Applications 306 (2002) 381–394.

[40]. Mitchell D, Selman B, Levesque H, Hard and easy distributions of SAT problems, in: AAAI, Vol. 92, 1992, pp. 459–465.

[41]. Zdeborová L, Statistical physics of hard optimization problems, Acta Physica Slovaca. Reviews and Tutorials 59 (3) (2009) 169–303.

[42]. Wilson RC, Nassar MR, Tavoni G, Gold JI, Correction: A mixture of delta-rules approximation to Bayesian inference in change-point problems, PLoS Computational Biology 14 (6) (2018) e1006210. [PubMed: 29944654]

[43]. Gerstner W, Kistler WM, Naud R, Paninski L, Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition, Cambridge University Press, 2014.

[44]. Schultz W, Dayan P, Montague PR, A neural substrate of prediction and reward, Science 275 (5306) (1997) 1593–1599. [PubMed: 9054347]

[45]. Goldman-Rakic PS, Cellular basis of working memory, Neuron 14 (3) (1995) 477–485. [PubMed: 7695894]

[46]. Gläscher J, Büchel C, Formal learning theory dissociates brain regions with different temporal integration, Neuron 47 (2) (2005) 295–306. [PubMed: 16039570]

[47]. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N, A hierarchy of temporal receptive windows in human cortex, Journal of Neuroscience 28 (10) (2008) 2539–2550. [PubMed: 18322098]

[48]. Bernacchia A, Seo H, Lee D, Wang XJ, A reservoir of time constants for memory traces in cortical neurons, Nature Neuroscience 14 (3) (2011) 366. [PubMed: 21317906]

[49]. Scott BB, Constantinople CM, Akrami A, Hanks TD, Brody CD, Tank DW, Fronto-parietal cortical circuits encode accumulated evidence with a diversity of timescales, Neuron 95 (2) (2017) 385–398. [PubMed: 28669543]

[50]. Runyan CA, Piasini E, Panzeri S, Harvey CD, Distinct timescales of population coding across cortex, Nature 548 (7665) (2017) 92. [PubMed: 28723889]

[51]. Meder D, Kolling N, Verhagen L, Wittmann MK, Scholl J, Madsen KH, Hulme OJ, Behrens TEJ, Rushworth MFS, Simultaneous representation of a spectrum of dynamically changing value estimates during decision making, Nature Communications 8 (1) (2017) 1942.

[52]. Joshi S, Gold JI, Pupil size as a window on neural substrates of cognition, Trends in Cognitive Sciences 24 (6) (2020) 466–480. [PubMed: 32331857]

[53]. Arnsten AFT, Wang MJ, Paspalas CD, Neuromodulation of thought: Flexibilities and vulnerabilities in prefrontal cortical network synapses, Neuron 76 (1) (2012) 223–239. [PubMed: 23040817]

[54]. Yerkes RM, Dodson JD, The relation of strength of stimulus to rapidity of habit-formation, Journal of Comparative Neurology and Psychology 18 (5) (1908) 459–482.

[55]. Cools R, D'Esposito M, Inverted-U–shaped dopamine actions on human working memory and cognitive control, Biological Psychiatry 69 (12) (2011) e113–e125. [PubMed: 21531388]

[56]. Aston-Jones G, Cohen JD, An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance, Annu. Rev. Neurosci 28 (2005) 403–450. [PubMed: 16022602]

[57]. Griffiths TL, Vul E, Sanborn AN, Bridging levels of analysis for probabilistic models of cognition, Current Directions in Psychological Science 21 (4) (2012) 263–268.

[58]. Fusi S, Asaad WF, Miller EK, Wang XJ, A neural circuit model of flexible sensorimotor mapping: Learning and forgetting on multiple timescales, Neuron 54 (2) (2007) 319–333. [PubMed: 17442251]

[59]. Kalman RE, Bucy RS, New results in linear filtering and prediction theory, Journal of Basic Engineering.

[60]. Welch G, Bishop G, An introduction to the Kalman filter, chapel Hill, NC, USA (1995).

[61]. Mathys C, Daunizeau J, Friston KJ, Stephan KE, A Bayesian foundation for individual learning under uncertainty, Frontiers in Human Neuroscience 5 (2011) 39. [PubMed: 21629826]

[62]. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, Stephan KE, Uncertainty in perception and the Hierarchical Gaussian Filter, Frontiers in Human Neuroscience 8 (2014) 825. [PubMed: 25477800]

[63]. Ossmy O, Moran R, Pfeffer T, Tsetsos K, Usher M, Donner TH, The timescale of perceptual evidence integration can be adapted to the environment, Current Biology 23 (11) (2013) 981–986. [PubMed: 23684972]

[64]. Efron B, Tibshirani RJ, An Introduction to the Bootstrap, CRC press, 1994.

[65]. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ, Bayesian model selection for group studies, Neuroimage 46 (4) (2009) 1004–1017. [PubMed: 19306932]

[66]. McDonnell JV, Martin J, Markant D, Coenen A, Rich AS, Gureckis TM, psiturk (version 1.02) [software]. New York, ny: New York University (2012).

[67]. De Leeuw JR, jspsych: A JavaScript library for creating behavioral experiments in a Web browser, Behavior Research Methods 47 (1) (2015) 1–12. [PubMed: 24683129]

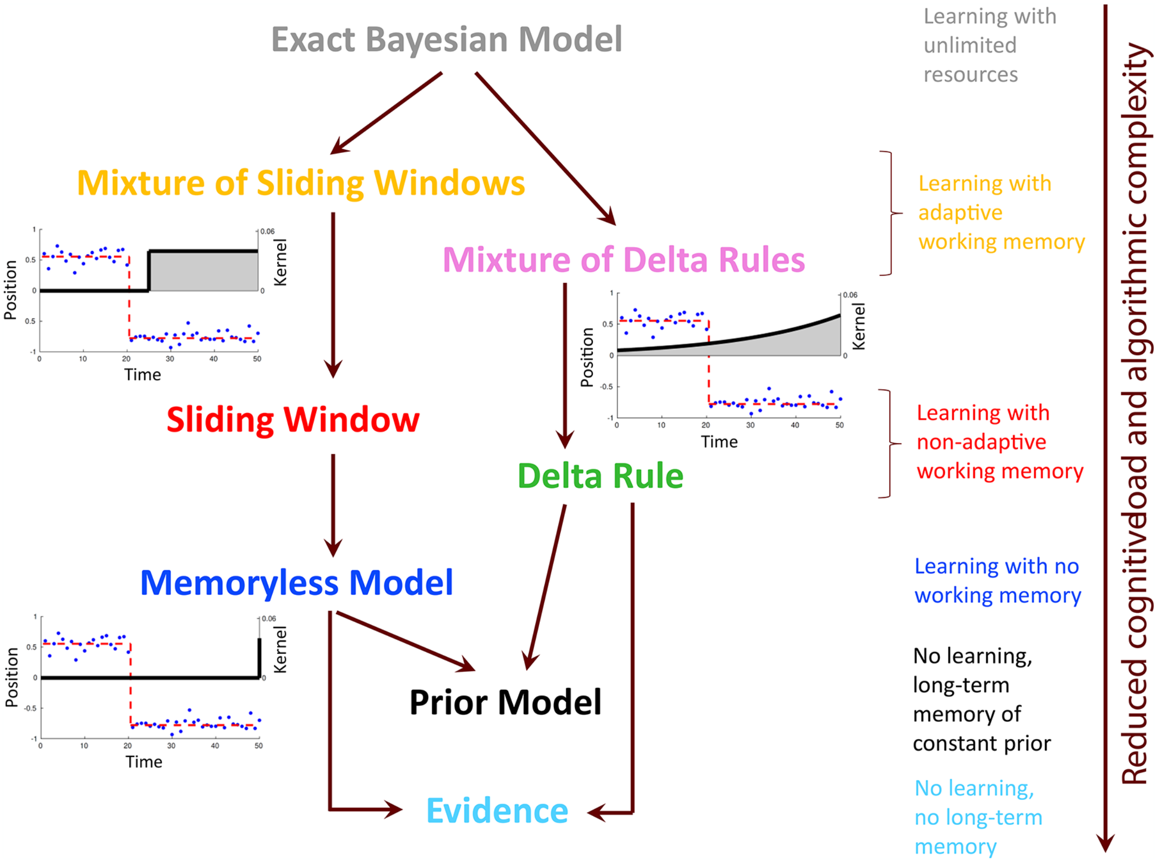
Exact Bayesian Model
Learning with unlimited resources
Mixture of Sliding Windows
Mixture of Delta Rules
Learning with adaptive working memory
Sliding Window
Delta Rule
Learning with non-adaptive working memory
Memoryless Model
Learning with no working memory
Prior Model
No learning, long-term memory of constant prior
Evidence
No learning, no long-term memory
Reduced cognitivdoad and algorithmic complexity
Position
Kernel
Time
0.06
0
1
0.5
-0.5
-1
10
20
30
40
50

Figure 1. A hierarchy of cognitive functions maps to a hierarchy of inference strategies.

Two nested families of inference strategies of decreasing algorithmic complexity can be derived from the exact Bayesian approach by progressively reducing requirements of memory and adaptivity (see also Supplementary Fig. 1). We illustrate this approach in the context of inference from noisy observations (blue dots) of a latent variable μ_t (red dashed lines). See text for model descriptions and Methods for model details. The decrease in algorithmic complexity over this hierarchy of strategies mirrors a corresponding decrease in cognitive load (legend on the right-hand side).

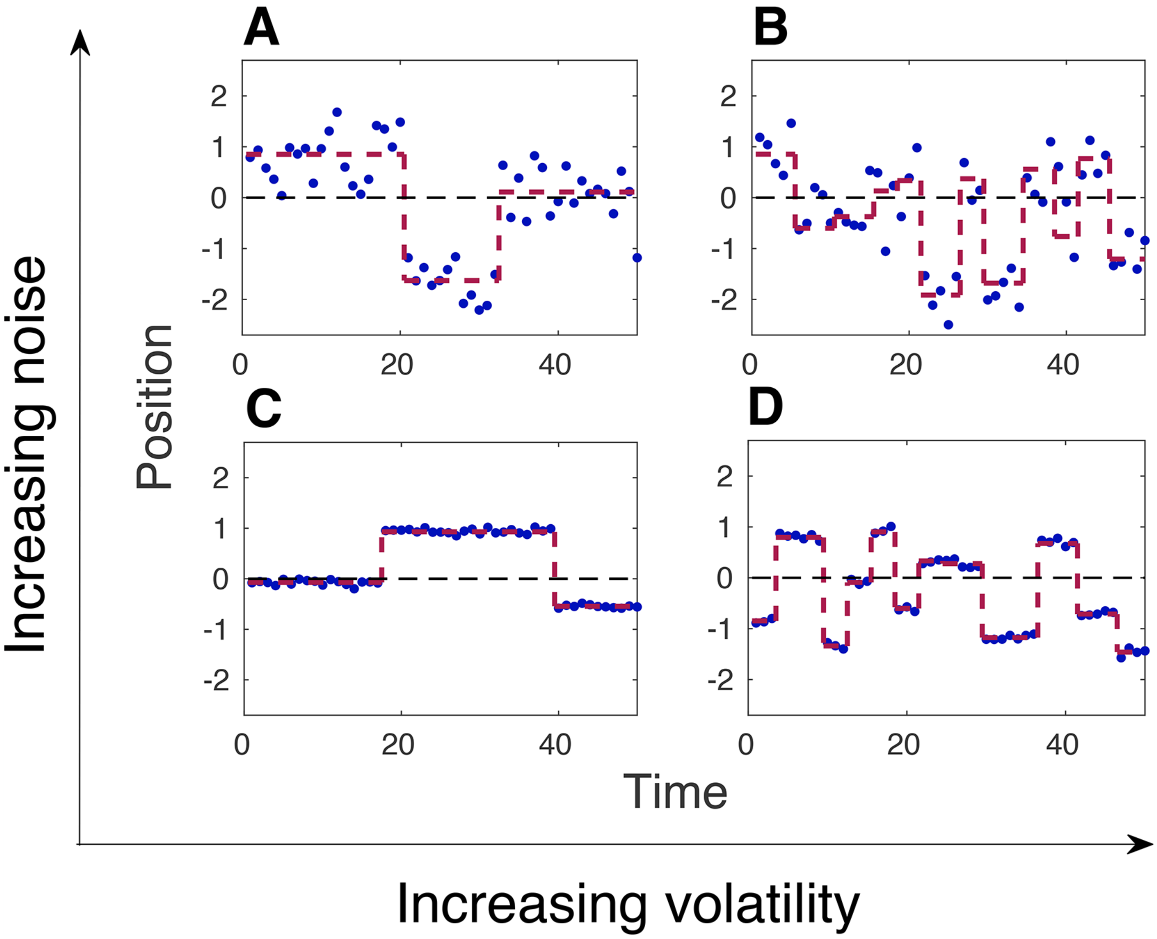

Figure 2. Gaussian change-point processes.

Observations x_t (blue dots) are generated from a source positioned at μ_t (dashed red line) with Gaussian noise (SD = σ). The source is hidden to the observer and undergoes change-points at random times with probability h (volatility). At the change-points, μ_t is resampled from a Gaussian distribution centered at $\bar{\mu}$ (dashed black line, stable over time) and with SD = σ_0 = 1. Different panels show processes with different volatility (increasing from left to right) and noise $R = \sigma/\sigma_0$ (increasing from bottom to top): (**A**): h = 0.06, R = 0.45; (**B**): h = 0.24, R = 0.45; (**C**): h = 0.06, R = 0.05; (**D**): h = 0.24, R = 0.05.

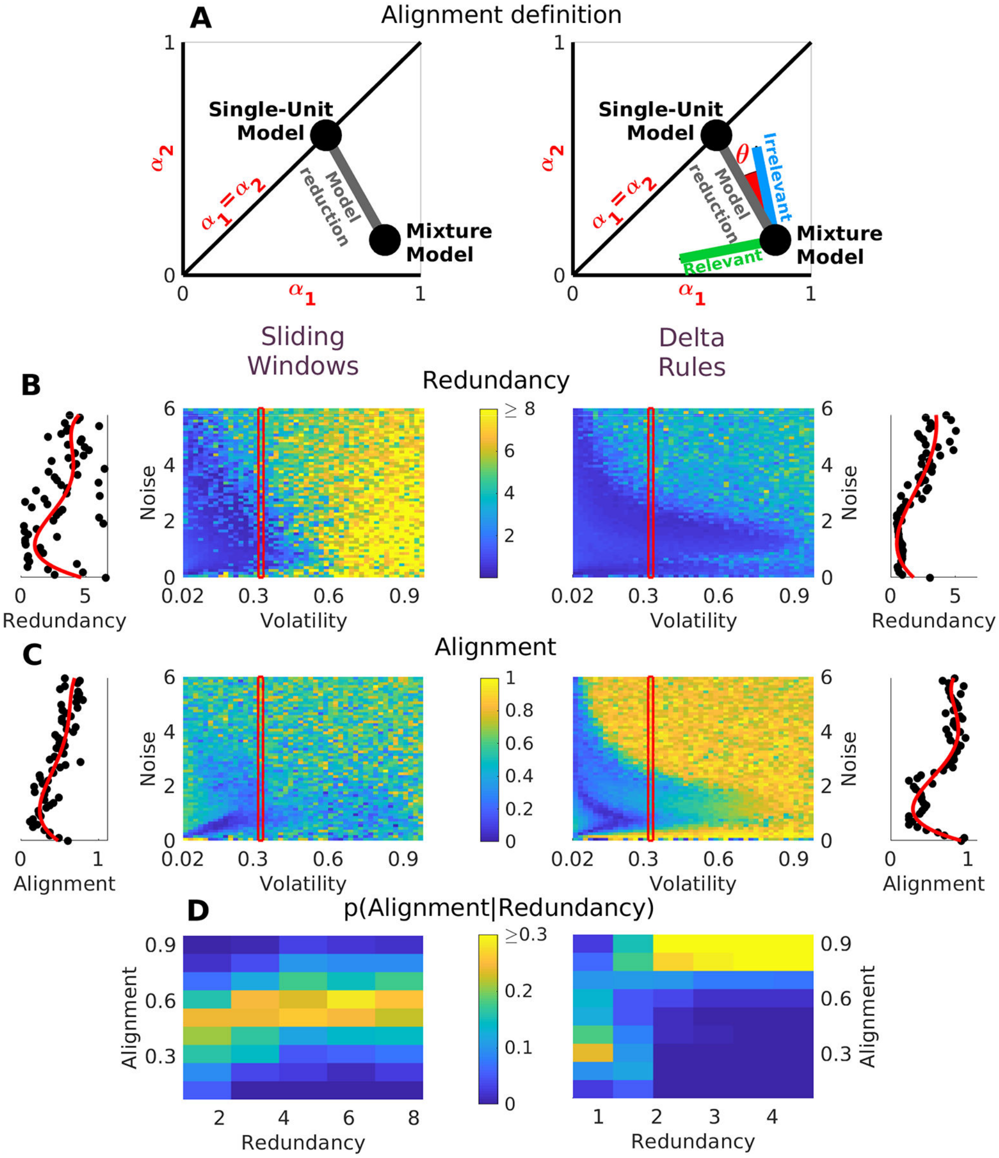

Figure 3. Adaptive models reduce to calibrated simpler strategies when variability is low or high. (**A**): Computation of the Alignment. (Left) Two-dimensional parameter space of the Mixture models with two units defined by learning rates α_1 and α_2, and the embedded unidimensional space of the nested single-unit models (diagonal line $\alpha_1 = \alpha_2$). The optimal Mixture model and optimal single-unit model (black dots) are indicated along with the parameter deformation leading from one to the other (gray line). (Right) Relevant and irrelevant parameter deformations that maximally or minimally change the prediction error moving away from the optimal adaptive Mixture model. Alignment is defined as the normalized angle θ between the irrelevant deformation and the direction to the best non-adaptive single-unit model. The prediction error used to compute Alignment is estimated over 5000 time steps of the process for each h/R values. (**B**): Redundancy of the adaptive Mixture models (left: Mixture of two Sliding Windows; right: Mixture of two Delta Rules) for a range of volatility and noise values in a change-point detection task (Fig. 2). The

same error function as in (A) is used to compute Redundancy. Slices through the red inset windows are shown to the left and right (red lines: 4th-order polynomial fits). (**C**): Alignment of the irrelevant parameter deformation towards the non-adaptive nested single-unit model, plotted as in B. (**D**): Probability distribution of Alignment values conditioned on Redundancy, sampled over tested volatility and noise values.

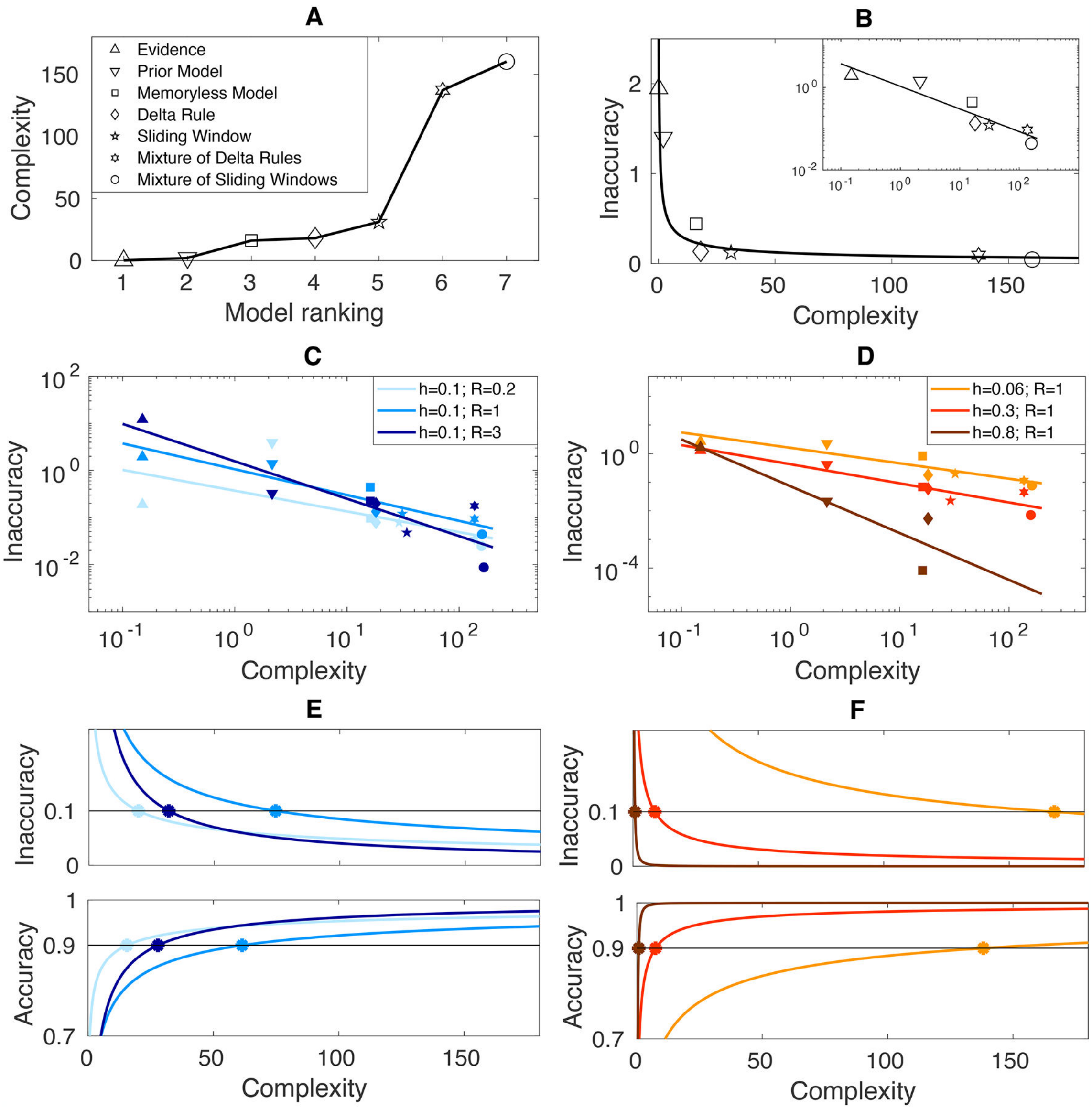

Figure 4. Diminishing returns from increasing complexity.

(**A**): Algorithmic complexity (Eq. 30) for models in Fig. 1. The exact Bayesian model has infinite complexity by our measure and is not shown. (**B**): Inaccuracy (Eq. 32) decreases as a power law in the complexity (Eq. 30), shown here for volatility and noise levels h = 0.1 and R = 1. Inset: linear fit on a log-log scale. See also Supplementary Fig. 3A–C for goodness-of-fit statistics. The exponent in the power law varies with (**C**) noise and (**D**) volatility. Inaccuracy is computed over ten 5000-time-long instances of the change point process. (**E**): Scaling of inaccuracy and accuracy (Eq. 34) with complexity for fixed volatility and varying noise (Eq. 33). Color code and scaling exponents for each condition taken from panel (C). Horizontal black lines indicate the threshold for performance within 10% of the Bayesian optimum. Intercept with the scaling curve for each task condition indicates the minimum model complexity required to reach the performance threshold. (**F**):

Same as panel (E) for fixed noise and varying volatility. Color code and scaling exponents taken from panel (D).

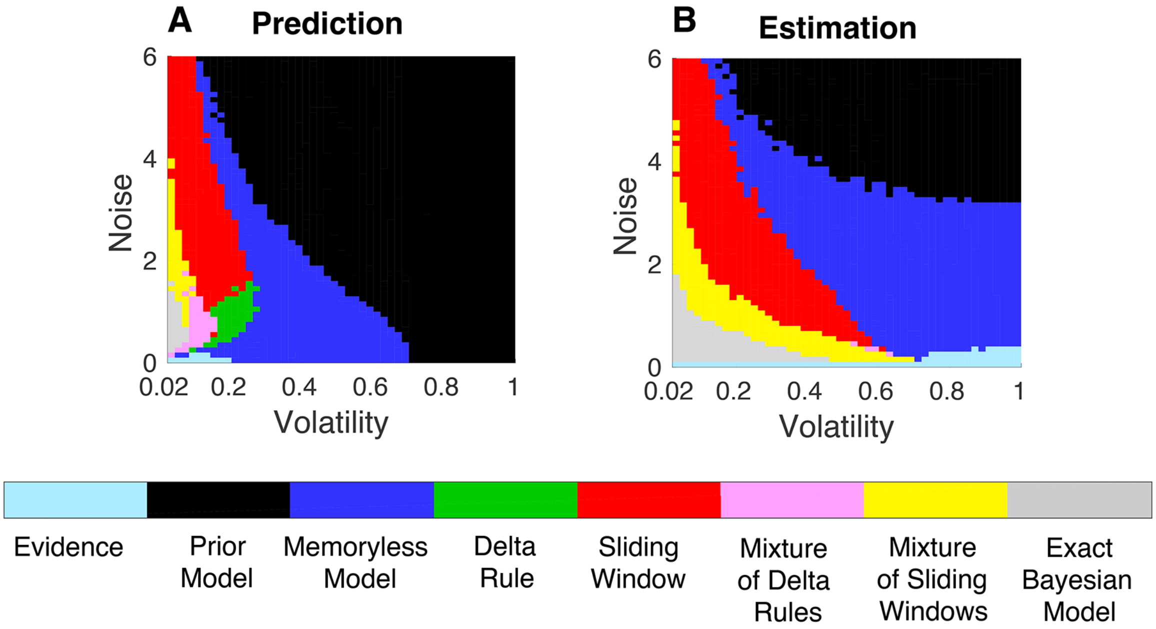

Figure 5. Simple inference strategies are usually sufficient.

The color map shows the simplest strategy achieving performance within 10% of the Bayesian optimum (inaccuracy < 0.1) for each combination of volatility and noise in the prediction (**A**) and estimation (**B**) tasks. See also Supplementary Fig. 4.

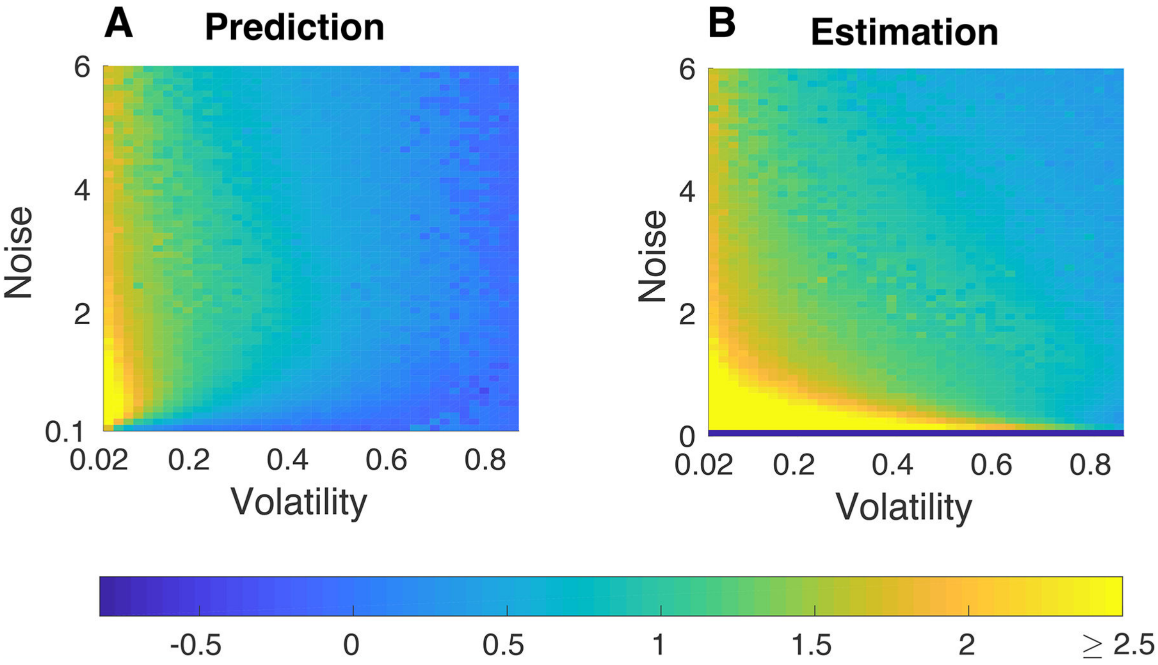

Figure 6. Optimal cognitive engagement.

The colormaps show $\log_{10} \mathscr{C}_{opt}$ (Eq. 4) as a function of volatility and noise, for the prediction (**A**) and the estimation (**B**) tasks; $\sigma_r = 0.1$. High cognitive engagement is optimal only at low volatility and intermediate noise.

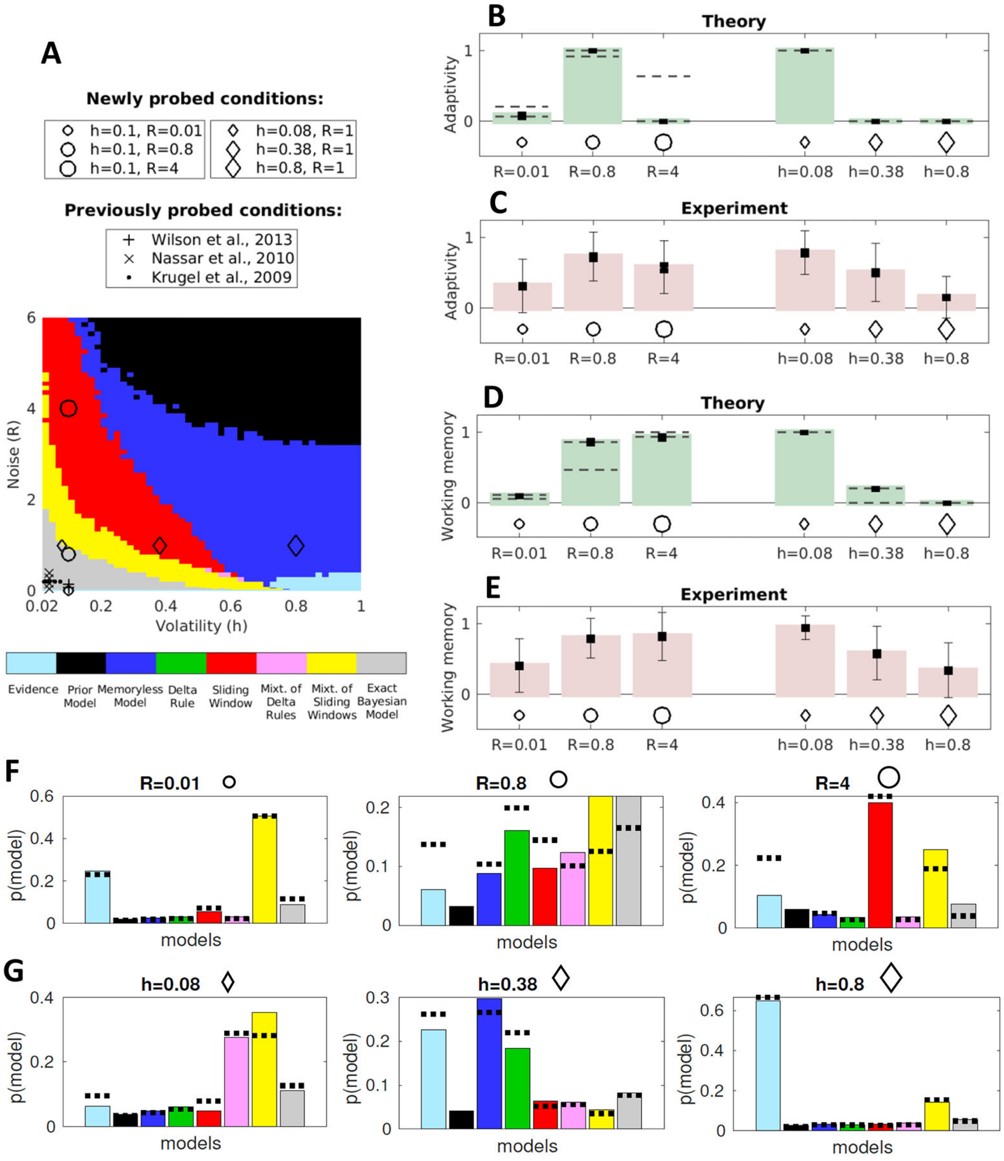

Figure 7. Subjects switch between simple and complex strategies as predicted by the theory in the Gaussian estimation task.

(**A**): Map of the volatility/noise conditions probed in this experiment compared to conditions probed in previous experiments (see legend). Background colors indicate the simplest model with inaccuracy $\mathcal{I} < 0.1$ (i.e., the most efficient model for tolerance = 0.1) at each point of the volatility/noise plane for the estimation task performed by the subjects. (**B**) and (**C**): Mean normalized adaptivity ± SEM (small error bars) for the theoretical most-efficient model (**B**) and 82 human subjects (**C**) performing the estimation task for each of the three noise (*R*)/volatility (*h*) conditions. Normalized adaptivity values were computed by fitting multiple linear regression models to data from 360 trials per subject and volatility/noise condition (details in Method). SEM were obtained by propagating the errors on the integration time scales estimated from the linear regressions (Methods). For the colored bars, the most-efficient model was defined as the simplest model with $\mathcal{I} < 0.1$; dashed gray lines represent the range of values obtained using different tolerances (0.02 – 0.2; note the broad range for high noise). Thin error bars in C represent the standard deviation of the normalized adaptivity across subjects. Both the theory and data showed peak adaptivity at intermediate noise (left, one-tailed t-test, $p = 7 \cdot 10^{-12}$ for the intermediate vs. low and $p = 0.0038$ for the intermediate vs. high noise comparisons) and low volatility (right, $p = 10^{-6}$ for the low vs. intermediate and $p = 10^{-28}$ for the low vs. high comparisons). (**D**) and (**E**): Mean normalized working-memory load from theory (**D**) and 82 human subjects (**E**) performing the estimation task for each of the three noise/volatility conditions (plotted as in (B) and (C)). For both the theory and the data, the working-memory load is smaller at low noise (one-tailed t-test, $p = 4{\cdot}10^{-12}$ for both low vs. intermediate and low vs. high noise comparisons) and decreases with increasing volatility (one-tailed t-test, $p = 5 \cdot 10^{-13}$ for low vs. intermediate, $p = 2{\cdot}10^{-24}$ for low vs. high, $p = 3{\cdot}10^{-5}$ for intermediate vs. high volatility comparisons). (**F**) and (**G**): Probabilities that each of the eight models of Fig. 1 (color code as in (A)) generated the data of a randomly chosen subject, in each noise (**F**) and volatility (**G**) condition. Bars indicate results for the best-performing subjects (with inaccuracy $\mathscr{I} <$ 75th percentile across all tested conditions); dotted lines represent the values obtained for all subjects.

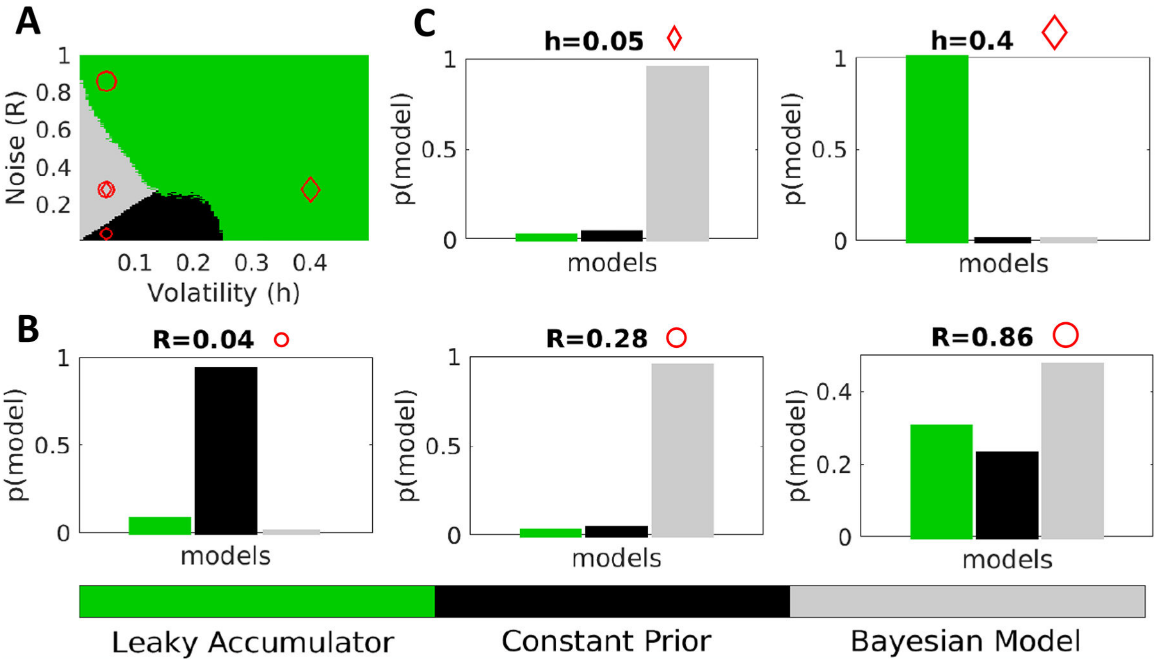

Figure 8. Subjects switch between simple and complex strategies as indicated by the theory in the Bernoulli prediction task.

(**A**): Theoretical predictions for this task. Three models of decreasing complexity are considered: the Bayesian model, the constant Prior, and the Leaky-Accumulator model. The colormap indicates the simplest model with inaccuracy $\mathcal{I} < 0.1$ (i.e., the most efficient model for tolerance = 0.1) at each point of the volatility/noise plane. Three conditions of increasing noise ($R = 0.04$, $R = 0.28$, $R = 0.86$, red circles) and two conditions of increasing volatility ($h = 0.05$, $h = 0.4$, red diamonds) were tested in the experiment. (**B**) and (**C**): Probabilities that each of the three models (color code as in (A)) minimized the sum of squared residuals of a randomly chosen subject, in each noise (**B**) and volatility (**C**) condition. Data from 53 subjects, 300 trials per subject and condition.