



# Human inference reflects a normative balance of complexity and accuracy

Gaia Tavoni<sup>1</sup>✉, Takahiro Doi<sup>2</sup>, Chris Pizzica<sup>3</sup>, Vijay Balasubramanian<sup>3,4,5</sup> and Joshua I. Gold<sup>3,5</sup>

**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.**

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 modelled in ways that differ widely in form, accuracy and complexity<sup>1–8</sup>. This diversity leaves open basic questions about the relevance of these models to cognition. The goal of the present study is 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<sup>6,9,10</sup>. We reasoned that computational complexity in models of inference can represent a cognitive cost (for example, 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 on the basis of all previous observations<sup>3,11–13</sup>. 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<sup>7</sup>, by constructing two nested families of models that generated increasingly simple approximations of 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)<sup>14–16</sup>. We used several complementary approaches to identify two fundamental principles. The first principle is a law of diminishing returns, whereby

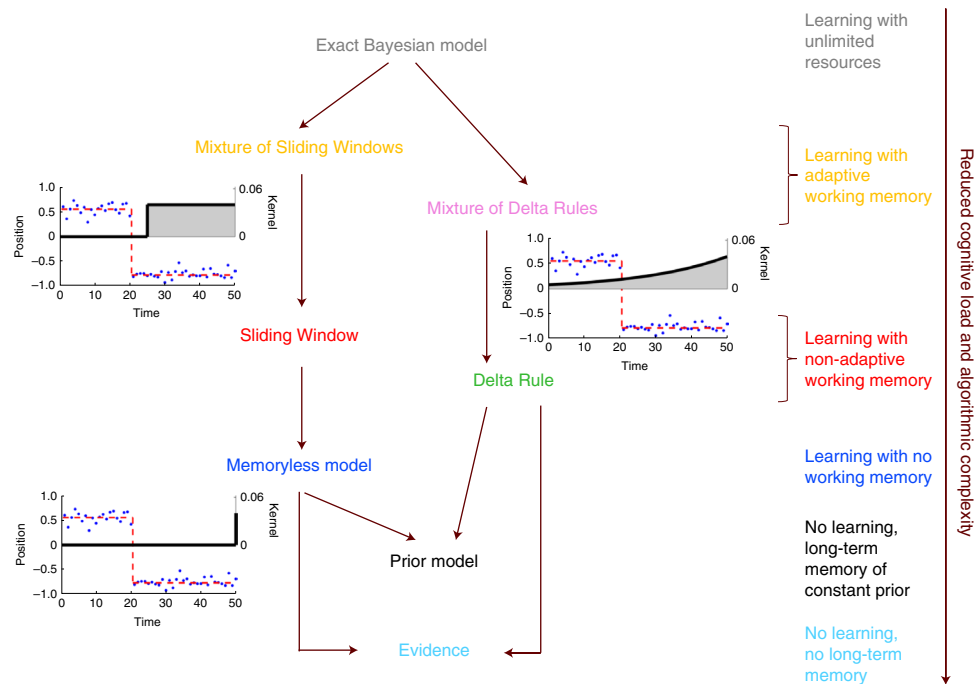
gains in accuracy become progressively smaller with increasing complexity, regardless of the amount of uncertainty in the environment. The second principle is a non-monotonic relationship between uncertainty and the complexity of the most efficient model: simple models are the most efficient when uncertainty 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 behavioural experiments to show that these principles apply to human behaviour. Overall, these results provide 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 such as non-adaptive delta rules that are often described as implementing a ‘model-free’ form of learning<sup>7,12,14,16,17</sup>. 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  $x_{1:t} = \{x_1, \dots, x_t\}$ . 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  $\mu_t$  at a time  $t$ , generates observations with Gaussian variability (Fig. 2)<sup>7,14</sup>. 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.

<sup>1</sup>Department of Neuroscience, Washington University in St. Louis, St. Louis, MO, USA. <sup>2</sup>Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA. <sup>3</sup>Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA. <sup>4</sup>Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, PA, USA. <sup>5</sup>These authors jointly supervised this work: Vijay Balasubramanian, Joshua I. Gold. ✉e-mail: [gaia.tavoni@wustl.edu](mailto:gaia.tavoni@wustl.edu)



**Fig. 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  $\mu_t$  (red dashed lines). See the text for the model descriptions and the Methods for the model details. The decrease in algorithmic complexity over this hierarchy of strategies mirrors a corresponding decrease in cognitive load (legend on the right side).

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  $\mu_t$  given the data observed over just the run-length  $r_t$ . By multiplying these probabilities and summing over the possible run-lengths, we can compute the probability that the source is  $\mu_t$  given all the data:

$$p(\mu_t|x_{1:t}) = \sum_{r_t=1}^t p(\mu_t|r_t)p(r_t|x_{1:t}). \quad (1)$$

The Bayesian model computes  $p(\mu_t|r_t)$  and  $p(r_t|x_{1:t})$  exactly<sup>3,12,13</sup>. The optimal Bayesian estimate of the source,  $\hat{\mu}_t$ , is then simply the expected value of  $\mu_t$  in the conditional distribution (equation (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

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

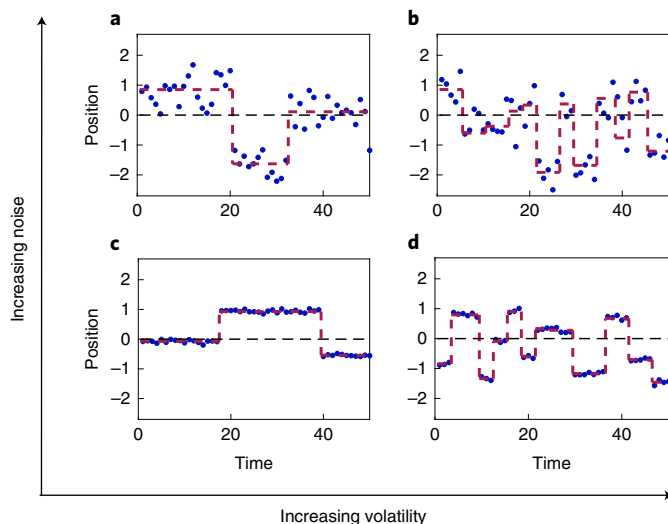
where  $\bar{\mu}$  is the asymptotic average value of the source (Fig. 2). The Bayesian model thus balances prior belief against evidence integrated over temporal windows of all possible lengths, with each window weighted by the likelihood that the latent variable has been stable over that duration. 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 (equation (1); ref. 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 the Methods for the 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) the 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 side of 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<sup>18,19</sup>. 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



**Fig. 2 | Gaussian change-point processes. a–d.** Observations  $x_t$  (blue dots) are generated from a source positioned at  $\mu_t$  (dashed red line) with Gaussian noise (s.d. =  $\sigma$ ). The source is hidden to the observer and undergoes change-points at random times with probability  $h$  (volatility). At the change-points,  $\mu_t$  is resampled from a Gaussian distribution centred at  $\bar{\mu}$  (dashed black line, stable over time) and with s.d. =  $\sigma_0 = 1$ . The different panels show processes with different volatility (increasing from left to right) and noise  $R = \sigma/\sigma_0$  (increasing from bottom to top):  $h = 0.06$ ,  $R = 0.45$  (a);  $h = 0.24$ ,  $R = 0.45$  (b);  $h = 0.06$ ,  $R = 0.05$  (c);  $h = 0.24$ ,  $R = 0.05$  (d).

*E.* The eigenvalues of  $H$  indicate how much  $E$  increases when moving away from the minimum error in parameter space in the direction of the eigenvectors of  $H^{20}$ . A high Redundancy thus indicates that the full model is reducible to a simpler one through a procedure that removes the least relevant degree of freedom (for example, the “manifold boundary approximation” method of ref. <sup>21</sup>). 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 the 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 is high so that inference is difficult, making complex strategies ineffective (Fig. 3b). Furthermore, when the models are redundant, their Alignment is also high (Fig. 3c,d). This relationship, which is particularly strong for Delta-Rule models, implies that the brain could, in principle, use local estimates of performance gradients<sup>22</sup> 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 online 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

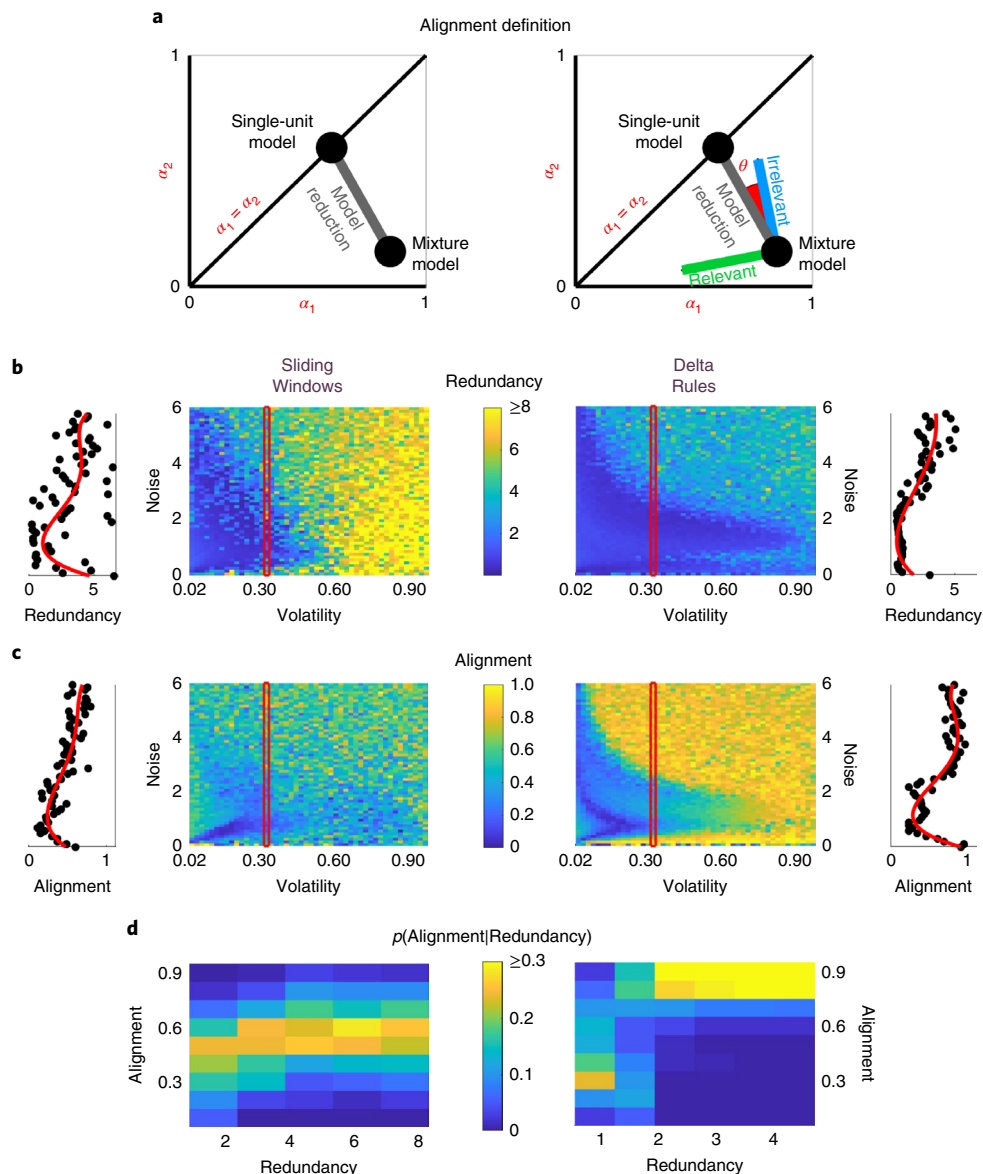
them (Fig. 4a and Supplementary Table 1). For the non-parametric Bayesian model, this cost grows linearly with the number of observations because the entire past provides a probabilistic context for each prediction or estimate, and thus the algorithmic complexity diverges to infinity. The other models 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<sup>18,19,23</sup>. 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 useful only 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. 4e,f) and to very different model implementations (we recapitulated the results from Fig. 4 using neural-network implementations<sup>24</sup>; 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 (for example, the 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.

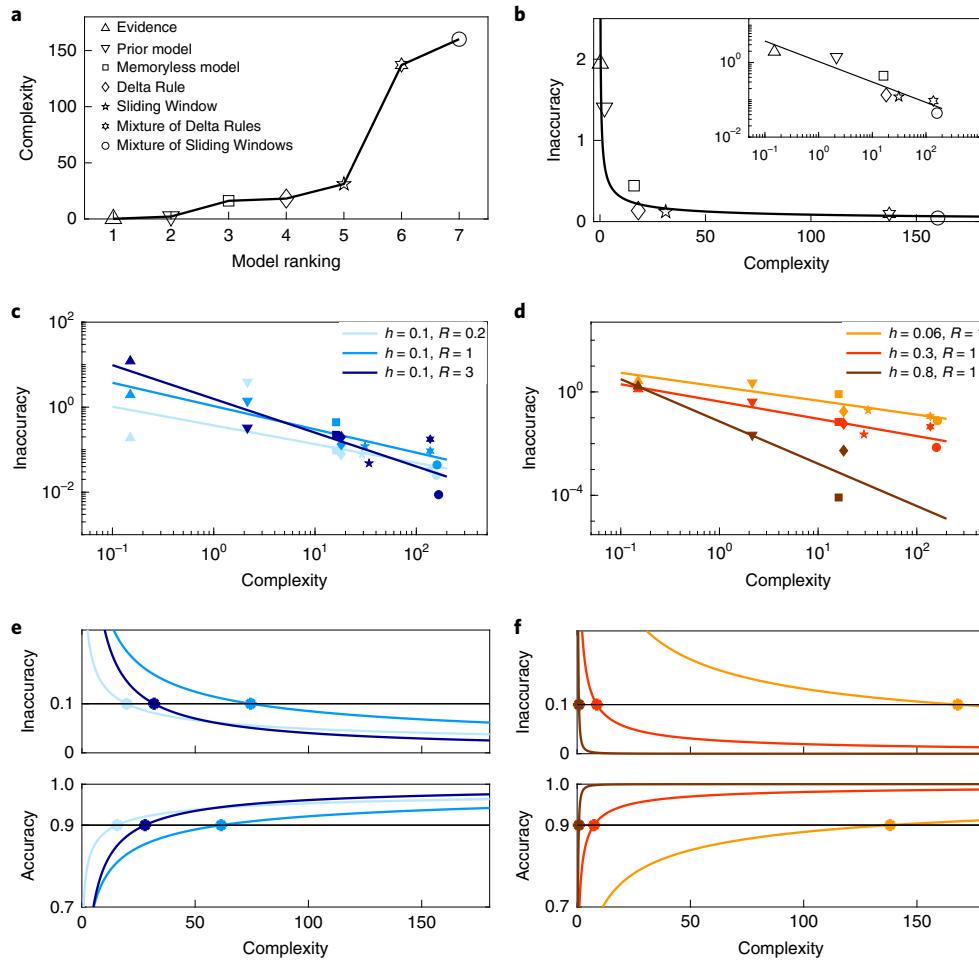
For estimation problems, slightly different patterns emerge (Fig. 5b). For high noise and volatility, and for low noise across volatilities, certain simple strategies are nearly as effective as complex



**Fig. 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  $\alpha_1$  and  $\alpha_2$ , 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 (grey 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  $\theta$  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 5,000 time steps of the process for each combination of  $h$  and  $R$  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 (the red lines indicate fourth-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.

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 compared with 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 one 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  $\sigma_r$ , so that substantial rewards are obtained when inaccuracy is  $O(\sigma_r)$  or smaller.



**Fig. 4 | Diminishing returns from increasing complexity.** **a**, Algorithmic complexity (equation (30)) for the models in Fig. 1. The exact Bayesian model has infinite complexity by our measure and is not shown. **b**, Inaccuracy (equation (32)) decreases as a power law in the complexity (equation (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. **c,d**, The exponent in the power law varies with noise (**c**) and volatility (**d**). Inaccuracy is computed over ten 5,000-time-long instances of the change-point process. **e**, Scaling of inaccuracy and accuracy (equation (34)) with complexity for fixed volatility and varying noise (equation (33)). The colour code and scaling exponents for each condition are taken from **c**. The horizontal black lines indicate the threshold for performance within 10% of the Bayesian optimum. The intercept with the scaling curve for each task condition indicates the minimum model complexity required to reach the performance threshold. **f**, Same as **e** for fixed noise and varying volatility. The colour code and scaling exponents are taken from **d**.

We can then use this function to derive an expression for expected performance ( $\mathcal{P}$ ) per unit complexity ( $\mathcal{C}$ ) for each noise–volatility pair:

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

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 engagement, the ratio in equation (3) represents a trade-off between reward and cognitive cost per prediction or estimation. Because algorithmic complexity (equation (30)) can also be thought of as a qualitative estimate of the time required to make an inference, the ratio (equation (3)) can also be interpreted as an estimate of the reward one can obtain per unit time<sup>25–28</sup>, which is a meaningful fitness function from an evolutionary and behavioural perspective.

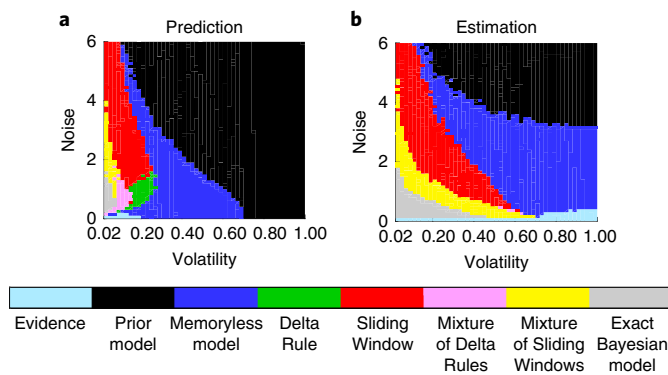
The performance per unit cost can be optimized by maximizing the expression on the right side of equation (3) with respect to the complexity  $\mathcal{C}$ . This gives

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

for the complexity, or cognitive cost, of the optimal inference strategy. Figure 6 represents  $\log \mathcal{C}_{\text{opt}}$  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  $\sigma_r$  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  $\mathcal{C}_{\text{reflex}}$  (equation (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).





**Fig. 5 | Simple inference strategies are usually sufficient.** **a,b**, The colour maps show 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.

**Human inferences are consistent with the theory.** We performed two distinct psychophysical experiments to relate our theory to human inference. In the first experiment, human participants 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 85 participants was tested in three conditions of fixed (low) volatility and variable noise (circles in Fig. 7a). A separate group of 84 participants 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<sup>7,14,29</sup> (small markers in Fig. 7a) that require complex, adaptive inference according to our theory. After training, participant behaviour was on average sensitive to the different values of noise and volatility that we tested (Supplementary Fig. 6).

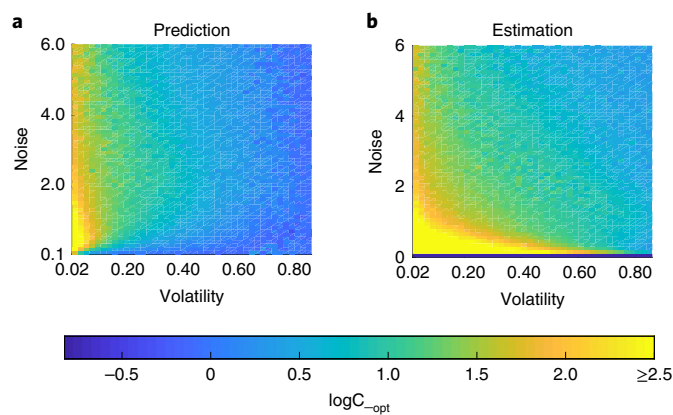
The participants tended to adjust both adaptivity (their use of flexible timescales for linear integration of past observations) and working memory (their maximum integration timescale) across changes in noise or volatility in a manner that reflected key features of our theory (Fig. 7b–e). In our theory, adaptivity is most useful for estimation tasks with intermediate noise and low volatility. Accordingly, the participants 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 participants 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 participant values, which probably 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 participants to solve inference problems.

We emphasize that the purpose of this analysis is to show that participants 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 in Fig. 1 best explained the participants' behaviour in each noise and volatility condition. We assumed that a given participant in each condition made inferences using a model in Fig. 1 with Gaussian decision noise, and we identified the values of the model parameters maximizing the log-likelihood of the participant's responses. The maximum log-likelihood values were used to compute the Bayesian information criterion approximation of the log-evidence for each model and participant. We aggregated the model log-evidence values across participants using a Bayesian approach described in ref.<sup>30</sup> and the Methods. The pattern of selected models matched key predictions of our theory (including when using different inaccuracy tolerances; Supplementary Fig. 4g–i; see also Supplementary Fig. 8 for confusion analyses). 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 or 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 behavioural data after including two additional models, the Kalman filter (KF) and the hierarchical Gaussian filter (HGF). In contrast to the models in 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<sup>31–33</sup>. Both models are adaptive and use 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 the KF and HGF to the model set. From a theoretical standpoint, the models in Fig. 1 were more effective than the KF and HGF in practically all volatility and noise conditions and across different tolerances to errors. From an experimental standpoint, the models that best explained the behavioural data in each condition remained those shown in Fig. 7f,g.

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, 53 participants were asked to predict the next observation. Thus, two important differences between this experiment and the previous one are the discrete as



**Fig. 6 | Optimal cognitive engagement.** **a, b**, The colour maps show  $\log C_{\text{opt}}$  (equation (4)) as a function of volatility and noise, for the prediction (**a**) and estimation (**b**) tasks;  $\sigma_r = 0.1$ . High cognitive engagement is optimal only at low volatility and intermediate noise.

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. The participants 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 behavioural results from this task also closely matched theoretical predictions. Here we considered three models that solve this task with decreasing accuracy and complexity<sup>34</sup>: (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 behavioural responses of almost every participant were best explained by the model predicted by the theoretical map (Fig. 8a). For high noise, behaviour 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 theoretically and experimentally supported 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 nested framework, we were able to highlight their interrelatedness while defining precisely how their particular, distinguishing

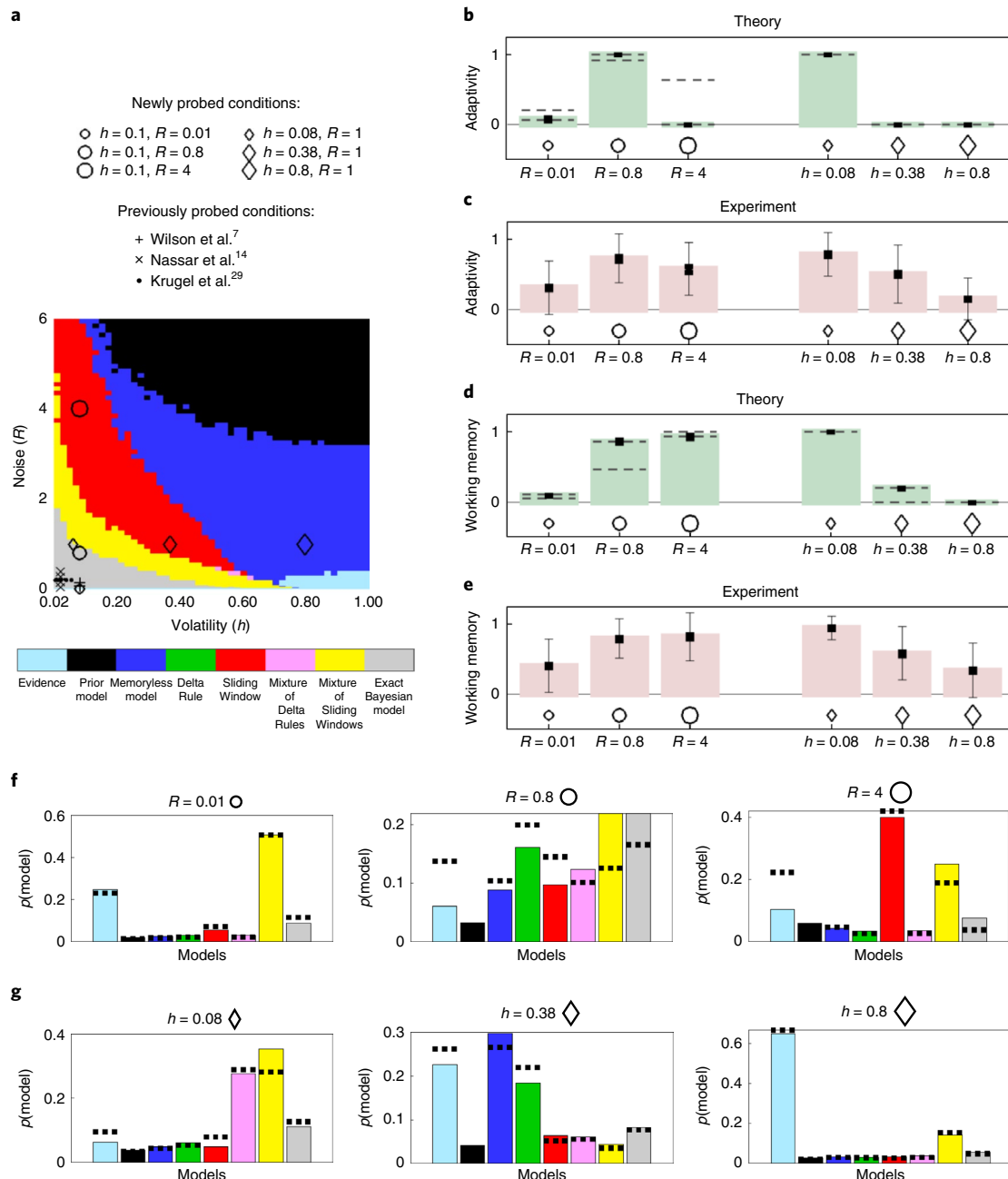
features contribute to this trade-off. We also focused on classical inference problems, for which we probed noise and volatility conditions that extended the ranges considered in previous studies<sup>7,14,29</sup> and more generally are relevant to tasks that include both forms of uncertainty, such as commonly used reversal-learning tasks<sup>16,35,36</sup>. Below we discuss our results relative to other computational, behavioural and neural findings, focusing on the insights provided by our work and potential future directions.

Each of our models is characterized by both its complexity and its inaccuracy compared with the exact Bayesian model. By analysing two nested families of models, we identified a power-law scaling of inaccuracy with complexity:  $\mathcal{I} \propto 1/C^b$ . 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  $\mathcal{R}$  (ref. 37). 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<sup>9</sup>, create a sort of information bottleneck<sup>38</sup>. Specifically, in simple contexts, the rate is measured in bits as the mutual information,  $I(X; \hat{X})$ , between the input  $X$  and output  $\hat{X}$  of an information channel<sup>37</sup>. In this formulation, the distortion for a Gaussian channel (similar to our Gaussian source) scales as  $\log \mathcal{D} \sim -\mathcal{R}$ , for distortions smaller than the variance of the samples. We similarly showed that inaccuracy scales as  $\log \mathcal{I} \sim -\log C$ , 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<sup>11</sup>. 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<sup>39</sup>.

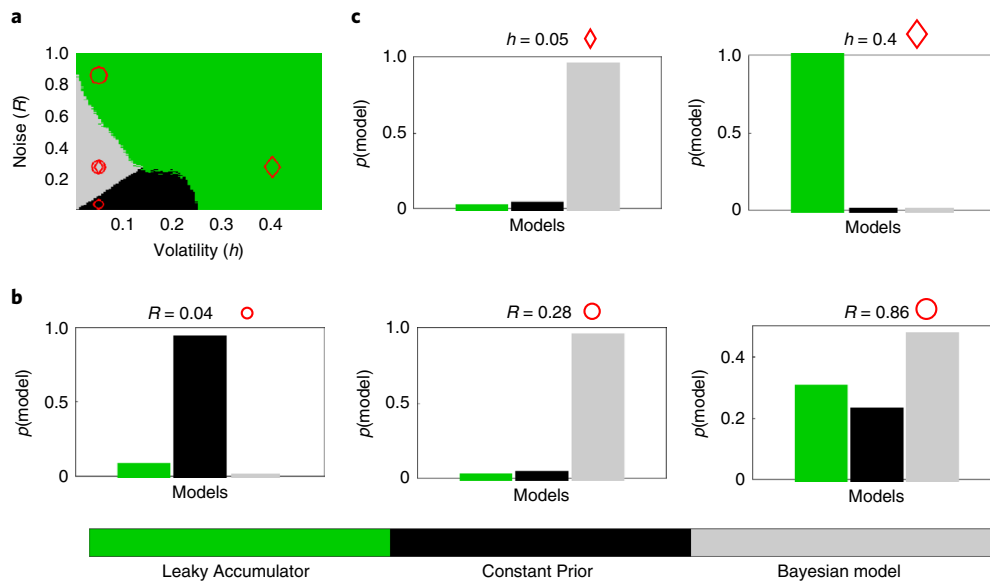
We also showed that complex strategies that use adaptive processes and/or working memory are necessary only for a restricted range of conditions, characterized by low volatility and moderate noise, with working memory being useful across a slightly wider set of uncertainty levels (particularly towards 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 colouring, the travelling 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<sup>40–43</sup>. In a broad sense, this order parameter plays a role similar to task difficulty or environmental uncertainty in our inference task.

Our behavioural 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).



**Fig. 7 | Human participants switch between simple and complex strategies as predicted by the theory in the Gaussian estimation task. a**, Map of the volatility and noise conditions probed in this experiment compared with conditions probed in previous experiments (see the legend). The background colours indicate the simplest model with inaccuracy  $\mathcal{I} < 0.1$  (that is, the most efficient model for tolerance = 0.1) at each point of the volatility–noise plane for the estimation task performed by the participants. **b,c**, Mean normalized adaptivity  $\pm$  s.e.m. (small error bars) for the theoretical most efficient model (**b**) and 82/83 human participants (**c**) performing the estimation task for each of the three noise or volatility conditions. Normalized adaptivity values were computed by fitting multiple linear regression models to data from 360 trials per participant and volatility or noise condition (the details are in the Methods). Standard errors of the mean were obtained by propagating the errors on the integration timescales estimated from the linear regressions (Methods). For the coloured bars, the most efficient model was defined as the simplest model with  $\mathcal{I} < 0.1$ ; the dashed grey lines represent the range of values obtained using different tolerances (0.02–0.2; note the broad range for high noise). The thin error bars in **c** represent the standard deviation of the normalized adaptivity across participants. Both the theory and data showed peak adaptivity at intermediate noise (left, one-tailed  $t$ -test,  $P = 7 \times 10^{-12}$  for the intermediate versus low and  $P = 0.0038$  for the intermediate versus high noise comparisons) and low volatility (right,  $P = 10^{-6}$  for the low versus intermediate and  $P = 10^{-28}$  for the low versus high comparisons). **d,e**, Mean normalized working-memory load from theory (**d**) and 82/83 human participants (**e**) performing the estimation task for each of the three noise or 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,  $P = 4 \times 10^{-12}$  for both low versus intermediate and low versus high noise comparisons) and decreases with increasing volatility (one-tailed  $t$ -test,  $P = 5 \times 10^{-13}$  for low versus intermediate,  $P = 2 \times 10^{-24}$  for low versus high and  $P = 3 \times 10^{-5}$  for intermediate versus high volatility comparisons). **f,g**, Probabilities that each of the eight models in Fig. 1 (colour coded as in **a**) generated the data of a randomly chosen participant, in each noise (**f**) and volatility (**g**) condition. The bars indicate the results for the best-performing participants (with inaccuracy  $\mathcal{I} < 75$ th percentile across all tested conditions); the dotted lines represent the values obtained for all participants.





**Fig. 8 | Human participants 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 colour map indicates the simplest model with inaccuracy  $\mathcal{I} < 0.1$  (that is, the most efficient model for tolerance = 0.1) at each point of the volatility–noise plane. Three conditions of increasing noise ( $R = 0.04$ ,  $R = 0.28$  and  $R = 0.86$ , red circles) and two conditions of increasing volatility ( $h = 0.05$  and  $h = 0.4$ , red diamonds) were tested in the experiment. **b, c**, Probabilities that each of the three models (colour coded as in **a**) minimized the sum of squared residuals of a randomly chosen participant, in each noise (**b**) and volatility (**c**) condition. The data are from 53 participants, with 300 trials per participant and condition.

For example, in refs. <sup>7,44</sup>, participants performed a Gaussian change-point task similar to the one considered here, with volatility  $\sim 0.1$  and noise  $\sim 0.1$ . 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 ( $h \sim 0.1$ ,  $R \sim 0.1$ ) point in the volatility–noise plane falls in the small region where the Two-Delta-Rule model is the most efficient strategy for prediction tasks (Fig. 5). Likewise, in ref. <sup>14</sup>, participants performed a similar Gaussian change-point task with volatility  $\sim 0.04$  and noise between  $\sim 0.06$  and  $\sim 0.4$ . Their behaviour 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 ref. <sup>29</sup>, participants performed a probabilistic object-reversal task with object-reversal probabilities (analogous to volatility) that ranged between  $\sim 0.008$  and  $\sim 0.08$ , 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 participants 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 on the basis of 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 establish whether 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 ‘ $\alpha$ -synapses’ used widely in biophysical models of neuron spiking dynamics<sup>45</sup>. 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<sup>16,46</sup>. This implementation, compared with 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 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<sup>47</sup>. 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<sup>48–52</sup>. In our formulation, the estimates obtained from these different integration timescales are weighted optimally and combined to produce a single output<sup>7,44</sup>. Consistent with this idea, learning rates with more relevance to an ongoing estimate of choice values have been shown to explain

more variance in functional MRI signals<sup>53</sup>. 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<sup>54,55</sup>. 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<sup>55–58</sup>. 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 the task demands.

An alternative hypothesis on how adaptive Bayesian inference might be approximated by the brain is based on particle filters and importance sampling<sup>3,4,26,59</sup>. 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 probabilities. 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 online statistical inference, showing that they can be regarded as special cases of a single formalism. This 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 participants research was reviewed by the University of Pennsylvania Institutional Review Board and determined to be exempt as authorized by 45 CFR 46.104, category no. 2. All participants provided informed consent prior to participating. The participants were compensated at a rate of US\$10 per hour.

**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)<sup>7,14</sup>. 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  $\mu_t$ . 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 ( $p(\mu_t) = \mathcal{N}(\mu_t | \bar{\mu}, \sigma_0^2)$ ). The goal of an observer is to infer the current position of the source  $\mu_t$  from the history of observations up to time  $t$  (the estimation problem) or to predict the position of the source at the next time step  $\mu_{t+1}$  (the prediction problem). For each simulation and task condition, the parameters  $\bar{\mu}$ ,  $\sigma_0$  and  $\sigma$  are constant. The ratio ( $R = \sigma/\sigma_0$ ) is the noise parameter of the process ( $R = 1/\sqrt{\text{signal-to-noise ratio}}$ ). 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  $\mu_{t+1}$  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  $\mu_t^i$  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 equation (2))<sup>12</sup>.

For Gaussian processes, the posterior probability of the source  $\mu_t$  given run-length  $r_t$  is

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

where we have used the Bayes rule  $p(\mu_t | x_{t-r_t+1:t}) \propto p(x_t | \mu_t) p(\mu_t | x_{t-r_t+1:t-1})$

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  $\mu_t$  with mean  $\bar{\mu} = \frac{\chi_p}{\nu_p}$  and variance  $\sigma_0^2 = \frac{\sigma^2}{\nu_p}$ . Using the relation

$\mathcal{N}(\mu | \mu_1, \sigma_1^2) \mathcal{N}(\mu | \mu_2, \sigma_2^2) \propto \mathcal{N}\left(\mu \mid \frac{\mu_1 \sigma_2^2 + \mu_2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$ , we obtain:

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

with

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

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

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

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

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:

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

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

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

Substituting equations (9a)–(9c) into equation (8), we obtain:

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

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

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

Equation (10) simplifies to:

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

In conclusion, we can compute:

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

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

$$\hat{\mu}_t = \langle \mu_t \rangle_{p(\mu_t|x_{1:t})} = \sum_{r_t=1}^t p(r_t|x_{1:t}) \mu_{t-1}^{r_t} \quad (14)$$

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:

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

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

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

**Building a hierarchy of approximations of Bayesian inference.** The Bayesian computation is simplified by considering only a fixed set of run-lengths  $\{r_1, \dots, r_N\}$  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  $\mu_t$  on the basis of a sliding-window integration of past observations over the duration  $r_i$ , combined with prior information on the average value of the source  $\bar{\mu}$ :

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

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 equation (17) corresponds to equation (7) in the Bayesian model, with  $r_i = r_i$ . Equation (17) can also be expressed recursively as:

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

with effective learning rate  $\alpha_i$  and initial condition

$\mu_{\lfloor r_i \rfloor+1}^{r_i} = \frac{1}{\nu_p + r_i} (\nu_p \bar{\mu} + dx_1 + \sum_{k=2}^{\lfloor r_i \rfloor+1} x_k)$ . Estimates computed by each unit are summed with a relative weight set adaptively by the posterior probabilities  $p(r_i|x_{1:t})$  (see the next section). Therefore, low (high) volatility in the world will lead to preferential use of long (short) sliding windows<sup>16,60</sup>. 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:

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

When we express 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:

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

with learning rates  $\alpha_i$  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  $\sim r_i$  time steps back in the past  $(1 - d)x_{t-\lfloor r_i \rfloor} + dx_{t-\lfloor r_i \rfloor-1}$  (equation (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 the 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  $\mu_t$  as a weighted average between the present observation  $x_t$  and the average source  $\bar{\mu}$  (Dirac delta kernel):

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

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 ( $\bar{\mu}$ ), on the basis of 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, and 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  $\mu_t$  and a prediction of  $\mu_{t+1}$ . 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 hierarchical 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  $\alpha_i$  that minimized the 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 5,000-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 ten different random instances (of 5,000 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  $\alpha = \frac{1}{\nu_p + 1}$  and is independent of volatility.

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

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

where the transition probabilities  $p(r_i|r_j)$  approximate  $p(r_i|r_{i-1})$  of the exact Bayesian model<sup>7,44</sup>:

$$p(r_i|r_j) = hp(r_i|r_j, \text{CP}) + (1-h)p(r_i|r_j, \text{no CP}) \quad (23)$$

We sort the  $N$  model run-lengths in ascending order:  $r_1 < r_2 < \dots < r_N$ . 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  $r_1$ :

$$p(r_i|r_j, \text{CP}) = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases} \quad (24)$$

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; and (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:

$$p(r_i|r_j, \text{no CP}) = \begin{cases} \frac{1}{r_{j+1} - r_j} & \text{if } i = j + 1 \\ 1 - \frac{1}{r_{j+1} - r_j} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

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

$$p(r_i|r_j, \text{no CP}) = \begin{cases} 1 & \text{if } i = j + 1 \\ 0 & \text{otherwise} \end{cases} \quad (26)$$

For  $j = N$ :

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

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

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

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

where  $\lambda_{\min}$  and  $\lambda_{\max}$  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 5,000 time steps of the process for each choice of  $h$  and  $R$ ; and 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 (equations (17) and (18)) and (2) the timescale of the exponential evidence-integration kernel in the Delta Rules (equations (19) and (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  $\delta\alpha = \alpha^{(1)} - \alpha^{(2)}$ , where  $\alpha^{(2)} = (\hat{\alpha}_1, \hat{\alpha}_2)$  is the two-component vector of the optimal parameters of the Mixture model, and  $\alpha^{(1)} = (\hat{\alpha}, \hat{\alpha})$  is the vector with both components equal to the optimal parameter of the non-adaptive nested model. The vector  $\delta\alpha$  is directed along the parameter transformation collapsing the best Mixture model into the best nested single-unit model. Thus, we can define

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

where  $0 \leq \theta \leq \pi$  is the angle between the irrelevant eigenvector and the direction of  $\delta\alpha$  (Fig. 3a). By definition,  $0 \leq \text{Alignment} \leq 1$ , and Alignment is a function of volatility  $h$  and noise  $R$ . To reduce numerical noise, in Fig. 3, Redundancy and Alignment were averaged over ten 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:

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

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  $C_{\text{reflex}}$  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 (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  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T N_t^i$ , 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  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T N_t^i = 0$  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  $\hat{\mu}_t = \bar{\mu} + \alpha(x_t - \bar{\mu})$ , which requires  $\langle N^A \rangle = 3$  (1 sum, 1 difference, 1 product),  $\langle N^S \rangle = 2$  (to store  $\bar{\mu}$  and  $\alpha$ ) and  $\langle N^R \rangle = 3$  (to read  $\bar{\mu}$  twice and  $\alpha$  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 the model 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  $\hat{\mu}_t = \hat{\mu}_{t-1} + \alpha(x_t - (1-d)x_{t-[r]} - dx_{t-[r]-1})$  (with  $\alpha = \frac{1}{\nu_p + r}$ ), 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  $t+[r]$  and  $t+[r]+1$ );  $\langle N^R \rangle = 6$  operations to read, at every  $t$ ,  $\hat{\mu}_{t-1}$ ,  $\alpha$ ,  $d$  (twice),  $x_{t-[r]}$  and  $x_{t-[r]-1}$ ; and finally  $\langle N^S \rangle = [r] + 4$  operations to store, at every  $t$ ,  $\alpha$ ,  $d$ ,  $\hat{\mu}_t$  (for one iteration) and  $x_t$  (for a duration of  $[r] + 1$  iterations). Because of the dependence on  $r(h, R)$  (the timescale 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 timescales, 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 (the Evidence, Prior and Memoryless models) or only a memory of the previous estimate (the 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 two summations, over  $N$  terms each, required to compute each of the  $N$  adaptive  $p(r_i|x_{1:t})$  (equation (22)): (1) the summation over  $N$  run-lengths  $j$  appearing at the numerator of equation (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  $\sim N$  (not as  $\sim N^2$ ) because only the  $N$  probabilities  $p(r_i|x_{i-1})$  are carried forward to the next iteration of the algorithm to compute the new  $p(r_i|x_{i+1})$ , whereas the individual addends of the summations mentioned above do not need to be memorized. Finally, the leading-order term of  $\langle N^F \rangle$  ( $2N^2$ ) arises because the computation of the adaptive  $p(r_i|x_{i-1})$  requires maintaining in memory, at every iteration, the  $N \times N$  matrices of the transition probabilities  $p(r_i|r_j, \text{CP})$  and  $p(r_i|r_j, \text{no CP})$  (equations (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  $\hat{\mu}_{t+1} = (1-h)\hat{\mu}_t + h\bar{\mu}$ . Each prediction thus requires four more algorithmic operations than each estimation, three more reading operations (to retrieve from memory  $h$  twice and  $\bar{\mu}$  once) and either two more storing operations for the Delta Rule and Sliding Window (to store both  $h$  and  $\bar{\mu}$ ) or just one more storing operation for the Memoryless model (to store  $h$ , as this model is already required 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 KF (algorithm of refs. <sup>61,62</sup>) requires  $\langle N^A \rangle = 9$ ,  $\langle N^W \rangle = 4$ ,  $\langle N^R \rangle = 12$  and  $\langle N^F \rangle = 2$  mean operations per inference. An HGF with  $n_i$  levels (algorithm of refs. <sup>31,32,63</sup>) requires  $\langle N^A \rangle = 29n_i - 22$ ,  $\langle N^W \rangle = 6n_i - 3$ ,  $\langle N^R \rangle = 31n_i - 22$  and  $\langle N^F \rangle = 6n_i - 1$ . We used  $n_i = 2$  in our analyses.

**Neural-spiking parallel implementation.** In Supplementary Fig. 2, we considered plausible neural-network implementations of the four arithmetic operations (addition, subtraction, multiplication and division). These neural circuits, called neural-spiking parallel systems, and their operating principles are presented in ref. <sup>24</sup>.

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

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

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

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

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

We obtained the neural cost of the square root operation ( $\sqrt{1+a}$ ) from the costs of the other four operations using the Taylor expansion of the square root up to the second order. We assigned a cost of two spikes to exponential operations, because in neural-spiking P systems any output is encoded in two 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 (that is, the specific numbers being summed, multiplied and so on (equations (31a)–(31d))). 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 (that is, 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 (which are unknown) but a proof of concept that these algorithms could be implemented in the brain and that, if the costs are computed on the basis of plausible neural-spiking costs, the scaling of inaccuracy versus complexity would still be a power law with the 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 and 8a and Supplementary Figs. 2, 4 and 9a–c, we used the ‘inaccuracy’ of the models. Inaccuracy is defined as the difference in mean squared error (computed over ten 5,000-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  $R$ :

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

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 in Fig. 1 (with optimally chosen parameters) versus their log-complexity and found a power-law relationship between the two quantities:

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

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

$$\mathcal{A}(h, R) = \frac{E_{\text{Bayes}}(h, R)}{E(h, R)} \quad (34)$$

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

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

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

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

**Psychophysics experiments.** We performed two psychophysics experiments.

In the first one, we tested the participants’ estimates in the continuous Gaussian change-point task; in the second one, we tested the participants’ predictions in the discrete Bernoulli change-point task.

In the first experiment, we recruited 169 participants using the Amazon Mechanical Turk crowdsourcing website. The participants 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 participant was shown a card number (corresponding to an observation  $x_i$  in Fig. 2) drawn from a card deck with Gaussian noise centred on the deck number. The deck number (corresponding to the mean  $\mu_i$  in Fig. 2) was hidden to the participants 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 2,500 ( $\bar{\mu}$  in Fig. 2). Non-integer values for card and deck numbers were approximated to the nearest integer. The participants were asked to guess, on each trial, from which deck the card was being picked (that is, 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, the participants were given written instructions about the statistics of card and deck numbers. First, they were informed that deck numbers were picked in a 0–5,000 range, that decks in the middle of the range (around 2,500) were the most likely and that decks at the extremes (around 0 and 5,000) were the 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 participants via thermometer screen icons during task performance.

All 169 participants 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 participants’ motivation to perform as well as possible over many trials, we provided feedback after each trial about the participant’s error relative to six different competitors (the Bayesian model, the two Mixture models, the Sliding-Window model, the Delta-Rule model and the Memoryless model) in a game-like setting. The same type of feedback was given for all noise and volatility conditions. Each participant performed three blocks of 400 trials in total (training + test). The blocks differed in terms of their values of noise or volatility (Fig. 7a): 85 participants performed three blocks at constant volatility ( $h = 0.1$ ) and variable noise ( $R = 0.01$ ,  $R = 0.8$  and  $R = 4$ ), and 84 participants performed three blocks at constant noise ( $R = 1$ ) and variable volatility ( $h = 0.08$ ,  $h = 0.38$  and  $h = 0.8$ ). Trials in which participants input a number outside the 0–5,000 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 participants using the Amazon Mechanical Turk crowdsourcing website. The participants performed the Bernoulli prediction task in the following game setting. Two jars (the two sources) contained

beads of different colours: 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  $h$ . The participants had to guess the colour 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 colour of the next bead, they had to infer and track the jar that was most likely going to produce the next bead. Each participant performed four blocks of 300 trials each, corresponding to the following four conditions: (1)  $h=0.05$  and  $p=0.98$ , (2)  $h=0.05$  and  $p=0.86$ , (3)  $h=0.05$  and  $p=0.57$ , and (4)  $h=0.4$  and  $p=0.86$  (conditions 1, 2 and 3 had fixed volatility and increasing noise; conditions 2 and 4 had fixed noise and increasing volatility). Trials in which participants 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<sup>3,5,64</sup> and large enough to ensure the validity of the  $t$ -tests used in Fig. 7.

**Data analysis.** *Gaussian change-point task.* Values of adaptivity and working-memory load for the participants and the models (Fig. 7b–e) were obtained with the following analyses.

**Integration kernels.** For each participant and noise–volatility condition, we computed an integration kernel (a linear weighting function) for each set of trials having the same lag from a change-point. We thus considered the set of participant responses  $\{R_t\}$  from all trials  $t_q = t_q^{\text{CP}} + \Delta t$ , with  $\Delta t$  a fixed lag,  $t_q^{\text{CP}}$  the  $q$ th change-point trial and  $q$  running from 1 to  $M$  (the number of change-points in a given block occurring at  $t_q^{\text{CP}} > n - \Delta t$ ; see below). From the response set  $\{R_t\}$ , we estimated the participant integration kernel for the lag  $\Delta t$  by finding the weights  $\{K_p, K_0, \dots, K_n\}$  of the multiple linear regression model

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

$K_p$  (the weight given to the prior  $\bar{\mu}$ ) and  $K_0, \dots, K_n$  (the weights given to the  $n+1$  most recent observations) were obtained using the MATLAB lsqmin 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  $0 \leq K_i \leq 1, i = \{p, 0, \dots, n\}$ . 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 participant, block and lag  $\Delta t$ . 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  $t_q^{\text{CP}} > n - \Delta t$ ), yielding underconstrained weights.

**Integration timescales.** For each integration kernel, we computed the normalized cumulative weight of the most recent  $\tau$  observations

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

and the timescale at which this normalized cumulative weight reaches a fixed threshold  $\theta$

$$\tau_{\theta} = \min\{\tau \mid c(\tau) > \theta\} \quad (38)$$

where  $\tau_{\theta}$  represents an integration timescale. For the results in Fig. 7, we used  $\theta=0.8$ , so that  $\tau_{\theta}$  is the timescale that explains 80% of the participant's integration over recently observed data. Conclusions were in general robust against changes in the threshold  $\theta$  (Supplementary Fig. 7). We computed one value of  $\tau_{\theta}$  for each integration kernel—that is, for each participant, block and lag  $\Delta t$ . We estimated the standard error on each  $\tau_{\theta}$  by bootstrapping the regression model (equation (36)). We used 200 bootstrap samples of the form  $\mathbf{b} = \{(R_{t_{i_1}}, \mathbf{x}_{t_{i_1}}), \dots, (R_{t_{i_m}}, \mathbf{x}_{t_{i_m}})\}$ , with  $\mathbf{x}_{t_{i_k}} = (\bar{\mu}, x_{t_{i_k}}, \dots, x_{t_{i_k}-n})$  and  $i_1, \dots, i_m$  a random sample (with replacement) of the integers 1 through  $M$  (see ref. <sup>65</sup> for more details about the bootstrap procedure).

**Adaptivity.** For each participant and each block of trials, we computed adaptivity as the variance of  $\tau_{\theta}$  across all lags  $0 \leq \Delta t \leq n$  (because the integration timescale  $\tau_{\theta}$  cannot be larger than the timescale  $n$  in the linear regression model). This metric of adaptivity quantifies how much the integration timescale 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 and volatility levels, we normalized each participant's adaptivity in any given condition by the maximum adaptivity for the same participant across the three conditions, and we then averaged the normalized values across participants. The few participants (three for the variable-noise conditions and one for the variable-volatility conditions) for whom adaptivity was zero in all three conditions were excluded from the analysis in Fig. 7b–e, because the normalized adaptivity was undefined. Standard errors on the average normalized adaptivity were computed as

s.e.m. =  $\sqrt{\sum_i \left( \frac{\partial(\text{normalized adaptivity})}{\partial \tau_{\theta,i}} \right)^2 (\text{s.e.}(\tau_{\theta,i}))^2}$  with the sum running over all the  $\tau_{\theta}$  obtained for different participants and lags  $\Delta t$ .

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 participants, in place of  $R_t$  in equation (36);  $\tau_{\theta}$  and normalized adaptivity values were then obtained in the same way as the participant's values. For the main bars in Fig. 7b, the most efficient model was defined as the simplest model in our hierarchy with inaccuracy  $\mathcal{I} < 0.1$  (as in Fig. 5b); theoretical predictions were qualitatively conserved across a relatively wide range of tolerance levels (for example, between 0.02 and 0.2; see Fig. 7b, dashed grey lines).

**Working-memory load.** For each participant or most efficient model and each block of trials, we computed working-memory load as the maximum  $\tau_{\theta}$  across all possible lags  $0 \leq \Delta t \leq n$ —that is, the maximum integration timescale (relative to the threshold  $\theta$ ) used for any given condition of noise and volatility. We then normalized this value by the maximum working-memory load across conditions for each participant and averaged the normalized values across participants. Error bars were computed as for adaptivity.

**Model fitting.** The probabilities that each of the models in Fig. 1 generated the data of a randomly selected participant (Fig. 7f,g) were obtained with the following method. We assumed that a given participant in a given condition made inferences  $\mu_t^{(s)}$  using one of the models in 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 participant, model ( $M$ ) and condition, we identified the values of the model parameters  $\theta$  maximizing the log-likelihood of the participants' responses under the model:

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

The parameters are  $\sigma_{\text{noise}}$  for the Evidence, Prior and Bayesian models;  $\sigma_{\text{noise}}$  and one learning rate for the single Delta-Rule, single Sliding-Window and Memoryless models; and  $\sigma_{\text{noise}}$  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 KF and a two-level HGF in the analysis. The parameters for the KF are  $\sigma_{\text{noise}}$  and the process noise (that is, the variance of the diffusion process)<sup>61,62</sup>. The parameters for the HGF are  $\sigma_{\text{noise}}$ , 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<sup>31,32,63</sup>.

**Model log-evidence values.** The maximum log-likelihood values were used to compute the Bayesian information criterion approximation of the log-evidence for each model and participant:

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

with  $k$  the number of parameters.

**Model probabilities.** We aggregated the model log-evidence values across participants using the method developed in ref. <sup>30</sup>, which uses a variational Bayesian approach to estimate the parameters of a Dirichlet posterior distribution over model probabilities. Figure 7f,g 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 participants' data are generated by the same model and is robust to the presence of outliers.

**Confusion analysis.** The confusion matrices in Supplementary Fig. 8 were generated as follows. First, we simulated each of the eight models in Fig. 1 using parameter values (including decision noise) fit to the participants' data. For each model, we ran a number of simulations equal to the number of participants, with each simulation composed of the same number of trials as in the experiment. Second, we fit each of the eight models to the data from each simulation. Third, 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. 7f,g.

**Bernoulli change-point task.** We identified the best-fitting model for each participant in two ways. In Fig. 8, we defined 'best fit' as the model that minimized the sum of squared residuals for a given participant. This definition is justified by the non-parametric nature of the three models considered for this task (Fig. 8a);  $p(\text{model})$  in Fig. 8b,c is the fraction of participants 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  $\sigma_{\text{noise}}$ ) in model inferences.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The experimental data used in Figs. 7 and 8 and Supplementary Figs. 6–10 are available on GitHub under the open-source license GNU-GPL-v3: <https://github.com/gaiat/Modeling-Human-Inference>. The data were collected using psiTurk (v.2.3.0)<sup>66</sup> (<http://psiturk.org/>), <https://github.com/NYUCCL/psiTurk>), jsPsych (v.6.0)<sup>67</sup> (<https://www.jspsych.org/>), <https://github.com/jspsych/jsPsych/>), chart.js (v.2.8.0) (<https://www.chartjs.org/>), <https://github.com/chartjs/Chart.js>) and papaparse.js (v.5.0) (<https://www.papaparse.com/>, <https://github.com/mholt/PapaParse>).

## Code availability

The codes are available on GitHub under the open-source license GNU-GPL-v3: <https://github.com/gaiat/Modeling-Human-Inference>.

Received: 5 May 2020; Accepted: 20 April 2022;

Published online: 30 May 2022

## References

- Rao, R. P. N. Bayesian computation in recurrent neural circuits. *Neural Comput.* **16**, 1–38 (2004).
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
- Fearnhead, P. & Liu, Z. On-line inference for multiple changepoint problems. *J. R. Stat. Soc. B* **69**, 589–605 (2007).
- Shi, L. & Griffiths, T. L. Neural implementation of hierarchical Bayesian inference by importance sampling. *Adv. Neural Inf. Process. Syst.* **22**, 1669–1677 (2009).
- Brown, S. D. & Steyvers, M. Detecting and predicting changes. *Cogn. Psychol.* **58**, 49–67 (2009).
- Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
- Wilson, R., Nassar, M. & Gold, J. A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Comput. Biol.* **9**, e1003150 (2013).
- Legenstein, R. & Maass, W. Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. *PLoS Comput. Biol.* **10**, e1003859 (2014).
- Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).
- Ortega, P. A. & Braun, D. A. Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A* **469**, 20120683 (2013).
- Glaze, C. M., Filipowicz, A. L. S., Kable, J. W., Balasubramanian, V. & Gold, J. I. A bias-variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nat. Hum. Behav.* **2**, 213–224 (2018).
- Adams, R. & MacKay, D. Bayesian online changepoint detection. Preprint at <https://doi.org/10.48550/arXiv.0710.3742> (2007).
- Wilson, R. C., Nassar, M. R. & Gold, J. I. Bayesian online learning of the hazard rate in change-point problems. *Neural Comput.* **22**, 2452–2476 (2010).
- Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* **30**, 12366–12378 (2010).
- Heilbron, M. & Meyniel, F. Confidence resets reveal hierarchical adaptive learning in humans. *PLoS Comput. Biol.* **15**, e1006972 (2019).
- Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
- Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 1998).
- Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **9**, 349–368 (1997).
- Barron, A., Rissanen, J. & Yu, B. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **44**, 2743–2760 (1998).
- Gutenkunst, R. et al. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, e189 (2007).
- Transtrum, M. K. & Qiu, P. Model reduction by manifold boundaries. *Phys. Rev. Lett.* **113**, 098701 (2014).
- Fan, Y., Gold, J. I. & Ding, L. Ongoing, rational calibration of reward-driven perceptual biases. *eLife* **7**, e36018 (2018).
- Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
- Zeng, X., Song, T., Zhang, X. & Pan, L. Performing four basic arithmetic operations with spiking neural P systems. *IEEE Trans. Nanobiosci.* **11**, 366–374 (2012).
- Shenhav, A. et al. Toward a rational and mechanistic account of mental effort. *Annu. Rev. Neurosci.* **40**, 99–124 (2017).
- Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. One and done? Optimal decisions from very few samples. *Cogn. Sci.* **38**, 599–637 (2014).
- Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* **2**, 230–247 (2010).
- Gold, J. I. & Shadlen, M. N. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299–308 (2002).
- Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S. C. & Heekeren, H. R. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc. Natl Acad. Sci. USA* **106**, 17951–17956 (2009).
- Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46**, 1004–1017 (2009).
- Mathys, C. & Weber, L. Hierarchical Gaussian filtering of sufficient statistic time series for active inference. In *International Workshop on Active Inference* (eds Verbeelen, T. et al.) 52–58 (Springer, 2020).
- Mathys, C. D. et al. Uncertainty in perception and the hierarchical Gaussian filter. *Front. Hum. Neurosci.* **8**, 825 (2014).
- Lee, S., Gold, J. I. & Kable, J. W. The human as delta-rule learner. *Decision* **7**, 55–66 (2020).
- Glaze, C. M., Kable, J. W. & Gold, J. I. Normative evidence accumulation in unpredictable environments. *eLife* **4**, e08825 (2015).
- Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H. & Rushworth, M. F. S. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron* **65**, 927–939 (2010).
- Sul, J. H., Jo, S., Lee, D. & Jung, M. W. Role of rodent secondary motor cortex in value-based action selection. *Nat. Neurosci.* **14**, 1202–1210 (2011).
- Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2012).
- Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. Preprint at <https://doi.org/10.48550/arXiv.physics/0004057> (2000).
- Canziani, A., Paszke, A. & Culurciello, E. An analysis of deep neural network models for practical applications. Preprint at <https://doi.org/10.48550/arXiv.1605.07678> (2016).
- Cheeseman, P. C., Kanefsky, B. & Taylor, W. M. Where the really hard problems are. *IJCAI (US)* **91**, 331–340 (1991).
- 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* **306**, 381–394 (2002).
- Mitchell, D., Selman, B. & Levesque, H. Hard and easy distributions of SAT problems. *AAAI* **92**, 459–465 (1992).
- Zdeborová, L. Statistical physics of hard optimization problems. *Acta Physica Slovaca Rev. Tutor.* **59**, 169–303 (2009).
- Wilson, R. C., Nassar, M. R., Tavoni, G. & Gold, J. I. Correction: a mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Comput. Biol.* **14**, e1006210 (2018).
- Gerstner, W., Kistler, W. M., Naud, R. & Paninski, L. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge Univ. Press, 2014).
- Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
- Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
- Gläscher, J. & Büchel, C. Formal learning theory dissociates brain regions with different temporal integration. *Neuron* **47**, 295–306 (2005).
- Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
- Bernacchia, A., Seo, H., Lee, D. & Wang, X. J. A reservoir of time constants for memory traces in cortical neurons. *Nat. Neurosci.* **14**, 366–372 (2011).
- Scott, B. B. et al. Fronto-parietal cortical circuits encode accumulated evidence with a diversity of timescales. *Neuron* **95**, 385–398 (2017).
- Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).
- Meder, D. et al. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nat. Commun.* **8**, 1942 (2017).
- Joshi, S. & Gold, J. I. Pupil size as a window on neural substrates of cognition. *Trends Cogn. Sci.* **24**, 466–480 (2020).
- Arnsten, A. F. T., Wang, M. J. & Paspalas, C. D. Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. *Neuron* **76**, 223–239 (2012).

56. Yerkes, R. M. & Dodson, J. D. The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* **18**, 459–482 (1908).
57. Cools, R. & D'Esposito, M. Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biol. Psychiatry* **69**, e113–e125 (2011).
58. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus–norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
59. Griffiths, T. L., Vul, E. & Sanborn, A. N. Bridging levels of analysis for probabilistic models of cognition. *Curr. Dir. Psychol. Sci.* **21**, 263–268 (2012).
60. Fusi, S., Asaad, W. F., Miller, E. K. & Wang, X. J. A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. *Neuron* **54**, 319–333 (2007).
61. Kalman, R. E. & Bucy, R. S. New results in linear filtering and prediction theory. *J. Basic Eng.* **83**, 95–108 (1961).
62. Welch, G. & Bishop, G. *An Introduction to the Kalman Filter* <https://perso.crans.org/club-krobot/doc/kalman.pdf> (1997).
63. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5**, 39 (2011).
64. Ossmy, O. et al. The timescale of perceptual evidence integration can be adapted to the environment. *Curr. Biol.* **23**, 981–986 (2013).
65. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC Press, 1994).
66. McDonnell, J. V. et al. psiTurk v.1.02 (New York University, 2012).
67. De Leeuw, J. R. jspsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behav. Res. Methods* **47**, 1–12 (2015).

## Acknowledgements

We thank A. Filipowicz for sharing the codes to run the psychophysical experiment for the Bernoulli task. We also thank K. Krishnamurthy and E. Piasini for interesting

discussions, and A. Cavagna and A. Ingrosso for pointing out a possible connection between one of our results and spin-glass systems. G.T. was supported by the Swartz Foundation (award no. 575556) and the Computational Neuroscience Initiative of the University of Pennsylvania, and is currently supported by Washington University in St. Louis. V.B. and J.I.G. are supported in part by NIH BRAIN Initiative grant no. R01EB026945. J.I.G. is also supported by grant nos 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.

## Author contributions

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

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41562-022-01357-z>.

**Correspondence and requests for materials** should be addressed to Gaia Tavoni.

**Peer review information** *Nature Human Behaviour* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

For data collection, we used:

Psiturk (v2.3.0):

McDonnell, J.V., Martin, J.B., Markant, D.B., Coenen, A., Rich, A.S., and Gureckis, T.M. (2012). psiTurk (Version 1.02) [Software]. New York, NY: New York University.

and the following javascript libraries:

jsPsych (v6.0):

de Leeuw, J.R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods, 47(1), 1-12. doi:10.3758/s13428-014-0458-y

chart.js (v2.8.0)

papaparse.js (v5.0)

Data analysis

Codes for numerical analyses are written in Matlab R2020b and are available on GitHub under the open-source license GNU-GPL-v3: <https://github.com/gaiat/Modeling-Human-Inference>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The experimental data used in Figs.7,8 and Supplementary Figs.6-10 are available on GitHub under the open-source license GNU-GPL-v3:

<https://github.com/gaiat/Modeling-Human-Inference>

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We used the Amazon Mechanical Turk crowdsourcing website to recruit 169 human subjects to perform a quantitative estimation task designed to test key features of our theoretical findings. The task was presented as a card game. In brief, subjects were shown on each trial a card number drawn from a card deck; their task was to guess which deck the card was drawn from.
Research sample	169 human subjects recruited via the Amazon Mechanical Turk crowdsourcing website. No prerequisites were necessary to participate in the task, thus the subjects who enrolled in the experiment presumably represented a diverse population. We did not record demographic information because it was not relevant to the scope of the study.
Sampling strategy	We collected data from a pilot set of 20 random subjects, which confirmed that the task was working correctly and that more subjects were needed to identify trends in the presence of individual variability. We then collected data from another 100 random subjects, re-analyzed the data, and saw trends beginning to emerge amidst the individual variability. From these findings we estimated that an additional 50 subjects were needed to identify those trends in a statistically reliable manner. We therefore used that target to collect a final data set, then conducted our analyses on the complete data set.
Data collection	In the Gaussian change-point task, each subject performed 3 blocks of 400 trials in total (40 training trials followed by 360 test trials). In the Bernoulli change-point task, each subject performed 4 blocks of 300 trials. The researcher was blind to experimental conditions during data collection. Subjects performed the task on their own computers. Their responses (keyboard input) were stored and compiled into a data base. No one was present besides the participants and the researcher.
Timing	Data on the Gaussian change-point task were collected in April, May 2019. Data on the Bernoulli task were collected in November 2020.
Data exclusions	We only excluded 4 subjects because they had zero adaptivity in all the three tested conditions, thus normalized adaptivity (one of the key quantities under examination) was undefined for these subjects. We also excluded few trials in which subjects were overtly not attending the task, as indicated by the following (predetermined) criterion: subject input a number outside the 0-5000 range prescribed in the task instructions, or responded in less than 100 ms.
Non-participation	Participants voluntarily enrolled in the experiment through the Amazon Mechanical Turk website and data from participants who dropped out before completing the task were not recorded. We do not have information on participants who dropped out.
Randomization	Subjects were allocated randomly and with equal probability into one of two groups: those in the first/second group were tested in conditions of variable noise/volatility, respectively.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics age, gender, genotypic or other information about human participants were not relevant to the scope of this study.

Recruitment

Participants were recruited through the Amazon Mechanical Turk crowdsourcing website. We are not aware of any potential self-selection bias or other biases that may be present.

Ethics oversight

The study protocol 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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.