ARTICLES
https://doi.org/10.1038/s41562-022-01352-4

nature human behaviour

Efficient coding of numbers explains decision bias and noise

Arthur Prat-Carrabin ✉ and Michael Woodford

Humans differentially weight different stimuli in averaging tasks, which has been interpreted as reflecting encoding bias. We examine the alternative hypothesis that stimuli are encoded with noise and then optimally decoded. Under a model of efficient coding, the amount of noise should vary across stimuli and depend on statistics of the stimuli. We investigate these predictions through a task in which the participants are asked to compare the averages of two series of numbers, each sampled from a prior distribution that varies across blocks of trials. The participants encode numbers with a bias and a noise that both depend on the number. Infrequently occurring numbers are encoded with more noise. We show how an efficient-coding, Bayesian-decoding model accounts for these patterns and best captures the participants' behaviour. Finally, our results suggest that Wei and Stocker's "law of human perception", which relates the bias and variability of sensory estimates, also applies to number cognition.

Human choices rely on representations, by the brain, of the variables relevant to the decision. How external variables are encoded into internal representations and how these are then decoded for decisions are central questions in cognitive neuroscience[1–6]. In many decision problems, several pieces of information are pertinent; how humans combine these in their decision-making process is a long-standing debate in economics and psychology[7–11]. Recently, a series of experimental studies have focused on averaging tasks, in which participants are presented with several stimuli (sometimes numbers, but sometimes visual stimuli characterized by their length, orientation, shape or colour) and asked to make a decision about the average magnitude of the presented stimuli[12–15]. Although the contribution of each stimulus to the average should, in theory, be proportional to its true magnitude, the weights attributed to stimuli by human participants appear to be nonlinear functions of their magnitudes. Participants asked to compare the averages of two series of digits, for instance, overweight larger digits when making a decision[12]. To account for this seemingly suboptimal behaviour, refs. [12,13] show that if the comparison of the average encoded values involves noise, then a nonlinear transformation of the presented stimuli can partially compensate for the performance loss induced by the noise. The nonlinear distortion of stimuli appears, under this proposal, as a consequence of an optimal encoding strategy, given unavoidable noise at decision time.

Here we investigate the alternative hypothesis that it is the encoding of a stimulus that is unavoidably noisy, while the decoded value is an optimal estimate of the magnitude that has been encoded by a noisy representation[16,17]. The resulting noisy estimate will generally be biased, to a degree that will vary as a function of the stimulus[18–21]. The average estimate will thus be a nonlinear function of the true stimulus magnitude, though for a different reason than in the model of ref. [12] mentioned above. This nonlinear function will depend on how the encoding noise varies over the stimulus space. Efficient coding theories suggest that the degree of noise with which a stimulus is encoded should be a decreasing function of its probability under the prior distribution of stimuli; in other words, less likely stimuli should be encoded with more noise[5,22–25]. This implies that the way that both estimation bias and estimation noise vary over the stimulus space should depend on the prior distribution over that space; we test for such dependence in a behavioural experiment.

In a related model that also combines efficient coding with Bayesian decoding, Wei and Stocker derive a relation between estimation noise and estimation bias, a "law of human perception" supported by evidence from numerous sensory domains[26]. In an experiment involving number comparisons, we find support for this law as well, suggesting that it applies to the semantic representation of numerosity carried by Arabic numerals and not only to the perception of physical stimuli. We also test other implications of both efficient coding and optimal decoding.

These theories highlight the potential role of the prior in human decision-making. We thus design a task in which participants are asked to compare the averages of two series of numbers, and we manipulate the prior distribution of the presented numbers across different blocks of trials: it is sometimes uniform but sometimes skewed towards smaller or larger numbers. In each of these three conditions, the weights of numbers in the participants' decisions appear to vary nonlinearly over the range of presented numbers. We compare the behaviour of our participants to the predictions of a family of purely descriptive statistical models—that is, models of the distribution of the estimate, conditional on the number presented, with no interpretation in terms of encoding and decoding—and find that the patterns in their responses are best captured by a model in which a nonlinear transformation of a presented number is observed with a degree of noise that itself depends on the number.

We then turn to a two-stage encoding–decoding model—that is, a model that decomposes the estimation process into an encoding stage and a decoding stage, each of which is further assumed to be optimized for the distribution of stimuli encountered in the task. We show how a model of efficient coding combined with Bayesian decoding accounts parsimoniously for both the nonlinear transformation and the noise, and best captures the participants' behaviour. We find that the encoding noise varies depending on the prior and that this allows our participants to increase their expected reward in the task.

Department of Economics, Columbia University, New York, NY, USA. ✉e-mail: arthur.p@columbia.edu

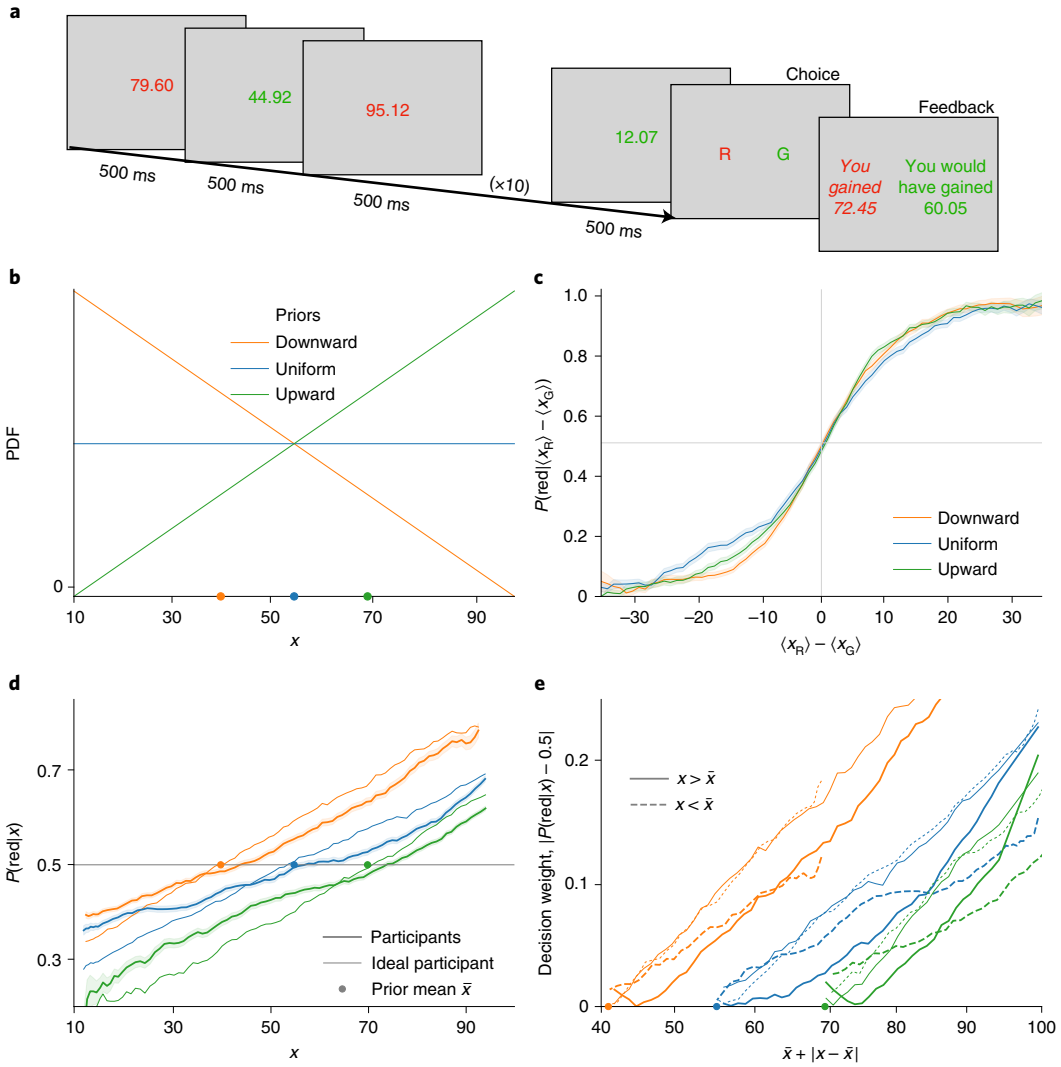

Fig. 1 | Illustration of the task and participants' choice behaviour. a, In each trial, ten numbers are sequentially presented, alternating in colour between red and green. The participant then chooses a colour, without a time constraint, by pressing the corresponding key. Feedback is given, which reports the two averages and emphasizes the chosen one. **b**, In each trial, both red and green numbers are drawn from the same prior distribution over the range [10.00, 99.99]. In different blocks of consecutive trials, the prior is Downward (a triangular distribution with the peak at 10.00, orange line), Uniform (blue line) or Upward (a triangular distribution with the peak at 99.99, green line). PDF, probability density function. **c**, Choice probability: the fraction of trials in which participants choose the 'red' average, as a function of the true difference in the averages of the two series of numbers. **d**, Probability $P(\text{red}|x)$ of choosing 'red' conditional on a red number x being presented, as a function of x, for the participants (thick lines) and the ideal participant (thin lines). **e**, Decision weight, defined as $|P(\text{red}|x) - 0.5|$, for the participants (thick lines) and the ideal participant (thin lines), as a function of the sum of the prior mean, $\bar{x}$, and the absolute difference between the number and the prior mean, $|x - \bar{x}|$. The participants' decision weights for numbers below the mean (dashed lines) and above the mean (solid lines) are appreciably different, whereas for the ideal participant there is only a small difference due to sampling error. In **c**, **d** and **e**, each point of the curves is obtained by taking the average, combining all participants' data, of the quantity of interest (ordinate) over a sliding window of length 10 of the quantity on the abscissa, incremented by steps of length 1. In **c** and **d**, the shaded areas represent the standard error of the mean.

Results

We first present our average-comparison experiment. In each trial of a computer-based task, ten numbers, each within the range [10.00, 99.99] and alternating in colour between red and green, are presented to the participant in rapid succession. The participant is then asked to choose whether the five red numbers or the five green numbers had the higher average (Fig. 1a). In different blocks of trials, the numbers are sampled from different prior distributions: the 'Uniform' distribution (under which all numbers in the [10.00, 99.99] range are equally likely), an 'Upward' distribution (under which higher numbers are more likely) and a 'Downward' distribution (under which lower numbers are more likely) (Fig. 1b). Within a block of trials, both red and green numbers are sampled from the same distribution. Additional details on the task are provided in the Methods. In our analyses below, we combine the responses of all the participants; the examination of individual data further supports our results (Supplementary Information).

Although the presented information would allow, in principle, an unambiguous identification of the colour of the numbers that have the higher average, participants make errors in their responses. These errors are not independent of the presented numbers: the proportion of trials in which they choose 'red' (the 'choice probability') is an increasing, sigmoid-like function of the difference between the average of the red numbers and that of the green

numbers, with about 50% red responses when this difference is close to zero. In other words, the participants make fewer errors when there is a marked difference between the two averages (Fig. 1c).

To isolate the influence of a single number on the decision, we can compute the probability of choosing 'red' conditional on a given (red) number x being presented, $P(\text{red}|x)$, and the absolute difference between this probability and 0.5: $|P(\text{red}|x)-0.5|$. This quantity, which we call the 'decision weight' (following ref. [12]), is a measure of the average impact of a given number on the decision. For an ideal participant who makes no errors, the decision weight is zero for x equal to the prior mean $\bar{x}$ and increases approximately linearly with the absolute distance $|x-\bar{x}|$ from the prior mean (Fig. 1d,e, thin lines). In our participants' data, instead, the value of x for which the decision weight is zero is somewhat larger than $\bar{x}$ for each prior, and the decision weight increases more steeply with increases in x above this value than it does with decreases in x below that value (Fig. 1d,e, thick lines).

We wish to explore models of noisy comparison that can account for these regularities. We begin by fitting models to our data that do not separate the processes of encoding and decoding, and instead simply posit that a given number x results in a noisy estimate $\hat{x}$ drawn from a conditional distribution $p(\hat{x}|x)$. We assume that the ten numbers presented are estimated with this procedure (with independent draws) and that the colour 'red' is chosen if the average of the estimates of the red numbers is greater than that of the green numbers—that is, if $\frac{1}{5}\sum_{i=1}^{5}\hat{x}_i^{\mathrm{R}} > \frac{1}{5}\sum_{i=1}^{5}\hat{x}_i^{\mathrm{G}}$, where $\hat{x}_i^{\mathrm{R}}$ and $\hat{x}_i^{\mathrm{G}}$ are the estimates for the red and green numbers. This kind of characterization is equally consistent with the theory of ref. [12] (in which $\hat{x}$ is a deterministic transformation of x plus a random noise term added at the time of the comparison process) and a model in which x is encoded with noise and $\hat{x}$ is then an estimate of the value of x based on the noisy encoded value.

We consider several assumptions about the form of the conditional distribution of estimates, $p(\hat{x}|x)$. The simplest kind of model consistent with a sigmoid choice probability (as in Fig. 1c) would be one in which the estimate is normally distributed around the true value of the number, with a constant standard deviation, s—that is, $\hat{x}|x \sim \mathrm{N}(x, s^2)$. In other words, numbers are perceived with constant Gaussian noise, but the estimates are unbiased. The probability of choosing the red colour, conditioned on the ten presented numbers, would then be

$$cP\left(\sum_{i=1}^{5}\hat{x}_i^{\mathrm{R}} > \sum_{i=1}^{5}\hat{x}_i^{\mathrm{G}}|x_{1:5}^{\mathrm{R}}, x_{1:5}^{\mathrm{G}}\right) = \Phi\left(\frac{\sum_{i=1}^{5}x_i^{\mathrm{R}}-\sum_{i=1}^{5}x_i^{\mathrm{G}}}{\sqrt{10s^2}}\right), \quad (1)$$

where Φ is the cumulative distribution function of the standard normal distribution. The choice probability in this model is thus a sigmoid function of the difference between the averages of red and green numbers, similarly to the choice probability of the participants (Fig. 2a). However, the decision weight of a number, in this model, is an approximately linear function of the absolute difference between the number and the prior mean, and two numbers equally distant from the mean have the same weight (Fig. 2c, top left). This simple model thus does not reproduce the participants' unequal weighting of numbers in decisions (Fig. 1e).

In the model just presented, the estimates are unbiased ($\mathbb{E}\hat{x} = x$), and all numbers are estimated with the same amount of noise ($\mathrm{Var}\hat{x} = s^2$). We also consider models that are more flexible in either or both of these respects. For example, we can allow for bias by assuming that the estimate, $\hat{x}$, of a presented number, x, is normally distributed around a nonlinear transformation, $m(x)$, of the number: $\hat{x}|x \sim \mathrm{N}(m(x), s^2)$. Such a model is equivalent to the one assumed in refs. [12,13], in which noise is added to the sum of nonlinearly transformed values of the stimuli. Alternatively, we might assume that the variability of estimates is not constant over the stimulus space, as is found to be true in many stimulus domains[20]. The simplest case of this kind would be one in which $\hat{x}|x \sim \mathrm{N}(x, s^2(x))$, where now the estimation noise, $s(x)$, varies with the number. Our most general model combines the features of these last two, positing that a transformation of the number is observed with varying noise: $\hat{x}|x \sim \mathrm{N}(m(x), s^2(x))$. In this model, the probability of choosing the red colour, conditioned on the ten presented numbers, will be

$$cP\left(\sum_{i=1}^{5}\hat{x}_i^{\mathrm{R}} > \sum_{i=1}^{5}\hat{x}_i^{\mathrm{G}}|x_{1:5}^{\mathrm{R}}, x_{1:5}^{\mathrm{G}}\right) = \Phi\left(\frac{\sum_{i=1}^{5}m(x_i^{\mathrm{R}})-\sum_{i=1}^{5}m(x_i^{\mathrm{G}})}{\sqrt{\sum_{i=1}^{5}s^2(x_i^{\mathrm{R}})+\sum_{i=1}^{5}s^2(x_i^{\mathrm{G}})}}\right). \quad (2)$$

The first three models are special cases of the fourth one, with an absence of bias (that is, an identity transformation $m(x)=x$) or a constant noise ($s(x)=s$) or both (equation (1)).

We fit these models to the behavioural data by maximizing their likelihoods. We flexibly specify the functions $m(x)$ and $s(x)$ by allowing them to be arbitrary low-order polynomials, with the order chosen to minimize the Bayesian information criterion (BIC). Running simulations of the fitted models, we find that each of them predicts that choice frequency should be a similar sigmoid function of the difference between the two averages. The models differ, however, in the implied graphs for the decision weights. The unbiased model with varying noise ($\mathrm{N}(x, s^2(x))$) still results in approximately linear decision weights, with equal weights for numbers at equal distances from the mean (Fig. 2c, top right). In contrast, the two models that include a transformation $m(x)$ of the presented numbers yield nonlinear decision weights resembling those of the participants: numbers below the mean and close to it have higher weights than numbers above the mean and equally distant from it, while the opposite occurs for numbers further from the mean (Fig. 2c, bottom panels; more details on how the transformation $m(x)$ affects the decision weights can be found in the Methods). In summary, the two models that feature a transformation of the number, whether with a constant or variable noise, better capture the patterns observed in the behavioural data.

To quantitatively compare the degrees of explanatory success of the models, we compute the BIC for the best-fitting model in each class. We can assume either that the parameters of the models are identical under the three priors ('homogeneous parameters') or, conversely, that the parameters depend on the prior ('prior-specific parameters'). The latter results in a larger number of parameters, which is penalized by the BIC. In spite of this penalty, in all four models the BIC is lower with prior-specific parameters than with homogeneous parameters, suggesting that the participants' decision-making process is adapted to the prior distribution (Fig. 2b). The two best-fitting one-stage models include a transformation $m(x)$ of the presented number, in accordance with our observations about the decision weights, and the best-fitting model also features a variable noise $s(x)$. We note that the BIC in this case is significantly lower than with a transformation but maintaining constant noise ($\Delta\mathrm{BIC}=81$), in spite of the additional parameters of the noise function $s(x)$. In the Supplementary Information, we discuss the finer features of the participants' behaviour that are better captured by allowing for variable noise, showing that a nonlinear transformation alone (as in refs. [12,13]) does not suffice to account for our data. In addition, to investigate the robustness of our results, we conduct a random-effect analysis that takes into account the participants' heterogeneity. We follow the Bayesian model selection procedure described in ref. [27] and use cross-validation to estimate the models' likelihoods. This analysis, detailed in the Supplementary Information, supports the results presented here.

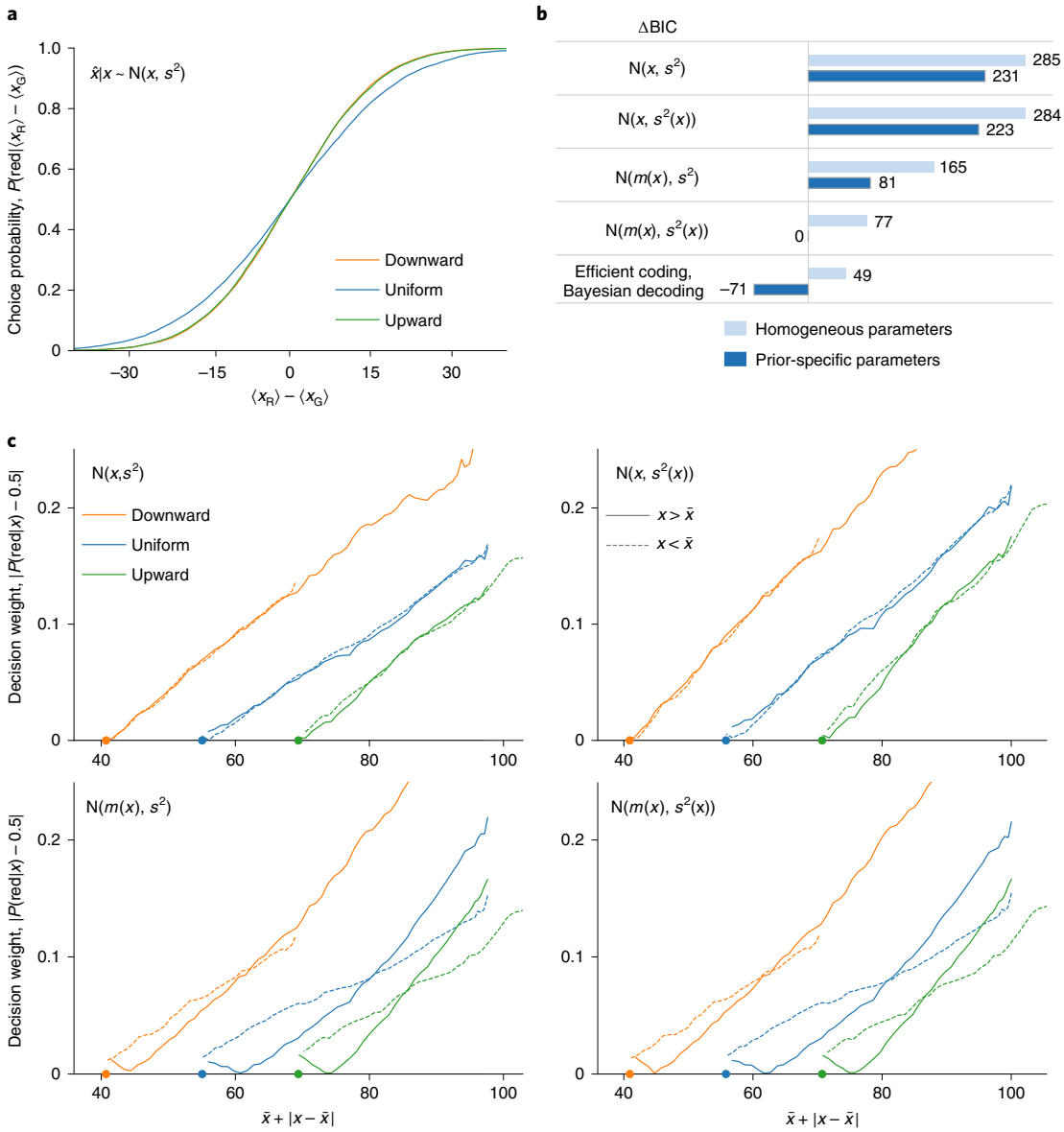

Fig. 2 | Models with both nonlinear transformation and varying noise best capture participants' behaviour. a, Choice probability under the unbiased, constant-noise model ($N(x,s^2)$) as a function of the difference in the averages of the presented numbers, for the three prior conditions. **b**, Model comparison statistics for the four one-stage models of noisy estimation and for the efficient-coding, Bayesian-decoding model, both with homogeneous parameters and with prior-specific parameters. For each model class, the difference in BIC from the least restrictive model class (general transformation, variable noise, prior-specific) is reported. The lower BIC for the efficient-coding, Bayesian-decoding model (a special case of the general one-stage model) indicates that it captures the participants' behaviour more parsimoniously. **c**, Decision weights, $|P(\text{red}|x) - 0.5|$, for the unbiased, constant-noise model (top left); the unbiased, varying-noise model (top right); the model with transformation of the number and constant noise (bottom left); and the model with transformation of the number and varying noise (bottom right), for numbers below the mean (dashed lines) and above the mean (solid lines). Only the two models with transformation of the number reproduce the behaviour of the participants shown in Fig. 1e. All models are fit to the participants' combined data.

The shape of the best-fitting noise function $s(x)$ for each prior is U-shaped: minimal at a value close to the mean of the prior (but slightly below it) and increasing as a function of the distance from this minimum (Fig. 3a). In the Downward-prior condition, larger numbers (which have a low probability) are perceived with a much larger amount of noise than smaller numbers (which have a high probability). The noise function in the Upward condition approximately mirrors that in the Downward condition, with small numbers perceived with more noise than large numbers. In other words, the least likely numbers are perceived with the greatest randomness, suggesting that the process by which the participants estimate numbers is adapted to the prior distribution from which they are drawn.

The finding that the variance of $\hat{x}$ differs for different numbers x is a natural one if we suppose that each estimate $\hat{x}$ results from an encoding–decoding process, whereby each individual number is encoded, with noise, into an internal representation r, and the estimate $\hat{x}(r)$ represents an inference about the likely value of x, based on its noisy internal representation r. In general, an optimal rule of inference will make the estimate $\hat{x}(r)$ a nonlinear function of r, so that the variance of $\hat{x}$ should depend on x even if we suppose that the variance of r does not. At the same time, an encoding–decoding model of this kind, in which decoding is assumed to be optimal given the nature of the noisy encoding, will imply that the function $m(x)$, indicating the bias in the average decoded value, will not be

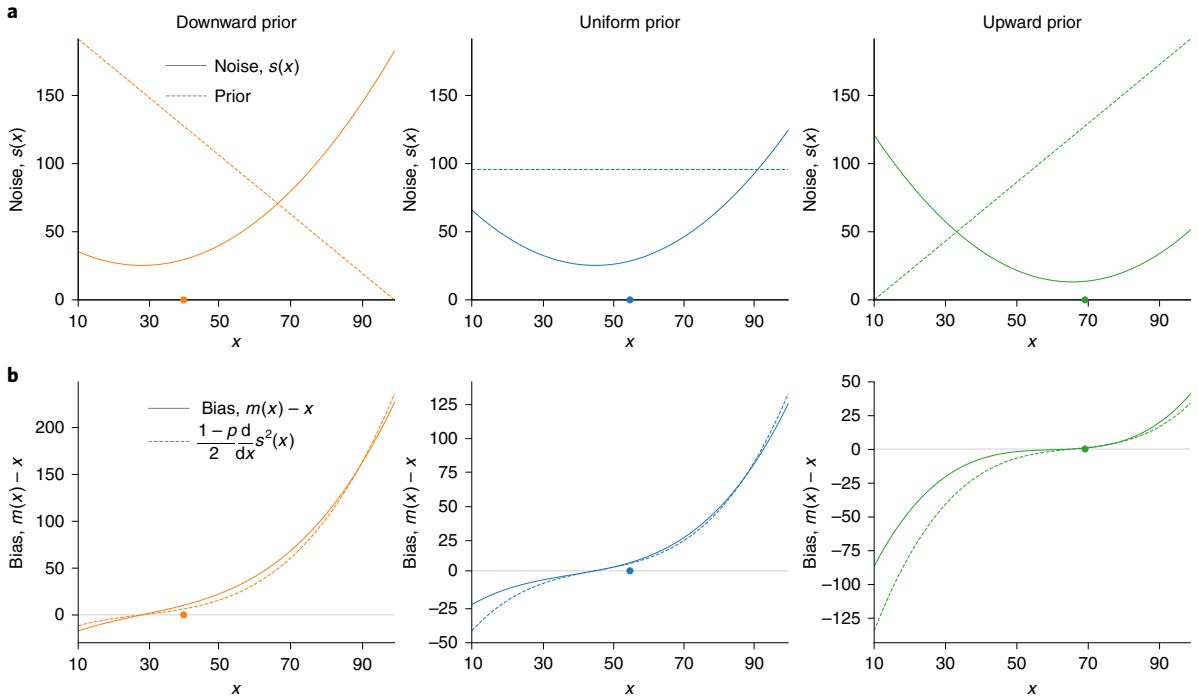

Fig. 3 | Best-fitting noise and bias: participants encode less frequent numbers with greater noise. a, The best-fitting noise function $s(x)$ in the $N(m(x), s^2(x))$ model (solid lines) and the prior distribution (dashed lines) in the Downward (left), Uniform (middle) and Upward (right) conditions. The scale refers to the noise (the scale of the prior PDF is not shown). **b**, The bias (solid lines) and the derivative of the noise variance (multiplied by $\frac{1-p}{2}$, dashed lines) for the best-fitting $N(m(x), s^2(x))$ model, as a function of x, in the three prior conditions. The efficient-coding, Bayesian-decoding model requires these two functions to be the same (equation (4)). The model was fit to the participants' combined data. In **b**, the ranges of the vertical axes are adapted to the values taken by the function.

independent of $s(x)$. This class of models thus represents a special case of the general specification $N(m(x), s^2(x))$, though a different restricted class than any of those considered above. In short, while we have hitherto investigated several assumptions about the form of the conditional distribution of estimates, $p(\hat{x}|x)$, we now derive this distribution from a more parsimonious model of encoding–decoding and examine its ability to capture the behavioural data.

The model that we consider, based on a proposal of Morais and Pillow[28] that we further discuss in ref. [29], combines a noisy but efficient coding of the presented number with a Bayesian-mean decoding of the noisy internal representation. More precisely, suppose that a presented number, x, elicits in the brain a series of n signals, $\mathbf{r} = (r_1, \ldots, r_n)$, each drawn independently from a distribution conditioned on the presented number, $p(r_i|x)$. The Fisher information, $I(x)$, of the vector $\mathbf{r}$ is n times that of one individual signal, r_i. Suppose that this encoding is efficient, in the sense that $I(x)$ solves the optimum problem

$$\min \int \tilde{\pi}(x) I(x)^{-p/2} \mathrm{d}x \qquad \text{s.t.} \int \sqrt{I(x)} \mathrm{d}x \leq K, \tag{3}$$

where $p > 0$ and where $\tilde{\pi}(x)$ is a subjective prior about the distribution of numbers, which may differ from the prior actually used in the experiment, $\pi(x)$. The constraint in equation (3) is the same as in ref. [30], though we consider a more general objective. The rationale for minimizing the objective in equation (3) is that it provides a lower bound on the mean L_p error of any point estimator of x[28]. In the limiting case $p \to 0$, minimizing this quantity is equivalent to maximizing an approximation of the mutual information between the number and its internal representation, which is the objective assumed by Wei and Stocker[30]. An objective with $p > 0$ is more plausible, however, in our setting, because in our task, larger estimation errors reduce a participant's reward to a greater extent than smaller errors do. The optimal Fisher information, derived from equation (3) through variational calculus[28,29], is proportional to a power of the subjective prior: $I(x) \propto \tilde{\pi}(x)^{\frac{2}{p+1}}$ (Methods).

As for the decoding, we assume that the estimate of the participants, $\hat{x}$, is the mean of the Bayesian posterior over the numbers, x, given the noisy signal, $\mathbf{r}$. With large n, this decoding rule, combined with the efficient coding scheme just presented, implies asymptotically a relation between the bias of the estimate and the derivative of the inverse Fisher information[29] (Methods), as

$$\text{Bias} \equiv \mathbb{E}(\hat{x} - x) \simeq \frac{1-p}{2} \frac{\mathrm{d}}{\mathrm{d}x} \left(\frac{1}{I(x)} \right). \tag{4}$$

The estimate variance, in addition, is approximately $1/I(x)$. We thus approximate the distribution of estimates as

$$\hat{x}|x \sim \mathrm{N}\left(x + \frac{1-p}{2} \frac{\mathrm{d}}{\mathrm{d}x} \left(\frac{1}{I(x)} \right), \frac{1}{I(x)} \right). \tag{5}$$

(These asymptotic approximations of the bias and of the variance are accurate to the order $1/n$; Methods.) As in the one-stage models considered above, the estimate is normally distributed around a transformation of the number. This is a special case of the general model considered above, $(N(m(x), s^2(x)))$, with the added constraint that both the nonlinear transformation $m(x)$ and the noise $s(x)$ are here determined by a single function, the Fisher information $I(x)$. In particular, the efficient-coding, Bayesian-decoding model predicts that the bias, $m(x) - x$, is proportional to the derivative of the noise variance, $s^2(x) = 1/I(x)$ (equation (4)).

To examine whether our data are consistent with this additional restriction, we look at the two sides of equation (4) implied by our

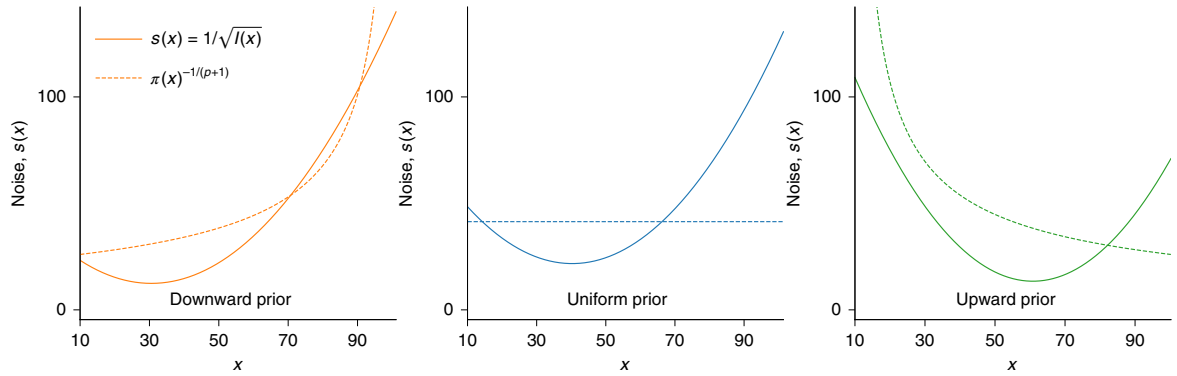

Fig. 4 | Efficient-coding, Bayesian-decoding model: the Fisher information, fitted to the participants' data, is adapted to the prior. The noise function, $s(x)$, equal to the inverse square root of the Fisher information, fitted to the participants' data (solid lines); and the prior $\pi(x)$ raised to the power $-1/(p+1)$, with the best-fitting value $p=0.7$ (dashed lines), in the Downward (left), Uniform (middle) and Upward (right) conditions. The ordinate scale refers to the noise function (the scale for the prior is not shown). The model is fit to the participants' combined data.

best-fitting estimates of $m(x)$ and $s(x)$ when this restriction is not imposed in the estimation. (Here we note that $m(x)$ can be identified from choice data only up to an arbitrary affine transformation; as explained in the Methods, we choose this transformation to make $m(x)-x$ more similar to the right side of the equation.) As noted above, the noise $s(x)$ is U-shaped; thus, the derivative of the noise variance, $\frac{\mathrm{d}}{\mathrm{d}x}s^2(x)$, increases as a function of x; it vanishes at a value close to the prior mean, is negative below this value and is positive above it (Fig. 3b, dashed lines). With an appropriate choice of normalization for $m(x)$, the bias vanishes at the same value, and for all three priors, the bias is also negative below this value and positive above it, and increases as a function of x (Fig. 3b, solid lines). In other words, numbers below the prior mean are underestimated, while numbers above the prior mean are overestimated. We thus find that the two functions, fitted to the data, are qualitatively consistent with the predicted relation (equation (4)).

We now directly estimate the model of estimation bias implied by equation (5), finding the Fisher information function $I(x)$ and the parameter p that minimize the BIC. (Here also, we allow the noise function to be a low-order polynomial; in the Methods, we consider a different functional form, closer to that used in ref. [12], with similar results.) We assume that, regardless of the distribution of numbers, the participants optimize the same efficient coding criterion, determined by p; thus, even when fitting with prior-specific parameters, we maintain the parameter p constant across conditions. As with our one-stage models, prior-specific parameters yield a lower BIC than homogeneous ones. Furthermore, the BIC is lower than that of the model in which the functions $m(x)$ and $s(x)$ are unrestricted (Fig. 2b, $\Delta\mathrm{BIC}=71$). We conclude that the model implied by equation (5) more parsimoniously captures the behaviour of the participants.

In this model, a single function, the Fisher information, $I(x)$, or equivalently the noise function, $s(x)=1/\sqrt{I(x)}$, together with the value of p completely determines the statistics of responses. Moreover, the assumption that the encoding is optimized, in our model (equation (3)), results in the prediction that the noise function should be proportional to the inverse of the subjective prior, $1/\tilde{\pi}(x)$, raised to the power $1/(p+1)$. As we do not have access to the subjective prior, we show (in Fig. 4) the noise functions alongside the correct priors raised to the power $-1/(p+1)$, with p equal to its best-fitting value, 0.7. The best-fitting noise function differs under the three priors. In the Downward condition, the noise is an increasing function of the number over most of the range of presented numbers (Fig. 4, left). In the Upward condition, conversely, the fitted noise function decreases as the number increases (except at large numbers; Fig. 4, right). Hence, in both the Downward and Upward conditions, the noise function is higher for numbers that are less likely under the prior. In the Uniform condition, the noise reaches a minimum around 40, and large numbers are observed with somewhat more noise than small numbers (Fig. 4, middle).

The estimation bias is proportional to the derivative of the squared noise function (equation (4)): it vanishes where the noise reaches a minimum; for smaller x, it is negative, while for higher x, it is positive. Moreover, the efficient-coding, Bayesian-decoding model reproduces the unequal weighting of numbers in the participants' decisions (Fig. 5c,d) and the sigmoid shape of the choice probability curve (Fig. 5e).

Our fitted model implies substantial estimation biases, in particular for the less probable numbers (Fig. 5b). In our model, these biases can exist only to the extent that noise is both substantial in magnitude and sufficiently variable over the range of numbers used; however, the degree of noise required is consistent with the variability of our participants' responses. Moreover, it is not implausible that our participants should encode individual numbers with considerable imprecision, given that they must process ten numbers, each presented for only 500 ms.

The theoretical predictions shown in Fig. 5 are based on a small-noise approximation, which need not be accurate when noise is substantial. Numerical simulations indicate that the bias implied by the small-noise approximation is greatest in the case of extreme values of x that occur with low prior probability, and it thus seems likely that the size of the predicted bias for extreme values of x is exaggerated in Fig. 5b. However, because these values do not often occur in our sample, we do not believe that our conclusions regarding the degree of fit of the Bayesian decoding model are greatly influenced by this approximation error; see the Supplementary Information for a detailed analysis.

Our estimates best fit the participants' data (even penalizing the additional free parameters) when the Fisher information is allowed to differ depending on the prior distribution. Context-dependent encoding of this kind is predicted by theories of efficient coding; this leads us to ask whether the differing encoding rules represent efficient adaptations to the different priors, in the sense of maximizing average reward in our task. We thus examine the performance, over numbers sampled from a given prior (say, Downward), of a model participant equipped with the Fisher information fitted to data in the context of another prior (say, Upward). In other words, we look at how successful the 'Upward encoding' (and associated Bayesian decoding) would be on 'Downward data' (we use these shorthands below). Defining the 'performance ratio' as the fraction of the value in each trial that is captured by the participant (Methods), we find that with Downward data, the Downward encoding yields the

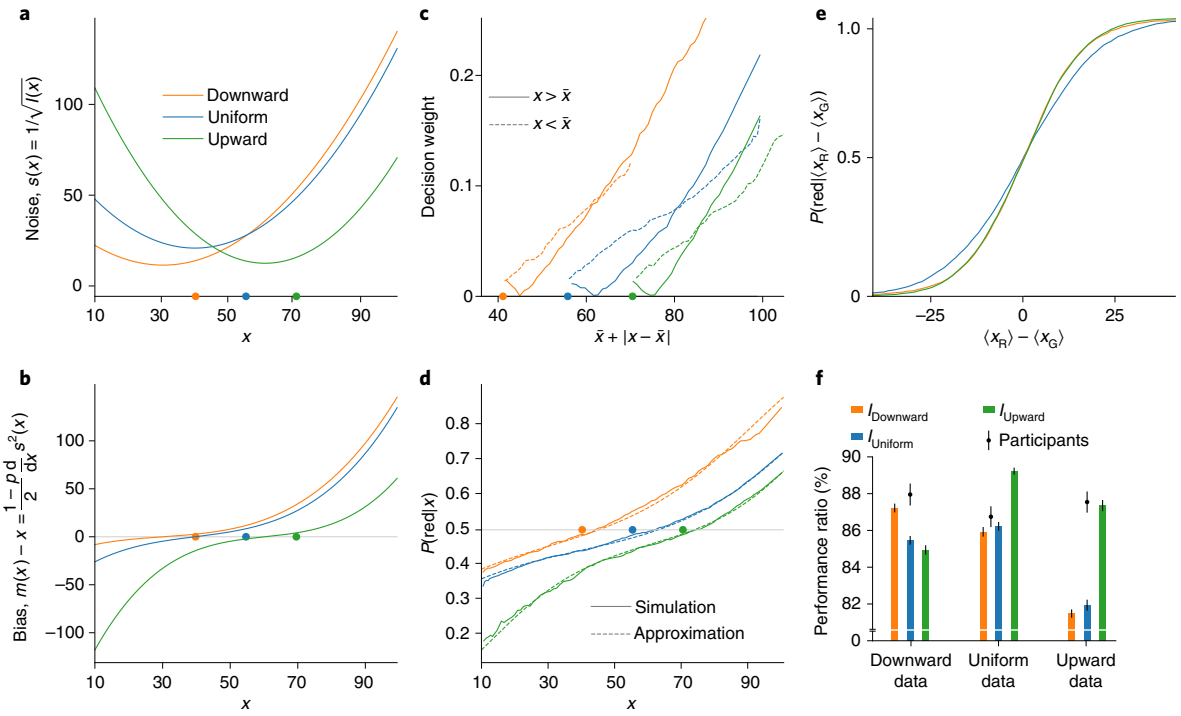

Fig. 5 | The efficient-coding, Bayesian-decoding model reproduces the participants' behaviour, and the Fisher information functions fitted to the participants' data improve the performance ratio in the Upward and Downward conditions. a, Noise $s(x)$ implied by the Fisher information in the three prior conditions. **b**, Bias $m(x) - x$ in the three prior conditions, also implied by the Fisher information. **c**, Decision weights $|P(\text{red}|x) - 0.5|$ predicted by the efficient-coding, Bayesian-decoding model. **d**, Probability of choosing 'red' conditional on a red number x being presented, in the model (solid lines) and our approximation of the model prediction (dashed lines; Methods). **e**, Choice probabilities implied by the model as a function of the difference between the averages of the numbers presented. **f**, Performance ratios over numbers sampled from the Downward (left), Uniform (middle) and Upward (right) priors, for the participants (black dots) and implied by the estimated encoding rules (bars) for each of the three prior conditions. With numbers sampled from the Downward prior, the Downward encoding rule, I_{Downward}, yields the best performance (left orange bar), whereas with numbers sampled from the Upward prior, the Upward encoding rule, I_{Upward}, results in the best performance (right green bar). The black vertical lines represent the standard deviations of the performance ratios (for the model, each standard deviation is obtained from 30,000 computations of the performance ratio of the model simulated with all the trials faced by the participants; for the participants, this quantity is bootstrap-estimated from the data, with 30,000 bootstrap samples in each case). All analyses use the participants' combined data.

largest performance ratio (87.2%; s.d., 0.47%), whereas the Upward encoding results in the lowest ratio (84.9%; s.d., 0.51%; Fig. 5f, left bars). Conversely, with Upward data, the Upward encoding outperforms the Downward encoding (87.4% versus 81.5%; s.d., 0.44% and 0.56%; Fig. 5f, right bars). In other words, the encoding rule fitted to the behaviour of the participants in the context of a given prior, Upward or Downward, results in higher rewards in the context of numbers sampled from this same prior, suggesting that the choice of encoding rule is efficient. With Uniform data, the performance ratio of the Uniform encoding is slightly better than that of the Downward encoding rule, but not quite as good as that of the Upward encoding rule (86.2% versus 85.9% and 89.2%; s.d., 0.47%, 0.47% and 0.39%). However, the participants' performance ratio in the Uniform condition (86.8%; s.d., 1.1%) is lower than that in the Upward condition (87.6%; s.d., 1.2%). Consistent with this, the fitted Uniform encoding rule is noisier than the Upward one; and it is primarily this fact (rather than the way that the precision of encoding varies for different numbers) that makes the fitted Uniform encoding less efficient even for Uniform data.

Discussion

We designed an average-comparison task in which we changed the prior distribution from which numbers were sampled across blocks of trials. In their choices, the participants seemed to differentially weight numbers that should be equally relevant to the correct decision. The participants' behaviour can be characterized by a model of noisy perception of the size of numbers, in which both the average estimation bias and the variability of estimates depend on the magnitude of the number. Furthermore, we introduced an encoding–decoding model, in which the variable precision of encoding is specified by a Fisher information function that is efficiently chosen with regard to a subjective prior, and the decoded value of each number is the Bayesian-mean estimate based on the encoded evidence. This is equivalent to a constrained version of the model of noisy perception (equation (5)), in which both the bias and the variable noise are determined by a single function, the Fisher information. This implies a relation between the bias and the derivative of the noise variance (equation (4)), for which we find qualitative support in the data. The efficient-coding, Bayesian-decoding model yields the lowest BIC among the models considered and reproduces the patterns observed in behaviour. Furthermore, the Fisher information fitted to the participants' data varies depending on the prior distribution. With the two skewed priors (Upward and Downward), the Fisher information is lower for numbers less likely to appear under the prior (Fig. 4). This resulted in a higher performance in the task (Fig. 5f), suggesting that the participants efficiently adapted their decision-making process to the prior distribution of the presented numbers. Our results are supported by a Bayesian random-effect analysis, which takes into account the participants' heterogeneity (Supplementary Information).

Spitzer et al.[12] had previously found that a model featuring a nonlinear transformation of presented numbers (that is, a bias) reproduced the apparent unequal weighting of numbers in participants' decisions, and they interpreted the bias as a strategy to compensate for decision noise. We provide a different account, which relies on two further observations. First, the noise in the perceived value of a number seems to vary with the number and to depend on the prior (Fig. 3a; see also the Supplementary Information, in which we exhibit behavioural patterns that are captured by the models featuring both nonlinear transformation and varying noise, but not by the model that features a nonlinear transformation only). Second, the way that the bias and the noise function vary from one condition to another is consistent with the theoretical relation between these two functions predicted by our model of efficient coding and Bayesian decoding. Under this account, estimation noise, estimation bias and thus the unequal weighting of numbers in average comparisons and the sigmoid-shape choice probability function all derive from the form of the Fisher information function. In turn, the Fisher information is an increasing function (at least roughly) of the prior density from which numbers are sampled, suggesting that the encoding is efficiently adapted to the prior.

We do not, however, reject the hypothesis that there may also be noise in the decision process. Our models, in fact, are not inconsistent with this hypothesis. Decision noise could be included in our models by adding a noise term to the decision variable (defined as the difference between the estimated empirical averages), as in ref. [12]. This would add a constant to the expression inside the square-root sign in equation (2), but the same effect would result from adding a constant to the function $s^2(x)$ in our model. The noise functions that we fit can be understood as implicitly already including this constant; note that if we assume that $s^2(x)$ is equal to $1/I(x)$ plus a constant (rather than simply $1/I(x)$, as stated above), the relationship between the bias and the variance of estimates implied by equation (4) remains the same, and our conclusions about the best-fitting function $I(x)$ for each prior will remain the same. We find it plausible that some amount of noise perturbs the decision and not only the estimation of the numbers[6,31]. But as introducing decision noise would not change our conclusions, we have focused, in the work reported above, on estimation noise and on how it adapts efficiently to the prior.

The idea that the nervous system encodes stimuli efficiently was first proposed by Barlow[1] and finds a modern formulation in the recent literature on neural coding[5,23–25,28,32]. Previous studies have provided experimental evidence that the encoding of stimuli is efficiently adjusted to the stimuli statistics[5,19,21,24,26,30,33]. With the exception of the "contextual modulation" presented in ref. [26], a common assumption is that the encoding strategy is permanently adapted to some unchanging distribution of stimuli, encountered in nature over long timescales. Here we show, conversely, that participants are able to adapt their encoding over a relatively short timescale, the one-hour time frame of the experiment (an adaptation to a change in the prior, within the time frame of an experimental session, is also found in ref. [34]). This raises the question, which we leave for future investigations, of how the brain dynamically tunes its representational capacity to changing environmental statistics.

Our study further differs from these psychophysics results in that the stimuli presented are Arabic numerals. The variables relevant for decisions are thus not physical magnitudes (such as the lengths of bars) but magnitudes represented in symbolic form, which one could expect to be processed in a different way (for example, '25' and '52' have the same size and luminosity but carry different semantic meanings). Studies on numerosity perception, however, suggest that there are important analogies with sensory perception[35–39]. For instance, participants exhibit scalar variability in various numerosity tasks[40–42], which has also been interpreted as resulting from an efficient semantic representation of numbers, adapted to a 'natural' number distribution[43]. Our results support the hypothesis that number processing can be understood using the same theoretical framework as accounts for sensory perception. Likewise, Polanía et al.[44] have used this framework to account for the subjective valuation of food items. In addition, we show that the encoding–decoding of numerals can quickly adapt to positively and negatively skewed distributions of numbers.

The efficient-coding, Bayesian-decoding model we examine predicts that estimates should be biased and that a simple mathematical relation exists between the bias and the derivative of the estimation variance (equation (4)). A quantity related to the estimation variance and often measured in perceptual tasks is the discrimination threshold, $\mathrm{DT}(x)$, which quantifies the sensitivity of an observer to small changes in the stimulus. From equation (4), we can derive (as shown in the Methods) a relation between the bias and the discrimination threshold, as

$$\mathbb{E}(\hat{x} - x) \propto \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{DT}^2(x). \tag{6}$$

Wei and Stocker[26] derive a relationship of this form, but under an assumption that the encoding maximizes the mutual information between the number and its internal representation. They show that it is supported by numerous empirical results obtained in perceptual tasks, and thus they call it a "law of human perception". Morais and Pillow[28] obtain this relation when encoding solves a Fisher-information optimization problem like the one that we consider (equation (3)), but assuming that estimates are given by a maximum-a-posteriori estimator, while we assume a Bayesian-mean estimator. Moreover, in the version of equation (6) that they derive from the maximum-a-posteriori estimator, the constant of proportionality is necessarily negative in sign, whereas a positive sign is implied by the Bayesian-mean estimator (when $p<1$), and this is required in order to fit our empirical results.

An assumption of the efficient-coding, Bayesian-decoding model is that the participants hold a subjective prior about the numbers. If the participants' prior matched the actual distribution of numbers used in each of the three prior conditions, the noise function, $s(x)$, should equal the inverse prior raised to the exponent $1/(p+1)$. Although in the Downward and Upward conditions the two functions are qualitatively comparable, discrepancies remain; and in the Uniform condition, the prior is flat, while the noise function is U-shaped and larger for large numbers (Fig. 4). A possible explanation is that the participants may not hold the correct prior. They might, for instance, start the experiment with a prior that corresponds to the frequencies of numbers that one usually encounters, and not fully update these prior beliefs during the task, in spite of our efforts to familiarize them with the correct distribution (Methods). The frequency of numeric quantities typically observed approximately follows a power law[43,45,46]—that is, a distribution skewed towards small numbers. In the Uniform condition, if the subjective prior is between the natural, power-law prior and the correct, Uniform prior, it should also be skewed towards small numbers to some extent. The efficient noise function, in this case, would be larger for large numbers, which is precisely what we observe in the participants' data (Fig. 4). Incorrect subjective priors are thus a plausible source of the observed discrepancies between the participants' noise functions and those predicted by the correct priors.

A deviation of the noise functions away from theoretical predictions is found near the endpoints of the interval of presented numbers: the noise increases, despite the rising prior density, near the lower endpoint in the Downward condition and near the upper endpoint in the Upward condition. The model we investigate relies on a small-noise assumption (in particular, the quantity minimized in equation (3) is a lower bound on the mean L_p error only in the limit of a vanishing noise[28]). When noise is large, the optimal Fisher

information differs from a simple power of the prior—in particular at the boundaries[25], if the mutual information is maximized (the $p \to 0$ limit). The optimal Fisher information, for a number x, presumably does not depend only on the local value of the prior at this number, $\pi(x)$, but depends instead on a larger neighbourhood around it. This should distort the small-noise solution, especially at the endpoints, where the prior vanishes abruptly after reaching its highest value. The participants' data seem consistent with this account, but it calls for further theoretical investigation of the case in which noise is not negligible.

The observed discrepancies might also be explained by principles other than efficient coding as we define it. For example, an alternative hypothesis might be that participants adapt their encoding strategy to the prior but choose to encode with higher precision the numbers that are close to the prior mean, instead of optimizing the program described by equation (3). The three best-fitting noise functions, indeed, reach their minima near the priors' means (Figs. 3a and 5a). This could be a general strategy used for stimulus encoding in various situations, or it could stem from the specifics of our tasks, in which the participants are asked to compare empirical averages. In any case, we note that the proportionality relation between the bias and the derivative of the noise variance (equation (4)) relies on the choice of a noise function that is efficient in the sense of our optimization program (equation (3)). A different choice of encoding, such as one that minimizes the noise near the mean, would result in a different relation (if any), and presumably not one of proportionality. However, when fitting the bias and the noise separately, we find that they are consistent with the proportionality relation (Fig. 3b); and the model that directly implies this relation is our best-fitting model. (Besides, in sensory domains, the proportionality relation is supported by many experimental results, as mentioned above.) We cannot, however, reject this alternative hypothesis; it calls for finer investigations on the influence of task reward and prior distribution on the choice of an encoding strategy.

Methods

Experiment, participants and reward. The procedures of this experiment comply with the relevant ethical regulations and were approved by the Institutional Review Board of Columbia University (Protocol Number IRB-AAAR9375). The experiment included 37 participants, 18 female and 19 male, aged 24.5 on average (s.d., 8.6). All participants provided informed consent. All but one participant participated in two blocks of trials, in each of which all numbers were samples from a single distribution (Uniform, Upward or Downward), so that each participant (except one) experienced two of the three conditions. Each session lasted about one hour. The participants were explicitly told the current distribution, and at the beginning of each new block, they were presented with a series of random samples to familiarize them with the distribution. On screen, the numbers were presented with two decimal digits and for a duration of 500 ms. In each trial, the colour of the first presented number was chosen between red and green with equal probability. The score of the participant was augmented at each trial by the chosen average and thus increased over the course of the experiment. At the end of the experiment, the participant received a financial reward, which was a linear function of the total score, with a US$10 'show-up' minimum. The expected reward, not taking into account the US$10 minimum, for a hypothetical participant providing random responses (that is, choosing 'red' with probability 0.5 at all trials) was US$10, and an accuracy of 80% correct responses yielded an average of US$25. The average reward over the 37 participants was US$28 (s.d., 3.8). Thirty-three participants experienced two blocks of 200 trials (so that $2 \times 200 = 400$ decisions were collected per participant), three participants experienced two blocks of 210 trials and one participant experienced one block of 210 trials (totalling 14,670 trials). Three trials were excluded from the analysis because the number 100.00 erroneously appeared in these trials. A total of 4,610 responses were collected in the Downward condition, 5,040 in the Uniform condition and 5,017 in the Upward condition. The task was coded in the Python programming language and with the help of the PsychoPy library[47]. The data analysis and the implementation of the models were conducted using Python and the libraries Numpy[48], SciPy[49] and Matplotlib[50].

The transformation $m(x)$ and decision weights. To shed light on the relation between the transformation $m(x)$ and the decision weights, we derive an approximation of the probability $P(\text{red}|x)$ of choosing 'red' conditional on a red number x being presented. Consider first the model with constant noise ($\mathrm{N}(m(x), s^2)$). The probability can be computed by marginalization of the probability of choosing 'red' conditional on ten numbers as

$$\begin{aligned} P(\text{red}|x) &= \int \dots \int P(\text{red}|x, x^{\mathrm{R}}_{2:5}, x^{\mathrm{G}}_{1:5}) \pi(x^{\mathrm{R}}_2) \dots \pi(x^{\mathrm{G}}_5) \mathrm{d}x^{\mathrm{R}}_2 \dots \mathrm{d}x^{\mathrm{G}}_5 \\ &= \int \Phi\left(\tfrac{m(x)+\Delta_m}{s\sqrt{10}}\right) \pi_\Delta(\Delta_m) \mathrm{d}\Delta_m, \end{aligned}$$

where

$$\Delta_m \equiv \sum_{i=2}^{5} m(x^{\mathrm{R}}_i) - \sum_{i=1}^{5} m(x^{\mathrm{G}}_i) \tag{7}$$

and π_Δ is the prior density for this random quantity. We approximate π_Δ by a Gaussian distribution with the same mean and variance, so that

$$\pi_\Delta \approx \mathrm{N}(-\bar{m}, 9\mathrm{Var}m), \tag{8}$$

where $\bar{m}$ is the mean of $m(x)$ under the prior and Var$m$ is the variance. Substituting this approximation in equation (7) results in

$$P(\text{red}|x) \approx \Phi\left(\frac{m(x) - \bar{m}}{\sqrt{10s^2 + 9\mathrm{Var}m}}\right). \tag{9}$$

We find this approximation to be fairly close to the conditional probabilities obtained through simulations of the model. In case of variable noise ($\mathrm{N}(m(x), s^2(x))$), we replace in equation (9) the noise variance s^2 by its average under the prior, $\mathbb{E}s^2(x)$, and despite this coarse approximation, we also obtain a close match to simulated data. Fig. 5d provides an example of the quality of these approximations in the case of the efficient-coding, Bayesian-decoding model.

Equation (9) implies that the decision weight, $|P(\text{red}|x) - 0.5|$, vanishes at approximately the number whose transformation equals the average transformation. In the absence of bias ($m(x) = x$), this number would be the prior mean; but it is slightly greater in the case of the fitted transformations. Hence, the decision weights at the prior mean do not fall to zero, and numbers above and below the prior mean consequently have different weights (Fig. 2c).

Model fitting and identification. When fitting the participants' data to the general one-stage model $\mathrm{N}(m(x), s^2(x))$, it is important to note that the transformation $m(x)$ can be identified from choice data only up to an arbitrary affine transformation: a model with transformation $\alpha m(x) + \beta$ and noise $\alpha s(x)$ makes the same predictions as the model with transformation $m(x)$ and noise $s(x)$, for any non-zero α and any β (see equation (2)). In calculating the bias plotted in Fig. 3b, we choose the 'scale and location' parameters α and β to satisfy two additional desiderata. First, for any choice of α, we choose β so that the left and right sides of equation (4) are exactly equal at the particular value of x where the right side is equal to zero. And second, given this, we choose α to minimize

$$\frac{\int (\mathrm{LHS}(x) - \mathrm{RHS}(x))^2 \mathrm{d}x}{\langle |\mathrm{RHS}| \rangle^2},$$

a measure of the relative difference between the two functions LHS(x) and RHS(x) defined by the two sides of the equation.

Optimal Fisher information. The Fisher information function $I(x)$ that solves the optimum problem defined in equation (3) is obtained through variational calculus. Let J be the functional

$$J[I] = \int \tilde{\pi}(x) I(x)^{-p/2} \mathrm{d}x + \lambda \int \sqrt{I(x)} \mathrm{d}x, \tag{10}$$

where λ is the Lagrange multiplier associated with the encoding constraint in equation (3). Taking the functional derivative with respect to I and setting to zero, we obtain

$$-\frac{p}{2} \tilde{\pi}(x) I(x)^{-p/2-1} + \frac{1}{2} \lambda I(x)^{-1/2} = 0, \tag{11}$$

and thus

$$I(x)^{\frac{p+1}{2}} \propto \tilde{\pi}(x). \tag{12}$$

Efficient-coding, Bayesian-decoding model: relations between bias, variance, Fisher information and Cramér–Rao bound. Reference [29] shows that with the assumed form of encoding (a series of n independent signals, characterized by a total Fisher information $I(x)$) and with a Bayesian-mean decoding that uses the prior $\tilde{\pi}(x)$, the bias of the decoder is asymptotically approximated by

$$b(x) = \frac{1}{I(x)} \left[\frac{\tilde{\pi}'(x)}{\tilde{\pi}(x)} - \frac{I'(x)}{I(x)}\right], \tag{13}$$

where $\tilde{\pi}'(x)$ and $I'(x)$ are the derivatives of $\tilde{\pi}(x)$ and $I(x)$. The Fisher information is of order n; thus, the bias is of order $1/n$. The variance is asymptotically approximated by $1/I(x)$; this approximation is accurate to order $1/n$. We use these approximations to define the bias and the variance of the efficient-coding, Bayesian-decoding model. The Cramér–Rao lower bound on the variance of biased estimators is

$$\frac{(1+b'(x))^2}{I(x)}, \tag{14}$$

where $b'(x)$ is the derivative of the bias. Because equation (13) implies that the bias $b(x)$ is of order $1/n$, this bound is equal to $1/I(x)$, to order $1/n$. Note that to order $1/n$, this is the variance of the Bayesian posterior mean decoder, as reported in equation (5). Thus, in our model, the difference between the Cramér–Rao bound and the variance of the estimates is within the magnitude of our approximation errors and is asymptotically vanishing.

As for the bias (equation (4)), the choice of the Fisher information (equation (12)) implies that

$$\frac{\tilde{\pi}'(x)}{\tilde{\pi}(x)} = \frac{p+1}{2}\frac{I'(x)}{I(x)}, \tag{15}$$

and thus the implied bias is

$$b(x) = \frac{p-1}{2}\frac{I'(x)}{I^2(x)} = \frac{1-p}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{I(x)}\right). \tag{16}$$

A different functional form: power function. In the main text, we have chosen polynomials to fit the functions $m(x)$ and $s(x)$ with a small number of parameters. Here we investigate another functional form in which the bias is a sign-conserving power function, similar to the nonlinear transformation used in ref. [6]. More precisely, we study the model implied by equation (5), this time assuming that the noise variance, $s^2(x)$, is written as

$$s^2(x) = s_0^2 + \frac{a}{k+1}|x-c|^{k+1}. \tag{17}$$

The implied bias (equation (4)) is then a sign-conserving power function:

$$\mathbb{E}(\hat{x}-x) = \frac{1-p}{2}\, a\, \mathrm{sign}(x-c)|x-c|^k. \tag{18}$$

We fit the model with this functional form to our data and find that its log likelihood is very close to (and slightly higher than) that of the same model fit with a polynomial function (−6,750.46 versus −6,751.87). The power-function approach, however, uses four free parameters (a, c, k and s_0^2)—that is, one more than the polynomial approach (in which $s(x)$ is a quadratic polynomial). As a result, the BIC with the power function is larger than that with the polynomial function (13,612 versus 13,589), so we conclude that the polynomial form is more parsimonious in its account of the data.

The two functional forms yield very similar results. In the Results, to be consistent with the other models and for the sake of concision, we present only the results obtained with the polynomial approach. Our results, however, do not depend on a specific choice of functional form.

Performance ratio. We measure the performance of a participant as follows. In each trial of our task, the score is augmented by the average that is chosen by the participant, so that the participant is guaranteed to receive at least the minimum of the two averages. The additional value to be captured in trial i is thus the absolute difference between the two averages—that is, $\Delta_i = |\langle x^{\mathrm{R}}\rangle_i - \langle x^{\mathrm{G}}\rangle_i|$. We define the 'performance ratio' as the fraction of this value captured by a participant—that is, $\sum\delta_i\Delta_i/\sum\Delta_i$, where δ_i is 1 if the participant makes the correct choice in trial i and 0 otherwise; and the sums are taken over all the trials in which the numbers are sampled from a given prior.

Relations between bias, variance and discrimination threshold. The discrimination threshold $\mathrm{DT}(x)$ is defined as the difference δ in stimulus magnitude for which a participant distinguishes two stimuli x and $x+\delta$ with a given success rate (for example, 75%). In a model in which an estimate $\hat{x}$ of presented number x is normally distributed around a transformation $m(x)$ of the number, with varying noise $s(x)$—that is, $\hat{x} \sim \mathrm{N}(m(x), s^2(x))$—the probability of telling $x_1 = x$ and $x_2 = x+\delta$ apart is (assuming δ is small)

$$\begin{aligned} P(\hat{x}_2 > \hat{x}_1) &= \Phi\left(\frac{m(x_2)-m(x_1)}{\sqrt{s^2(x_2)+s^2(x_1)}}\right) \\ &\approx \Phi\left(\frac{m'(x)}{s(x)}\delta\right) \\ &\approx \tfrac{1}{2} + \frac{1}{\sqrt{2\pi}}\frac{m'(x)}{s(x)}\delta. \end{aligned} \tag{19}$$

This implies

$$\mathrm{DT}(x) \propto \frac{s(x)}{m'(x)}, \tag{20}$$

where the proportionality factor depends on the chosen target success rate. In the efficient-coding, Bayesian-decoding model, the bias $m(x)-x$ is proportional to the derivative of the inverse of the Fisher information; therefore, the bias is of order $\mathcal{O}(1/n)$, and $m'(x) \approx 1$. Equations (20) and (4) then immediately result in Wei and Stocker's law of human perception[26] (equation (6)):

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{DT}^2(x) &\propto \tfrac{\mathrm{d}}{\mathrm{d}x}s^2(x) \\ &\propto m(x) - x = \mathbb{E}(\hat{x}-x). \end{aligned} \tag{21}$$

We also note that the converse derivation appears in the supporting information of ref. [26]: from the relation involving the discrimination threshold (equation (6)), the authors derive a relation involving the noise variance, as in equation (4).

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

Data availability

The data relating to this study are available at https://doi.org/10.7916/tn94-qn62 (ref. [51]).

Code availability

Scripts for analysing the data, implementing the models, fitting their parameters and producing the results and figures are available at https://doi.org/10.7916/tn94-qn62 (ref. [51]).

Received: 21 February 2020; Accepted: 12 April 2022;
Published online: 30 May 2022

References

1. Barlow, H. B. in *Sensory Communication* (ed. Rosenblith, W. A.) 217–234 (MIT Press, 1961).
2. Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).
3. Ma, W. J., Beck, J. M. & Pouget, A. Spiking networks for Bayesian inference and choice. *Curr. Opin. Neurobiol.* **18**, 217–222 (2008).
4. Savin, C. & Denève, S. Spatio-temporal representations of uncertainty in spiking neural networks. *Adv. Neural Inf. Process. Syst.* **27**, 2024–2032 (2014).
5. Ganguli, D. & Simoncelli, E. P. Neural and perceptual signatures of efficient sensory coding. Preprint at https://arxiv.org/abs/1603.00058 (2016).
6. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1398–1411 (2016).
7. Tversky, A. Elimination by aspects: a theory of choice. *Psychol. Rev.* **79**, 281–299 (1972).
8. Payne, J. W., Bettman, R. & Johnson, E. J. *The Adaptive Decision Maker* (Cambridge Univ. Press, 1993).
9. Gigerenzer, G. & Goldstein, D. G. Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* **103**, 650–669 (1996).
10. Johnson, E. J. & Ratcliff, R. in *Neuroeconomics* (eds Glimcher, P. & Fehr, E.) 35–48 (Elsevier, 2014).
11. Summerfield, C. & Tsetsos, K. Do humans make good decisions? *Trends Cogn. Sci.* **19**, 27–34 (2015).
12. Spitzer, B., Waschke, L. & Summerfield, C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat. Hum. Behav.* **1**, 0145 (2017).
13. Li, V., Castañon, S. H., Solomon, J. A., Vandormael, H. & Summerfield, C. Robust averaging protects decisions from noise in neural computations. *PLoS Comput. Biol.* **13**, e1005723 (2017).
14. De Gardelle, V. & Summerfield, C. Robust averaging during perceptual judgment. *Proc. Natl Acad. Sci. USA* **108**, 13341–13346 (2011).
15. Tsetsos, K. et al. Economic irrationality is optimal during noisy decision making. *Proc. Natl Acad. Sci. USA* **113**, 3102–3107 (2016).
16. Knill, D. C. & Richards, W. *Perception as Bayesian Inference* (Cambridge Univ. Press, 1996).
17. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
18. Stocker, A. A. & Simoncelli, E. P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).
19. Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.***14**, 926–932 (2011).

20. Petzschner, F. H., Glasauer, S. & Stephan, K. E. A Bayesian perspective on magnitude estimation. *Trends Cogn. Sci.* **19**, 285–293 (2015).
21. Wei, X.-X. & Stocker, A. A. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).
22. Clarke, B. S. & Barron, A. R. Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Stat. Plan. Inference* **41**, 37–60 (1994).
23. Brunel, N. & Nadal, J. P. Mutual information, Fisher information, and population coding. *Neural Comput.* **10**, 1731–1757 (1998).
24. Ganguli, D. & Simoncelli, E. P. Implicit encoding of prior probabilities in optimal neural populations. *Adv. Neural Inf. Process. Syst.* **2010**, 658–666 (2010).
25. Wei, X.-X. & Stocker, A. A. Mutual information, Fisher information, and efficient coding. *Neural Comput.* **326**, 305–326 (2016).
26. Wei, X.-X. & Stocker, A. A. Lawful relation between perceptual bias and discriminability. *Proc. Natl Acad. Sci. USA* **114**, 10244–10249 (2017).
27. 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).
28. Morais, M. & Pillow, J. W. Power-law efficient neural codes provide general link between perceptual bias and discriminability. *Adv. Neural Inf. Process. Syst.* **31**, 5076–5085 (2018).
29. Prat-Carrabin, A. & Woodford, M. Bias and variance of the Bayesian-mean decoder. *Adv. Neural Inf. Process. Syst.* **34**, 23793–23805 (2022).
30. Wei, X.-X. & Stocker, A. A. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. *Adv. Neural Inf. Process. Syst.* **25**, 1304–1312 (2012).
31. Castañón, S. H. et al. Human noise blindness drives suboptimal cognitive inference. *Nat. Commun.* **10**, 1719 (2019).
32. McDonnell, M. D. & Stocks, N. G. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Phys. Rev. Lett.* **101**, 058103 (2008).
33. Stocker, A. A. & Simoncelli, E. P. Sensory adaptation within a Bayesian framework for perception. *Adv. Neural Inf. Process. Syst.* **18**, 1291–1298 (2006).
34. Heng, J. A., Woodford, M. & Polania, R. Efficient sampling and noisy decisions. *eLife* **9**, e54962 (2020).
35. Moyer, R. S. & Landauer, T. K. Time required for judgements of numerical inequality. *Nature* **215**, 1519–1520 (1967).
36. Parkman, J. M. Temporal aspects of digit and letter inequality judgments. *J. Exp. Psychol.* **91**, 191–205 (1971).
37. Hinrichs, J. V., Yurko, D. S. & Hu, J. M. Two-digit number comparison: use of place information. *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 890–901 (1981).
38. Pinel, P., Dehaene, S., Rivière, D. & LeBihan, D. Modulation of parietal activation by semantic distance in a number comparison task. *NeuroImage* **14**, 1013–1026 (2001).
39. Kutter, E. F., Bostroem, J., Elger, C. E., Mormann, F. & Nieder, A. Single neurons in the human brain encode numbers. *Neuron* **100**, 753–761.e4 (2018).
40. Whalen, J., Gallistel, C. R. & Gelman, R. Nonverbal counting in humans: the psychophysics of number representation. *Psychol. Sci.* **10**, 130–137 (1999).
41. Izard, V. & Dehaene, S. Calibrating the mental number line. *Cognition* **106**, 1221–1247 (2008).
42. Dehaene, S., Izard, V., Spelke, E. & Pica, P. Log or linear? Distinct intuitions of the number scale in western and Amazonian indigene cultures. *Science* **320**, 1217–1220 (2008).
43. Cheyette, S. J. & Piantadosi, S. T. A unified account of numerosity perception. *Nat. Hum. Behav.* **4**, 1265–1272 (2020).
44. Polanía, R., Woodford, M. & Ruff, C. C. Efficient coding of subjective value. *Nat. Neurosci.* **22**, 134–142 (2019).
45. Dehaene, S. & Mehler, J. Cross-linguistic regularities in the frequency of number words. *Cognition* **43**, 1–29 (1992).
46. Piantadosi, S. T. & Cantlon, J. F. True numerical cognition in the wild. *Psychol. Sci.* **28**, 462–469 (2017).
47. Peirce, J. et al. PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).
48. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
49. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
50. Hunter, J. D. Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
51. Prat-Carrabin, A. & Woodford, M. Efficient coding of numbers explains decision bias and noise: data and code. Columbia University Academic Commons https://doi.org/10.7916/tn94-qn62 (2022).

Acknowledgements

We thank B. Ho for his outstanding help as a research assistant, the National Science Foundation for research support (grant no. SES DRMS 1949418, M.W.) and the Italian Academy for Advanced Studies in America at Columbia University for research support (Spring 2021 Fellowship, A.P.-C.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.W. conceptualized the study. M.W. and A.P.-C. designed the experiment. A.P.-C. implemented the task and collected the data. M.W. and A.P.-C. analysed the data and wrote the computational models. A.P.-C. implemented the models. M.W. and A.P.-C. interpreted the results and wrote the manuscript.

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/s41562-022-01352-4.

Correspondence and requests for materials should be addressed to Arthur Prat-Carrabin.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers 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.

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

nature portfolio

Corresponding author(s): Arthur Prat-Carrabin

Last updated by author(s): Apr 8, 2022

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	Custom code in Python 2.7, using the open-source library PsychoPy 2 (v1.90.2).
Data analysis	Custom code in Python 3.7, using the open-source libraries NumPy 1.16.2, SciPy 1.2.1, and Matplotlib 3.0.3.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The manuscript includes a data availability statement. The data is available at https://doi.org/10.7916/tn94-qn62 .

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

Behavioural & social sciences study design

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

Study description	Quantitative experimental.
Research sample	37 subjects (18 female), aged 18-63, (mean 24.5, s.d. 8.6), recruited on Columbia University campus. The sample is representative of a population of healthy adults on an American university campus, and is thus skewed towards younger ages.
Sampling strategy	Subjects were recruited via fliers. Sample size was chosen to be comparable to a similar study (Spitzer et al., 2017).
Data collection	Data were collected through a computer-based task in which visual stimuli were presented to subjects, and responses (key presses on a keyboard) were recorded. The experimental sessions took place in an office in which only the subject and the experimenter were present. The experimenter was aware of the study hypothesis, but did not know in advance the experimental condition that a subject would face.
Timing	From July 10 to September 5, 2018.
Data exclusions	No participant was excluded from the study. Out of the 14,670 responses collected, 3 were excluded from the analysis because of an error in the presented stimuli (the number 100.00 was presented when the maximum should have been 99.99).
Non-participation	No participants dropped out.
Randomization	The random order of arrival of each participant was used to cycle through the list of possible conditions, in order to determine which condition this subject would be assigned to.

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
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Subjects were recruited via fliers on Columbia University campus. A potential bias is that the population recruited may be more used to reasoning about numbers than the average population. This may have a quantitative impact (e.g., on the performance of subjects) if compared with the general population, but it is not likely to change our qualitative results nor our main conclusions.
Ethics oversight	Institutional Review Board of Columbia University.

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