ARTICLES
DOI: 10.1038/s41562-017-0215-1
nature human behaviour

The idiosyncratic nature of confidence

Joaquin Navajas[1,2*], Chandni Hindocha[1,3], Hebah Foda[1], Mehdi Keramati[4], Peter E. Latham[4] and Bahador Bahrami[1]

Confidence is the 'feeling of knowing' that accompanies decision-making. Bayesian theory proposes that confidence is a function solely of the perceived probability of being correct. Empirical research has suggested, however, that different individuals may perform different computations to estimate confidence from uncertain evidence. To test this hypothesis, we collected confidence reports in a task in which subjects made categorical decisions about the mean of a sequence. We found that for most individuals, confidence did indeed reflect the perceived probability of being correct. However, in approximately half of them, confidence also reflected a different probabilistic quantity: the perceived uncertainty in the estimated variable. We found that the contribution of both quantities was stable over weeks. We also observed that the influence of the perceived probability of being correct was stable across two tasks, one perceptual and one cognitive. Overall, our findings provide a computational interpretation of individual differences in human confidence.

Understanding the computational basis of individual differences in human cognition has fundamental implications for medical and biological sciences, as well as for economics and the social sciences. A prime example is confidence, which plays a key role in a wide range of aspects in life, including learning to make better decisions[1], monitoring our actions[2], cooperating effectively with others[3,4] and displaying good political judgement[5]. One of the most intriguing features of confidence is that humans tend to communicate this feeling in a largely idiosyncratic way: although confidence reports are typically stable within each person, they tend to be variable across the population[6,7]. For instance, different individuals performing the same task generate distributions of confidence ratings with different means and shapes[7]. In addition, the correlation between confidence and objective performance varies for different people, and is related to individual variations in brain structure[8] and connectivity[9,10]. While a vast literature has focused on the biological correlates of individual differences in human confidence[8–10], the computational roots of this phenomenon remain unclear.

Previous research in sensory psychophysics[8,11] and value-based decision-making[10] assumed that confidence is a function solely of the perceived probability of being correct. This assumption is reasonable: confidence should reflect only this subjective probability[12–14]. Driven by this normative framework, previous studies explained differences among people as measurement noise[15], or as individual differences in the ability to report the probability of being correct[8,9]. This may have been an oversimplification: there is extensive literature showing that confidence is influenced by factors other than the probability of being correct[16], such as the reliability of sensory stimuli[2,13], the magnitude of sensory data[11], post-decisional biases[17] and even personality traits[7].

Here we set out to determine what probabilistic quantities, besides perceived probability of being correct, contribute to individual differences in human confidence. We focused on a categorical task, in which subjects had to decide whether the mean of a set of items was above or below a decision boundary, and then report their confidence. For about half of the subjects, confidence did depend solely on the perceived probability that they were correct. However, for the other half, confidence also depended on a different statistical quantity: their uncertainty in the estimate of the mean[18,19]. Moreover, the dependence of confidence on the perceived probability of being correct and uncertainty was stable across experiments performed weeks apart. Finally, the dependence of confidence on the perceived probability of being correct was stable across tasks involving uncertainty in the perceptual and cognitive domain, but the dependence on the perceived uncertainty was not. This is consistent with the predictions of a recent theoretical account arguing that uncertainty is encoded by domain-specific neural populations[14]. Overall, these findings provide a computational interpretation of individual differences in the human sense of confidence.

Results

In a perceptual task (experiment 1), participants observed a sequence of 30 tilted Gabor patches presented at the fovea in rapid (4 Hz) serial visual presentation (Fig. 1a). At the end of the sequence, participants decided whether the mean orientation of the patches was clockwise or anticlockwise relative to vertical. Participants then reported how confident they were in their decision on a scale from 1 to 6. To manipulate uncertainty, we pseudo-randomly drew the orientation samples from uniform distributions with exactly the same mean (+3 degrees or −3 degrees) but different variances on different trials (Fig. 1b). Participants performed better as variance decreased (Fig. 1c, one-way repeated-measures analysis of variance (rm-ANOVA), $F(3,29) = 231.4$, $p < 10^{-10}$).

To fit the choices of each participant, we assumed that they keep track of the mean orientation, which they update after each stimulus presentation. To update their estimate of the mean within each trial, we considered a model in which participants combine a noisy estimate of the current sample with their previous estimate of the mean,

$$\mu_i = (1-\lambda)\ \mu_{i-1} + \lambda\theta_i + \gamma\theta_i\xi_i \qquad (1)$$

where μ_i is the estimate of the mean after i samples ($\mu_0 = 0$), $0 < \lambda < 1$ determines the relative weighting of recent versus more distant samples, θ_i is the actual orientation of the ith sample in the sequence, ξ_i is sampled from the standard normal distribution and γ is a free parameter indicating the strength of the noise. The multiplicative

[1]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, WC1N 3AZ, UK. [2]Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, Buenos Aires, C1428BCW Argentina. [3]Clinical Psychopharmacology Unit, University College London, Gower Street, London WC1E 6BT, UK. [4]Gatsby Computational Neuroscience Unit, University College London, 25 Howland Street, London W1T 4JG, UK. *e-mail: joaquin.navajas@utdt.edu

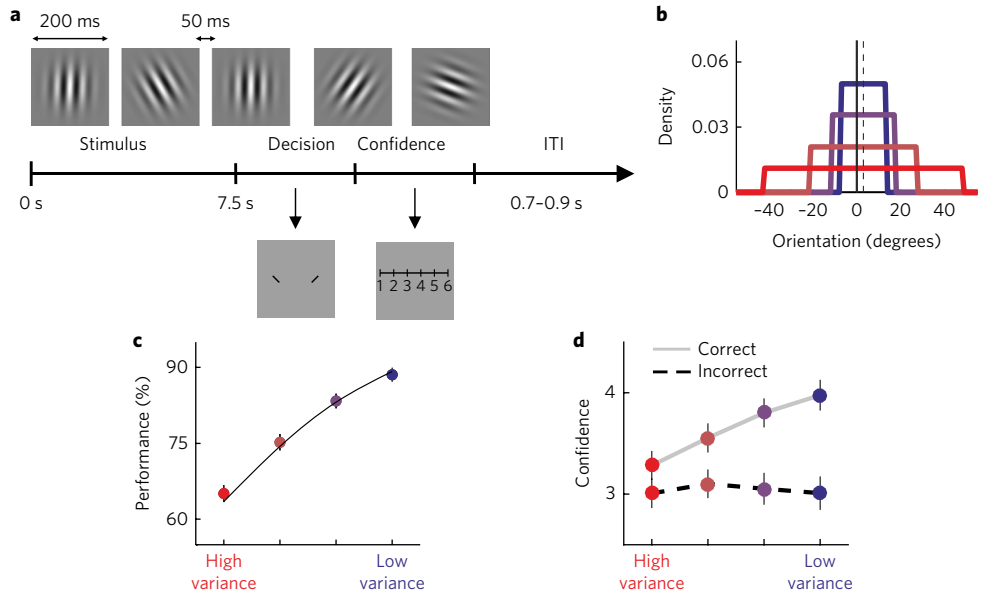

Fig. 1 | Tracking mean evidence in rapid serial visual presentations. a, Thirty tilted Gabor patches were serially flashed at the fovea, updated at 4 Hz. Participants made a binary decision about whether the mean in the sequence was tilted to the right or left, followed by a confidence rating. After an inter-trial interval (ITI), which was uniformly distributed between 0.7 and 0.9 seconds, a new trial began. Full details of the task are available in the Methods section. **b**, The samples were drawn from a uniform distribution with mean, m, set to either exactly +3 degrees or exactly −3 degrees. The dashed line shows $m=+3$. The endpoints of the uniform distributions were $m \pm v$, with $v=10$, 14, 24 or 45 degrees, yielding four conditions with four different variances. **c**, Performance increased with decreasing variance. The dots show the average performance across subjects, and the vertical lines depict the s.e.m. The solid black curve shows the best fit of the stochastic updating model (equations (1) and (2)). **d**, Confidence reports averaged over all subjects. The vertical lines show s.e.m. At the population level, confidence in incorrect trials remains approximately constant as a function of variance.

nature of the noise ensures that the uncertainty in the update of the estimate scales with the size of the observed sample, θ_i. At the end of the sequence, choice is determined by the sign of the final value of the mean (μ_{30}): the agent chooses clockwise if μ_{30} is positive, and anticlockwise if μ_{30} is negative.

This model explains two important quantitative patterns observed in our behavioural data. First, all items in the sequence had a significant influence on choice (regression weights against zero, $t(29)>3.17$, $p<0.003$ for all items), but later samples had more influence than earlier ones (slope of regression weights against zero, $t(29)=4.70$, $p=10^{-6}$). This recency effect was modulated by the learning rate, λ (Supplementary Fig. 1). Second, we observed that items in high-variance sequences had smaller influence on choice ($F(3,29)=57.8$, $p\sim0$), indicating larger integration noise in these trials. The last term in equation (1), modulated by γ, captures this pattern (Supplementary Fig. 2).

We also tested an alternative model that tracks the mean of the sequence in a deterministic way, and then makes stochastic decisions. This model, however, failed to explain the trend in Fig. 1c, which shows that performance increases as variance decreases (see Supplementary Fig. 3 for details and model comparison).

Computation of confidence. In this task, confidence should reflect the perceived probability of being correct, for which participants need to have an estimate of the variance of μ_{30}. We assumed that they are able to compute the true variance associated with equation (1) (although our findings do not require this assumption, see Supplementary Notes). Thus, perceived variance, denoted σ^2_{30}, is given by

$$\sigma^2_{30}=\gamma^2\sum_{i=1}^{30}(1-\lambda)^{2(30-i)}\theta_i^2. \qquad (2)$$

The model described by equations (1) and (2), which we call the stochastic updating model, is illustrated in Fig. 2a. Given μ_{30} and σ^2_{30}, subjects can compute, on each trial, the perceived probability of being correct, $\widehat{p}$ (correct) (shaded area under the Gaussian distribution in Fig. 2a).

Using this model, we estimated the expected values of $\widehat{p}$ (correct) for different variance conditions (see Methods, equation (9) and Fig. 2b). When we separated these values by correct and incorrect trials, we observed a pattern that has been suggested on the basis of normative arguments[15,20]: confidence on correct trials should increase as the variance decreases, whereas confidence on error trials should show the opposite effect, and decrease as the variance decreases. We did not, however, observe this pattern in our data, at least not on average: as shown in Fig. 1d, confidence on correct trials did indeed increase as variance dropped, but on error trials confidence was relatively independent of variance ($F(3,29)=0.57$, $p=0.63$).

This last observation indicates that, again on average, subjects were mis-estimating confidence: they should have been less confident on low-variance error trials than in high-variance error trials, as their probability of being correct was lower (dashed curve in Fig. 2b). This suggests that subjects partially based their confidence on the uncertainty in the value of the mean orientation—a reasonable, if suboptimal, heuristic. Under this heuristic, low-variance trials would raise their confidence relative to high-variance ones. An appropriate weighting of perceived probability of being correct, shown in Fig. 2b, and a function of uncertainty such as the observed Fisher information (the inverse of σ^2_{30}), shown in Fig. 2c, could, therefore, explain the confidence ratings observed in Fig. 1d.

To formally test this proposal, we compared the normative model of confidence based on only $\widehat{p}$ (correct) with seven alternative models based on different linear combinations of $\widehat{p}$ (correct), mean, standard deviation, variance and Fisher information (Supplementary Fig. 4). We evaluated which combination provided a better fit to confidence ratings using ordinal logistic regressions (see Methods).

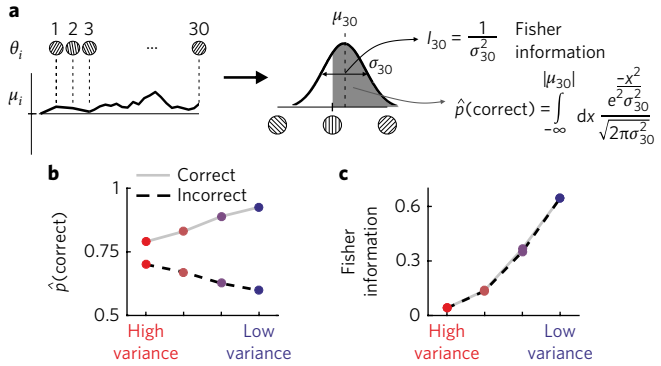

Fig. 2 | Estimating confidence. a, Each trial consists of 30 presentations of tilted Gabor patches. At each presentation (θ_i), the mean (μ_i) is updated by combining the estimate on the previous sample with a noisy version of the current Gabor patch. The black line represents one realization of the model. At the end of the sequence, the subject makes a decision based on the sign of μ_{30}. The subjective probability of being correct and the observed Fisher information are then computed according to the equations shown on the right; see Methods for full details. **b**, The perceived probability of being correct, $\hat{p}$(correct), averaged over variance condition for correct trials (solid grey line) and incorrect trials (dashed black line), and also averaged across participants. For correct trials, this quantity increases with decreasing variance (solid grey line); for incorrect trials, it shows the opposite pattern (dashed black line; see ref. 15 for more details). **c**, The uncertainty in the estimate of μ_{30}, quantified by the observed Fisher information, increases both for correct and incorrect trials (same markers as **b**).

The normative model based on just $\hat{p}$(correct) had one parameter per subject, whereas the alternative models had two parameters for each participant. Our data supported extending the normative model by adding a second parameter, uncertainty in the estimated mean, quantified by standard deviation, variance or Fisher information (Wilcoxon signed-rank test for deviance: $z = 4.78$, $p = 10^{-6}$ for standard deviation; $z = 4.73$, $p = 10^{-6}$; for variance; $z = 4.73$, $p = 10^{-6}$ for Fisher information). These three models were statistically indistinguishable from each other ($z < 1.7$, $p > 0.1$ for all pairwise comparisons; see Supplementary Fig. 4 for more details).

This analysis indicates that uncertainty in orientation does indeed influence confidence. To analyse this finding in more detail—and, in particular, to quantitatively examine inter-subject differences—we need to choose a particular function of uncertainty. Because standard deviation, variance and Fisher information are related by invertible transformations, it is fundamentally impossible to determine which function is used by the brain (see Supplementary Notes). Instead, we ask which quantity is the best linear predictor of confidence in an ordinal regression model.

To do that, we conducted a separate experiment in which the perceived probability of being correct played no role. We asked participants to estimate the average orientation in the sequence of Gabor patches and to rate their confidence (see the Control experiment section in the Methods). This experiment was very similar to experiment 1: on each trial, the angles of the Gabor patches were drawn from uniform distributions with one of four different variances (the same used in experiment 1). However, rather than just two possible means, the mean was randomly chosen from a uniform distribution over the whole range of orientations. Consequently, participants did not make a categorical decision, as in the previous experiment; instead, they estimated the value of the mean. Therefore, their reported confidence was not about the probability that they were correct, but about their uncertainty in the estimate of the mean. As the variance in the sequence decreased, responses were more accurate ($F(3,9) = 13.21$, $p = 10^{-5}$) and more confident ($F(3,9) = 37.4$, $p = 10^{-9}$, see Supplementary Fig. 5). We regressed confidence against single-trial estimates of either Fisher information, variance or standard deviation. These fits were significantly better when using Fisher information rather than variance (Wilcoxon signed-rank test for difference in log-likelihood, $z = 2.8$, $p = 0.005$) or standard deviation ($z = 2.9$, $p = 0.004$). These results suggest that it is reasonable to use Fisher information to quantify uncertainty. (For additional details, see Methods and Supplementary Fig. 5.)

Individual differences and their stability over time. The analysis presented so far is based on population-averaged data (Fig. 1d), so it is uninformative about differences among individuals. To determine whether, and how, $\hat{p}$(correct) and Fisher information influence confidence within subjects, we looked at the data of each individual. As expected[6,7], we observed substantial inter-individual differences (Fig. 3). Some subjects did indeed base confidence solely on $\hat{p}$(correct). However, in approximately half of them, confidence appeared to be influenced—at least to some degree—by Fisher information. To quantify this, we regressed[21] confidence reports against model-based estimates of $\hat{p}$(correct) and information. Figure 3 shows a scatter plot of the regression weights for $\hat{p}$(correct) and Fisher information. In 13 out of the 30 participants, confidence significantly reflected $\hat{p}$(correct) but not information. In 14 other participants, however, confidence significantly reflected both $\hat{p}$(correct) and information. One participant's confidence conveyed only information but not $\hat{p}$(correct), and finally, for two participants, confidence did not reflect either of the two quantities.

The ordinal regression identified seven parameters for each individual (see Methods, equation (10)): a weight for $\hat{p}$(correct), denoted β_p; a weight for information, denoted β_I; and five parameters α_j ($j = 1,\ldots,5$). The five parameters are the average log odds of observing a confidence rating greater than j; from these we selected the mid-value, α_3, which is based on splitting the confidence scale into halves. The parameter α_3 was correlated with the average confidence across the entire experiment ($r = 0.84$, $p < 10^{-8}$), and so indicates how under- or overconfident a given participant is; we thus refer to α_3 as the overall confidence. We confirmed that individual differences in these parameters (β_p, β_I and α_3) are not simply explained by how well our model fitted decisions (see Supplementary Notes). The three selected variables were uncorrelated with each other across the population ($r < 0.35$, $p > 0.1$ for all pairwise comparisons between β_p, β_I and α_3).

Finally, we note that while subjects were required to report confidence, they did not explicitly use it to, for example, regulate learning[1] or make collective decisions[3]. Thus, we know only that β_p and β_I link perceived probability of being correct and Fisher information to confidence reports, which could in principle differ from internal computations of confidence[11]. To explore this issue, we regressed reaction time against perceived probability of being correct and Fisher information, as previous studies have shown that reaction time correlates with the computation of confidence[22,23]. The regression coefficients based on reaction time were highly correlated with β_p and β_I (Supplementary Fig. 6), suggesting that confidence ratings reflected the computation of confidence.

This analysis would be no more than a model-fitting exercise if a different profile—that is, a different relationship between confidence, $\hat{p}$(correct) and Fisher information—emerged when the same participants were retested. To test for stability, in experiment 2 we retested 14 of the participants from experiment 1 approximately one month later. We observed that the three variables (β_p, β_I and α_3) were correlated across experiments (Fig. 4), indicating that this decomposition is stable across time and informative of the identity of the participants. To further validate this observation, we found that the distance in the three-dimensional space defined by (β_p, β_I and α_3)

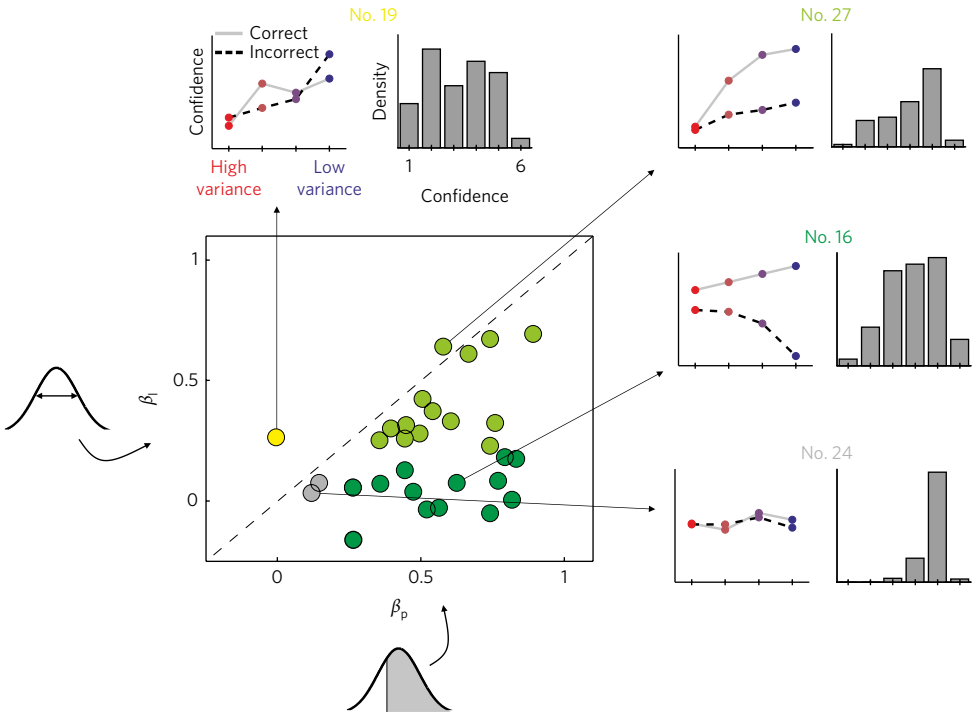

Fig. 3 | Analysis of confidence across individuals. The main panel in the lower left shows regression weights on confidence for different individuals. *x* axis: weight of the probability of being correct (β_p); *y* axis: weight of information (β_I). Each dot is a different participant, and the colour codes for significance (at the 0.05 level) as follows: dark green, only β_p was significant; light green, both β_p and β_I were significant; yellow, only β_I was significant; grey, neither was significant. Insets along the top and right margins show average confidence and confidence distributions for four representative participants. Left plots: mean confidence across different variance conditions, split by correct (solid grey line) and incorrect (dashed black line) trials. Right plots: probability distribution over confidence. For participant 19 (yellow dot), confidence reflected only information: confidence increased with variance for incorrect trials. For participant 16 (dark green dot), confidence reflected only the perceived probability of being correct: confidence in error trials decreased with increasing variance. For participant 27 (light green dot), confidence reflected a mixture of both computations. For participant 24 (grey dot), confidence was not modulated by either of these quantities. Note that there are large differences in confidence distributions, with subjects 24 and 27 showing far more confidence than subjects 16 and 19. Because α_3 is the fraction of trials with confidence larger than 3, that quantity is larger for subjects 24 and 27 than for subjects 16 and 19.

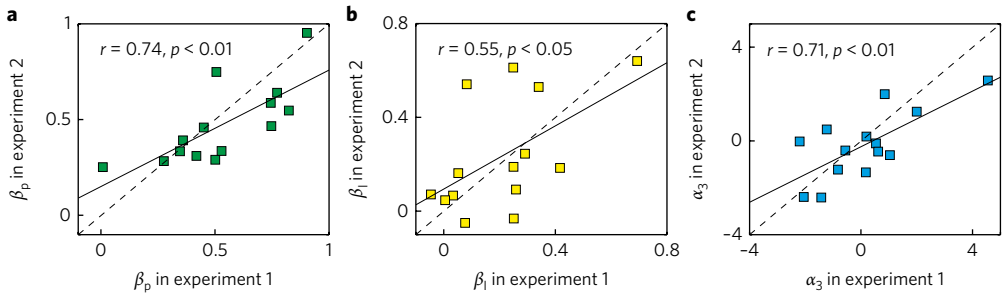

Fig. 4 | Stability across time. Fourteen participants of experiment 1 were retested approximately one month later (35.2 ± 2.4 days; range = 23–49 days). We probed stability by asking how much our three parameters (β_p, β_I and α_3) changed across experiments. **a**–**c**, Correlation across experiments for β_p (**a**), β_I (**b**) and α_3 (**c**). Each square is a different participant, the dashed line is the identity, and the value of *r* given in each box is the Pearson correlation coefficient. The three variables were significantly correlated across experiments, suggesting that this decomposition is stable across time. A non-parametric method to measure rank correlation across experiments yielded similar results (Spearman's rank correlation, $r_s = 0.82$, $p < 0.001$ for β_p, $r_s = 0.54$, $p < 0.05$ for β_I, and $r_s = 0.55$, $p < 0.05$ for α_3). A robust regression that underweights potential outliers further supported these findings (β_p: regression coefficient 0.59 ± 0.14, $p = 0.001$; β_I: regression coefficient 0.74 ± 0.27, $p = 0.02$; α_3 regression coefficient 0.60 ± 0.18, $p = 0.005$).

within participants (across the two experiments) was smaller than the distance between different participants within an experiment (Wilcoxon rank sum test, $z = 4.0$, $p < 10^{-4}$). This shows that our computational model of confidence is stable across different experimental sessions (see Discussion for comparison with previous studies).

Consistency across tasks. To determine whether subjects compute confidence the same way across tasks—that is, whether they give the same weight to $\hat{p}$ (correct) and Fisher information, and have the same overall confidence—we repeated our experiments on a cognitive task: averaging a sequence of numbers. In experiment 3, a new group of 20 participants performed, in counterbalanced order,

the visual task described above and a numerical averaging task (Fig. 5). In the numerical task, we presented two-digit numbers, updated at the same rate as in experiment 1 (4 Hz). The task was to decide whether the mean of the sequence was greater or smaller than 50. Uncertainty was manipulated in the same way as in experiment 1, using a set of variances that ensured comparable performance across tasks (see Methods).

In both tasks, accuracy increased with decreasing variance (Fig. 5a,b). A two-way rm-ANOVA with the factors 'variance' and 'task' showed a significant main effect of variance ($F(3,19)=194.3$, $p<10^{-10}$) but a non-significant effect of task ($F(1,19)=2.5$, $p=0.13$) or interaction ($F(3,19)=0.84$, $p=0.47$). Importantly, replicating experiment 1, variance did not modulate confidence in error decisions ($F(3,19)=0.2$, $p=0.89$ for the visual task; $F(3,19)=1.1$, $p=0.4$ for the numerical task). Confidence in the visual task was not statistically different from confidence in the numerical task ($F(1,19)=1.58$, $p=0.22$, Fig. 5c,d).

As in the visual task, later numbers had more influence on choice than earlier numbers ($F(5,19)=18.0$, $p=10^{-12}$) (Supplementary Fig. 1), and numbers in the high-variance condition had a smaller influence on choice than number in the low-variance condition ($F(3,19)=19.4$, $p=10^{-9}$) (Supplementary Fig. 2). We therefore used the same stochastic updating model (equations (1) and (2)) to fit the data in experiment 3. Also consistent with the visual task, decisions were better fitted by this model than the alternative model we considered in the visual task (log-likelihood of the difference against zero: $t(19)=5.2$, $p<10^{-4}$ for the cognitive task; $t(19)=6.4$, $p<10^{-5}$ for the perceptual task). We regressed confidence against $\widehat{p}$ (correct) and Fisher information, and, as in experiment 1, about half the subjects based confidence solely on $\widehat{p}$ (correct), and about half also took into account Fisher information (see Supplementary Figs. 7 and 8). We also provided independent evidence that, in the numerical task, Fisher information was more linearly predictive of confidence reports than other functions of variance (Supplementary Fig. 5).

We asked if our three regressors (β_p, β_I and α_3) were consistent across the numerical and visual tasks. The within-participants distance in the three-dimensional space was smaller than the between-participants distance (Wilcoxon rank sum test, $z=3.3$, $p<0.001$), suggesting that they were—at least in aggregate. And indeed, the weight of perceived probability of being correct, β_p, and the overall confidence, α_3, were significantly correlated across tasks ($r=0.74$, $p<0.001$ and $r=0.63$, $p<0.01$, respectively). However, the weight of Fisher information, β_I, was uncorrelated across tasks ($r=0.20$, $p=0.37$), indicating that Fisher information has quantitatively different effects on confidence in visual and numerical tasks (Fig. 6). This result is in agreement with a recent theoretical account arguing that the inverse variance is represented by domain-specific neural populations[14] (see Discussion).

Discussion

The computations underlying confidence have attracted considerable attention over the last several years, in part due to recent developments in model-based approaches[12–14] combined with neurophysiological recordings in non-human animals[24–26] and neuroimaging in humans[8–10,27]. The standard approach consists of fitting a model to the entire population and treating inter-individual variability as noise[11,15]. However, if such individual differences are robust over time, and consistent across tasks[7], then treating them as noise limits our understanding of the computational processes underlying confidence. Here we found that inter-individual differences in confidence ratings are meaningful in terms of their underlying computations. In particular, we found that different individuals used different weightings for two probabilistic quantities: their perceived probability of being correct, and their uncertainty in their estimate of the task-relevant variable[14], the latter quantified by the observed Fisher information[18,19]. We isolated the contribution of each of these two quantities to confidence, and measured, for each individual: the influence of the perceived probability of being correct on confidence (β_p); the influence of Fisher information on confidence (β_I); and the participants' overall confidence (α_3). All three variables were stable across several weeks (Fig. 4), and two of them (β_p and α_3) were stable across different tasks—one in the perceptual domain; the other in the cognitive domain (Fig. 6).

Normative theories of decision-making postulate that confidence should depend solely on the probability of being correct[12–14].

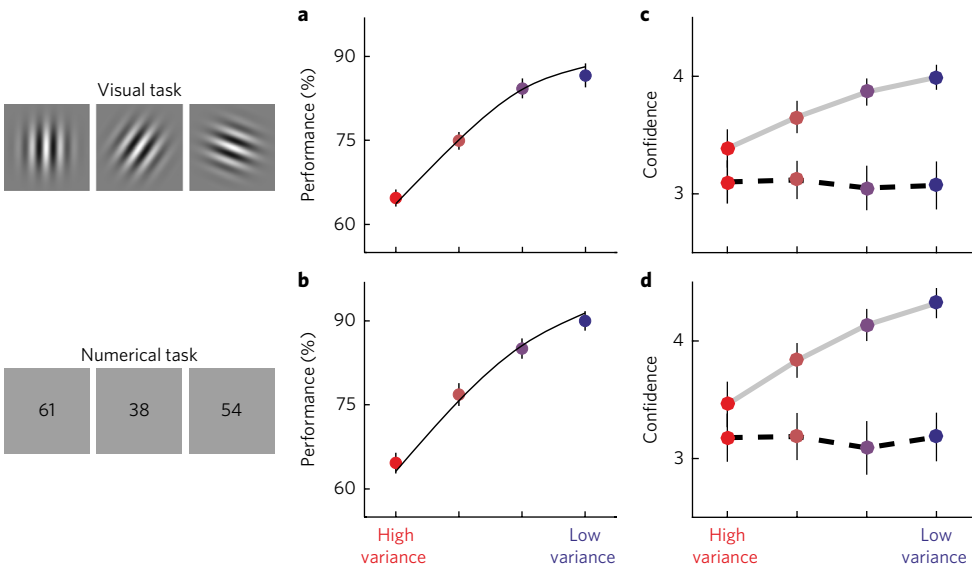

Fig. 5 | Decisions and confidence in experiment 3 ($N=20$). a, Visual task (replication of experiment 1 with different participants; panel **a** corresponds to Fig. 1c). Dots show average performance across participants and vertical lines represent s.e.m. **b**, Same as **a**, but for the numerical task. **c**, Average confidence across participants. Vertical lines depict s.e.m. (replication of experiment 1 with different participants, panel **c** corresponds to Fig. 1d). **d**, Same as **c**, but for the numerical task. The similarity between panels **a** and **b**, and between panels **c** and **d**, indicates that, at least on average, the visual and numerical tasks lead to remarkably similar behaviour, despite the fact that one is perceptual and the other is cognitive.

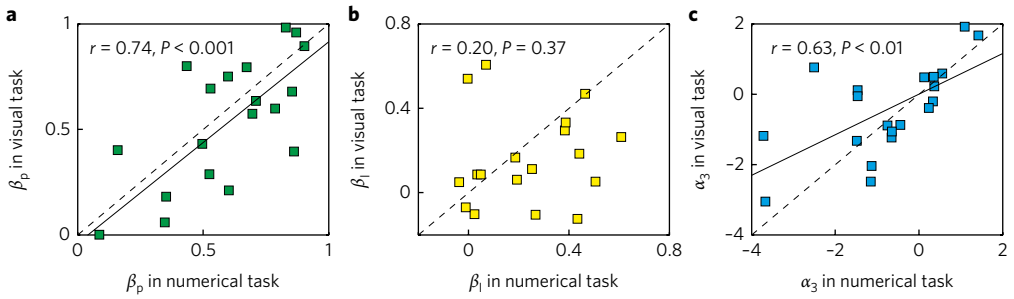

Fig. 6 | Consistency across tasks involving uncertainty in the perceptual and cognitive domain. Twenty participants that were not tested in experiments 1 or 2 performed one visual and one numerical task (experiment 3). As in Fig. 3, we decomposed confidence in terms of the weight of $\hat{p}$(correct) (β_p), the weight of information (β_I), and the overall confidence (α_3). **a–c**, Correlation across tasks for β_p (**a**), β_I (**b**) and α_3 (**c**). Each square is a different participant, the dotted line is the identity, and the value of r given in each box indicates the Pearson correlation coefficient. β_c and α_3 were positively correlated across tasks; however, the weights of Fisher information, β_I, were uncorrelated across tasks. A non-parametric method to measure the correlation across experiments yielded similar results ($r_s = 0.68$, $p < 0.01$ for β_p, $r_s = 0.22$, $p = 0.35$ for β_I, and $r_s = 0.62$, $p < 0.01$ for α_3).

We speculate that the perceived uncertainty about task-relevant variables could serve as a mental shortcut—a convenient heuristic—that provides a proxy for the probability of being correct[28]. This shortcut is reasonable, as uncertainty correlates with decision performance in our experiments (Figs. 1c and 2c). Previous research in our group showed that confidence can reflect the magnitude of sensory data[11], a choice-independent quantity that also correlates with behavioural performance. Our finding that a heuristic computation modulates confidence judgements about categorical decisions is in line with this study.

Our model of confidence assumes that subjects linearly combine the normative computation of $\hat{p}$(correct) with a function of variance. However, we cannot rule out the possibility that subjects compute $\hat{p}$(correct) suboptimally—for example, by partially basing it on the uncertainty in the task-relevant variable—and then computing confidence based solely on their suboptimal estimate of $\hat{p}$(correct). While further experiments are needed to disentangle these alternatives, we consider the former explanation to be more likely than the latter. Indeed, many studies suggest that confidence is a multivariate function that depends on factors such as the structure of the task[11], the social context[29] and post-decisional biases[17].

Previous research has shown reliable individual differences in the mean and shape of the distribution of confidence ratings[6,7], and in the extent to which confidence predicts behavioural accuracy[7,8]. These properties are believed to be idiosyncratic and correlate with individual variations in personality trait[7], brain structure[8] and resting-state functional connectivity[9]. For example, individual differences in the correlation between confidence and accuracy were systematically linked to a frontal network including the anterior prefrontal cortex, ventro-medial prefrontal cortex and rostro-lateral prefrontal cortex[8,10,30,31]. These findings were based on decisions in a wide range of contexts, including visual[8] and value-based[10] judgements. Although these studies provided interesting insights into the brain regions that correlate with individual differences in confidence, none of them explicitly asked what probabilistic quantities influence this variability.

Here, we provide empirical evidence that the idiosyncratic nature of confidence is due to differences in the computation of confidence; more specifically, different individuals place different weighting on the perceived probability of being correct and the perceived uncertainty in the estimate of the task-relevant variable. In principle, we could have used any function of variance to quantify uncertainty, and indeed all tested functions provide equally good fits in our categorical task (see Supplementary Fig. 4). We chose to model the influence of uncertainty as linear changes in Fisher information (inverse variance) only because it provided the best linear fits to confidence in a separate experiment (see Supplementary Fig. 5).

The idea that the inverse variance could modulate confidence has been previously proposed and tested in several studies[1,2,17,32,33]. In ref. [32], subjects judged the mean orientation of a set of lines, and it was found that confidence reports underweighted the stimulus variance[32]. However, whether the model parameters of that study were stable over time or consistent across domains remains unknown. In ref. [33], participants observed random-dot motion in two conditions: one with low mean and low variance, and the other one with high mean and high variance[33]. Although performance was the same for both conditions, some participants gave higher confidence ratings in one condition or the other. A model in which different subjects gave different weights to signal-to-noise ratio and inverse variance fitted these data but, critically, the fit was unstable over time (the weight of the signal-to-noise ratio was uncorrelated across a test and retest). In principle, this is at odds with our finding that the weight of $\hat{p}$(correct) was stable over time. However, we should emphasize that the signal-to-noise ratio is different from $\hat{p}$(correct): while the signal-to-noise ratio is an objective quantity that depends only on stimulus properties, $\hat{p}$(correct) is a subjective quantity that depends on the decision and how the subject learned about the stimulus (see equations (5)–(9) in Methods).

Here, instead of fitting confidence against physical properties of the stimuli, we focused on a normative theory based on the perceived (rather than the actual) probability of being correct, and explained individual differences in confidence as systematic deviations from this theory. This decomposition fitted our data better than a linear combination of the stimulus mean and variance (Supplementary Fig. 4). Our work thus provides a robust model of individual differences in confidence, with all parameters stable over time (Fig. 4). Finally, we evaluated the reliability of this computational model of confidence across domains, which suggested a relationship between specific model components and their neural encoding.

An implication of our behavioural findings is that neurons representing confidence should receive input both from populations encoding the perceived probability of being correct and from populations encoding uncertainty. Because of differences in connectivity (which are likely to arise during learning and development), different individuals should have different weightings for these two quantities; that is, different values of β_p and β_I. That is exactly what we found (Fig. 3). Furthermore, if connectivity changes slowly—a reasonable assumption in the absence of learning—β_p and β_I would be stable over time. Again, that is exactly what we found (Fig. 4).

This does not, however, explain the fact that β_p is invariant across tasks whereas β_I is not (Fig. 6a,b). For that, we need to consider how $\hat{p}$(correct) and uncertainty are encoded. Because the probability of being correct is a dimensionless quantity, and is universal

across different sources of uncertainty, it is reasonable to assume that it is encoded by a domain-general circuitry—for instance, by neurons in the prefrontal cortex[8,10,30,31]. In contrast, uncertainty—whether it is Fisher information, variance or standard deviation (see Supplementary Notes)—is a quantity with dimension, and so is likely to be encoded by domain-specific populations[14]. For example, in the case of the visual task, uncertainty could be represented by neurons in the primary visual cortex that are tuned to orientation[34]; and indeed, sensory uncertainty can be decoded from activity in the visual cortex[35]. In the same manner, numerical uncertainty could be represented by neurons in the parietal cortex tuned to different numerical quantities[36], although this has not yet been tested.

Under the assumption that the perceived probability of being correct is encoded by domain-invariant populations, the influence of this quantity on confidence should be stable across domains. This would explain our results in Fig. 6a: β_p was correlated across the visual and numerical tasks. Likewise, under the assumption that uncertainty is encoded by domain-invariant populations, the influence of this quantity on confidence should vary across domains. This would explain our results in Fig. 6b: β_I was not correlated across the visual and numerical tasks.

These are, of course, hypotheses. They do, though, make testable predictions. First, neural circuits encoding confidence should show different functional connectivity with those encoding visual versus numerical uncertainty. Second, different participants should have different relative strengths of these two forms of connectivity, co-varying with their behavioural differences. Future experiments combining behavioural data, computational modelling and neural recordings could test these predictions.

The value of investigating individual differences in human behaviour and cognition was first recognized in the psychological sciences, with a special interest in high-level aspects such as intelligence[37] and personality[38]. More recently, technical advances in magnetic resonance imaging have made it possible to develop a cognitive neuroscience of individual differences[39,40]. Findings include neural correlates of individual differences in motor behaviour[41], visual perception[42], mood[43], social network size[44] and confidence[8–10]. While these studies provide valuable insights into the neural basis of inter-individual differences in human cognition, the mechanisms responsible for such differences remain unknown. To overcome this limitation, the next challenge is to build a computational neuroscience of individual differences. A first step in this direction is to understand the computations performed by healthy adults leading to inter-individual variability in behaviour. Our study provides a computational model of consistent individual differences in confidence, paving the way towards determining how these computations change under development[45], aging[46] and psychiatric disorders[47].

Methods

Participants. Sixty healthy adults (aged 18–45, 43 right-handed, 31 female) with normal or corrected-to-normal vision participated in this study. All participants were recruited through advertisement at University College London, and gave written informed consent. We collected data from 94 experimental sessions lasting approximately 90 min each. Participants were paid £10 per hour. All experimental procedures were approved by the research ethics committee at University College London.

Display. Stimuli were generated using the Cogent Toolbox (http://www.vislab.ucl.ac.uk/cogent.php) for MATLAB (Mathworks Inc.). Participants observed an LCD display (21-inch monitor; refresh rate: 60 Hz; resolution: 1,024 × 768 pixels) at a viewing distance of approximately 60 cm.

Experiment 1: visual task. Thirty participants performed experiment 1, which consisted of an orientation-averaging task (Fig. 1). Observers viewed a sequence of 30 tilted Gabor patches over a middle-grey background (standard deviation of the Gaussian envelope: 0.63 deg; spatial frequency: 1.57 cycles deg^{-1}; contrast: 25%) flashed in rapid succession at the centre of the screen. Each patch was presented for 200 ms with an inter-stimulus interval of 50 ms, resulting in an update rate of 4 Hz. Once the sequence finished, participants were asked to judge whether the mean orientation of the patches was tilted clockwise or anticlockwise relative to the vertical. The response alternatives consisted of two tilted lines presented in the left and right visual field (size: 2.2 deg, location: 11.3 deg left or right to the centre of the screen). The position of the response alternatives was randomly assigned and counterbalanced across trials. To select the option displayed in the left, participants pressed the 'Q' button of a QWERTY keyboard using the left hand; to select the option on the right, they pressed the 'P' button. Participants were then asked to report their confidence on a rating scale from 1 to 6. A horizontal line was presented at the centre of the screen (length: 18.9 deg) with six equally spaced marks signalling different levels of confidence. Participants moved a cursor to the left or right of the scale by pressing the 'Q' or 'P' buttons respectively. The initial point in the scale was randomly chosen on a trial-by-trial basis. Once the participants selected a confidence rating, they pressed the space bar to continue. After an inter-trial interval (which was uniformly distributed between 0.7 and 0.9 s), a new trial began.

The orientations of the patches were drawn from uniform distributions with mean m and endpoints $m \pm v$. We used distributions with two different means ($m = +3$ or -3 degrees) and four different variances (given by their different endpoints: $v = 10$, 14, 24 or 45 degrees). Uniform distributions were pseudo-randomly sampled such that the mean was exactly ± 3 degrees on every trial. This generated weak correlations, but multi-collinearity analyses indicated that presentations could not be predicted from previous samples ($R^2 < 0.07$). Orientations were randomly shuffled to define the presentation order. The experiment consisted of 400 trials: 50 trials for each of the 8 distributions. Blocked feedback was given every 20 trials by a message displaying the number of correct trials in that block. Each block comprised five trials of each variance condition presented in random order. Therefore, performance for different variance conditions could not be learned from feedback.

Experiment 2: stability across time. All participants of experiment 1 were invited to perform the visual task a second time, approximately one month later. Fourteen participants accepted the invitation and were re-tested. Experiment 2 was performed 35.2 ± 2.4 days after experiment 1 (range: 23–49 days). Experimenters were blind to the results of experiment 1 when testing participants in experiment 2.

Experiment 3: stability across the perceptual and cognitive domain. Twenty healthy adults who did not participate in experiment 1 or 2 performed experiment 3. Participants performed two sessions: the visual task described in experiment 1 and a numerical averaging task. Half of the participants performed the visual task first. The second session was performed 9.7 ± 2.9 days (range: 1–27 days) after the first one. Experimenters were blind to the results of the first session when testing the participants in the second session.

The numerical task was identical in structure to the visual task but, instead of Gabor patches, two-digit numbers (size: 3.8 deg; font: Arial) were presented. The colour of the numbers (black or white over a middle-grey background) was randomly chosen at each presentation. Participants were instructed to decide whether the mean of the sequence was greater or smaller than 50. Numbers were sampled from uniform distributions with mean $m = 47$ or $m = 53$, and endpoints $m \pm v$ were defined by $v = 7$, 9, 11 or 33. These values were chosen, through pilot experiments with a different set of participants, to obtain performances similar to that observed in experiment 1. Uniform distributions were pseudo-randomly sampled such that the mean of the sequence was exactly m on each trial. We performed the same multi-collinearity analysis of experiment 1, and found that presentations could not be predicted from previous samples ($R^2 < 0.06$). Decisions were collected in the same way as in experiment 1: a response screen with two options ('smaller' and 'greater') was presented on both sides of the visual field. Participants gave their answer, and indicated confidence, using the same keys as in the visual task.

Control experiment. Ten healthy adults (aged 20–45, 6 female, all right-handed) who had not participated in experiment 1, 2 or 3 participated in the control experiment. The experiment consisted of one visual and one numerical task that subjects performed in a single session of approximately 90 min. Half of the participants performed the visual task first. Participants observed a sequence of items serially flashed at the fovea at 4 Hz, and were asked to provide their analogue estimate of the mean. To rate their confidence, participants moved a cursor over a continuous horizontal line. All other parameters (length of the sequence, colour, contrast, brightness, viewing distance and so on) were identical to our main study.

In the visual task, participants observed tilted Gabor patches. The mean of the distribution was uniformly sampled across the entire circle. After observing 30 items, we presented a line in the centre of the screen, initialized at a random orientation. Participants then moved the mouse horizontally to change its orientation until they matched the perceived mean in the sequence. In the numerical task, participants observed two-digit numbers. We uniformly sampled the mean between 44 and 66 (to ensure that all numbers were between 11 and 99 in the condition with higher variance). Participants typed their answer using a keyboard.

Model fitting. To fit the stochastic updating model (equations (1) and (2)) to the participants' decisions, we find, for each individual, the parameters λ and γ that maximize the log-likelihood,

$$\log L(\lambda,\gamma)=\sum_{k=1}^{N_{tr}}\frac{1+d_k}{2}\log\Phi\left(\frac{\overline{\mu}_{30,k}(\lambda)}{\sigma_{30,k}(\lambda,\gamma)}\right)+\frac{1-d_k}{2}\log\left[1-\Phi\left(\frac{\overline{\mu}_{30,k}(\lambda)}{\sigma_{30,k}(\lambda,\gamma)}\right)\right] \quad (3)$$

where Φ is the standard cumulative normal function, d_k is the decision on trial k (+1 if clockwise, −1 if anticlockwise), $\sigma_{30,k}(\lambda,\gamma)$ is obtained from equation (2), N_{tr} is the number of trials and

$$\overline{\mu}_{30,k}(\lambda)=\lambda\sum_{i=1}^{30}(1-\lambda)^{30-i}\theta_{i,k} \quad (4)$$

is the mean value of μ_{30} on trial k. (A minor technical point: equation (4) describes the visual task; the cognitive task is the same except that the mean is offset by 50.)

Estimating the Fisher information and the perceived probability of being correct. On the basis of the best fitting parameters λ and γ derived from the stochastic updating model (the values of λ and γ that maximize $L(\lambda,\gamma)$ in equation (2)), we estimated, on a trial-by-trial basis, the observed Fisher information and the expected perceived probability of being correct. The observed Fisher information is just the inverse variance of the participants' estimate, the latter computed via equation (2) (Fig. 2a). The expected perceived probability of having made a correct decision, d, is given by

$$\hat{p}\left(\text{correct}\,\middle|\,\overline{\mu}_{30},\sigma_{30},d\right)=\int_{-\infty}^{+\infty}d\mu_{30}\hat{p}\left(\text{correct}\,\middle|\,\mu_{30},\sigma_{30}\right)p\left(\mu_{30}\,\middle|\,\overline{\mu}_{30},\sigma_{30},d\right) \quad (5)$$

The first term inside the integral, $\hat{p}\left(\text{correct}\,\middle|\,\mu_{30},\sigma_{30}\right)$, is the shaded area under the Gaussian in Fig. 2a; consequently, it is given by the cumulative normal distribution,

$$\hat{p}\left(\text{correct}|\mu_{30},\sigma_{30}\right)=\Phi\left(\frac{|\mu_{30}|}{\sigma_{30}}\right). \quad (6)$$

The second term in the integral, $p\left(\mu_{30}\,\middle|\,\overline{\mu}_{30},\sigma_{30},d\right)$, is the probability of observing μ_{30} given $\overline{\mu}_{30}$, σ_{30} and, importantly, the decision, d. If the decision is clockwise ($d=+1$), μ_{30} must be positive, whereas if the decision is anticlockwise ($d=-1$), μ_{30} must be negative. We can take these constraints into account using the Heaviside step function, $\Theta(x)$ (which is 1 if $x>0$ and 0 otherwise), yielding

$$p\left(\mu_{30}\,\middle|\,\overline{\mu}_{30},\sigma_{30}\right)=\frac{1}{Z}\frac{e^{\frac{-\left(\mu_{30}-\overline{\mu}_{30}\right)^2}{2\sigma_{30}^2}}}{\sqrt{2\pi\sigma_{30}^2}}\Theta\left(\mu_{30}d\right) \quad (7)$$

where Z is the normalization constant,

$$Z=\int_{-\infty}^{+\infty}d\mu_{30}\Theta\left(\mu_{30}d\right)\frac{e^{\frac{-\left(\mu_{30}-\overline{\mu}_{30}\right)^2}{2\sigma_{30}^2}}}{\sqrt{2\pi\sigma_{30}^2}}=\Phi\left(\frac{\overline{\mu}_{30}d}{\sigma_{30}}\right). \quad (8)$$

Combining these two expressions, we have

$$\hat{p}\left(\text{correct}\,\middle|\,\overline{\mu}_{30},\sigma_{30},d\right)=\frac{1}{Z}\int_{-\infty}^{+\infty}d\mu_{30}\frac{e^{\frac{-\left(\mu_{30}-\overline{\mu}_{30}\right)^2}{2\sigma_{30}^2}}}{\sqrt{2\pi\sigma_{30}^2}}\Theta\left(\mu_{30}d\right)\Phi\left(\frac{|\mu_{30}|}{\sigma_{30}}\right). \quad (9)$$

On each trial, $\hat{p}\left(\text{correct}\,\middle|\,\overline{\mu}_{30},\sigma_{30},d\right)$ was computed numerically using Matlab. Note that the expected perceived probability of being correct (equation (9)) is dependent on the decision, d, whereas the Fisher information (equation (2); Fig.2a) does not depend on d, and so is choice-independent.

Ordinal regression of confidence reports. We ran for each individual a multivariate ordinal regression[21]. For each of the five possible splits in the rating scale, this regression fits a logistic model with fixed effects and different offsets,

$$\log\left(\frac{p(c>j)}{1-p(c>j)}\right)=-\alpha_j+\beta_p Z_p+\beta_I Z_I \quad (10)$$

where $1\le j\le 5$, c denotes confidence, and Z_p and Z_I are z-scored estimates of the perceived probability of being correct and Fisher information on each trial. The outputs of this regression are the offsets $\alpha_1,\ldots,\alpha_5$, and the weights β_p and β_i. To summarize the computations underlying confidence, we selected α_3 (the offset when splitting the scale into halves, which we refer to as the overall confidence), β_p (the weight of the probability of being correct on confidence) and β_i (the weight of information on confidence).

Statistical analyses. In experiment 1, we computed the average performance for each variance condition and each participant. These values were submitted to a repeated-measures one-way analysis of variance (rm-ANOVA) with factor 'variance condition' (4 levels) and 'participant' (30 levels) as repeated measure (Fig. 1). The normality assumption of this test was checked using the Lilliefors test ($k=0.7$, $c=0.8$, $p=0.07$). We also computed the average confidence rating for each variance condition and each participant, conditioned on correct or incorrect trials, and submitted those values to a two-way rm-ANOVA with the factors 'variance condition' (4 levels), 'outcome' (2 levels: correct or incorrect) and 'participant' (30 levels) as repeated measures (Fig. 2c). The normality assumption of this test was checked using the Lilliefors test ($k=0.04$, $c=0.06$, $p>0.5$). The goodness of fit for each model and subject (Supplementary Fig. 1b), quantified by the negative log-likelihood (equation (3)), was submitted to a two-sided paired t-test (29 degrees of freedom). The normality assumption of this test was checked using the Lilliefors test ($k=0.08$, $c=0.11$, $p>0.5$).

In experiment 2, we compared the within-participants distances in the space defined by $(\beta_p,\beta_i,\alpha_3)$ with the between-subjects distances. Because we have 14 participants, this defines 14 within-subjects distances and $14\times13/2=91$ between-subjects distances. We z-scored each dimension and used the Euclidean metric to compute distance. The Lilliefors test rejected the null hypothesis that these values were normal ($k=0.1$, $c=0.08$, $p=0.01$); therefore, we used a non-parametric test, the Wilcoxon ranked sum test. This test is unpaired and the reported p value is two-sided.

In experiment 3, we computed the average performance for each variance condition, task and participant (Fig. 5a,b). We submitted these values to a two-way rm-ANOVA with the factors 'variance condition' (4 levels), 'task' (2 levels) and 'participants' (20 levels) as repeated measures. The normality assumption of this test was checked using the Lilliefors test ($k=0.07$, $c=0.09$, $p=0.36$). We computed the average confidence rating across all conditions and participants and performed the same rm-ANOVA used in experiment 1 (Fig. 5c,d). As in experiment 1, average confidence was normally distributed (Lilliefors test, $k=0.06$, $c=0.07$, $p=0.17$). To evaluate the stability of $(\beta_p,\beta_i,\alpha_3)$ across domains, we computed the within- and between-subjects distances following the same procedure of experiment 2, and compared these values using the same non-parametric test.

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

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

Received: 28 July 2017; Accepted: 29 August 2017;
Published online: 25 September 2017

References

1. Meyniel, F., Schlunegger, D. & Dehaene, S. The sense of confidence during probabilistic learning: a normative account. *PLoS Comput. Biol.* **11**, e1004305 (2015).
2. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Phil. Trans. R. Soc. Lond. B* **367**, 1310–1321 (2012).
3. Bahrami, B. et al. Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
4. Bahrami, B. et al. What failure in collective decision-making tells us about metacognition. *Phil. Trans. R. Soc. Lond. B* **367**, 1350–1365 (2012).
5. Tetlock, P. in *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press, Princeton, New Jersey, 2005).
6. Graziano, M. & Sigman, M. The spatial and temporal construction of confidence in the visual scene. *PLoS ONE* **4**, e4909 (2009).
7. Ais, J., Zylberberg, A., Barttfeld, P. & Sigman, M. Individual consistency in the accuracy and distribution of confidence judgments. *Cognition* **146**, 377–386 (2016).
8. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541–1543 (2010).
9. Barttfeld, P. et al. Distinct patterns of functional brain connectivity correlate with objective performance and subjective beliefs. *Proc. Natl Acad. Sci. USA* **110**, 11577–11582 (2013).
10. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
11. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput. Biol.* **11**, e1004519 (2015).
12. Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Phil. Trans. R. Soc. Lond. B* **367**, 1322–1337 (2012).
13. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian probability: from neural origins to behavior. *Neuron* **88**, 78–92 (2015).

14. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
15. Sanders, J., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
16. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
17. Navajas, J., Bahrami, B. & Latham, P. E. Post-decisional accounts of biases in confidence. *Curr. Opin Behav. Sci.* **11**, 55–60 (2016).
18. Pouget, A., Deneve, S. & Latham, P. E. in *Visual Attention and Cortical Circuits* (eds Braun, J., Koch, C. & Davis, J.) 265–283 (MIT Press, Cambridge, 2001).
19. Moreno-Bote, R. et al. Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
20. Hangya, B., Sanders, J. I. & Kepecs, A. A mathematical framework for statistical decision confidence. *Neural Comput.* **28**, 1840–1858 (2016).
21. McCullagh, P. Regression models for ordinal data. *J. Roy. Statist. Soc. B* **42**, 109–142 (1980).
22. Petrusic, W. M. & Baranski, J. V. Judging confidence influences decision processing in comparative judgments. *Psychon. Bull. Rev* **10**, 177–183 (2003).
23. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864 (2010).
24. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
25. Lak, A. et al. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
26. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
27. van den Berg, R. et al. A common mechanism underlies changes of mind about decisions and confidence. *Elife* **5**, e12192 (2016).
28. Barthelme, S. & Mamassian, P. Flexible mechanisms underlie the evaluation of visual confidence. *Proc. Natl Acad. Sci. USA* **107**, 20834–20839 (2010).
29. Pescetelli, N., Rees, G. & Bahrami, B. The perceptual and social components of metacognition. *J. Exp. Psychol. Gen.* **145**, 949 (2016).
30. Fleming, S. M., Dolan, R. J. & Frith, C. D. Metacognition: computation, biology and function. *Phil. Trans. R. Soc. Lond. B* **367**, 1280–1286 (2012).
31. Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811–2822 (2014).
32. Zylberberg, A., Roelfsema, P. R. & Sigman, M. Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious. Cogn.* **27**, 246–253 (2014).
33. de Gardelle, V. & Mamassian, P. Weighting mean and variability during confidence judgments. *PLoS ONE* **10**, e0120870 (2015).
34. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
35. van Bergen, R. S., Ma, W. J., Pratte, M. S. & Jehee, J. F. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728 (2015).
36. Nieder, A. & Dehaene, S. Representation of number in the brain. *Annu. Rev. Neurosci.* **32**, 185–208 (2009).
37. Neisser, U. et al. Intelligence: knowns and unknowns. *Am. Psychol.* **51**, 77–101 (1996).
38. Goldberg, L. R. The structure of phenotypic personality traits. *Am. Psychol.* **48**, 26–34 (1993).
39. Kanai, R. & Rees, G. The structural basis of inter-individual differences in human behaviour and cognition. *Nat. Rev. Neurosci.* **12**, 231–242 (2011).
40. Dubois, J. & Adolphs, R. Building a science of individual differences from fMRI. *Trends Cogn. Sci.* **20**, 425–443 (2016).
41. van Gaal, S., Scholte, H. S., Lamme, V. A., Fahrenfort, J. J. & Ridderinkhof, K. R. Pre-SMA gray-matter density predicts individual differences in action selection in the face of conscious and unconscious response conflict. *J. Cogn. Neurosci.* **23**, 382–390 (2011).
42. Schwarzkopf, D. S., Song, C. & Rees, G. The surface area of human V1 predicts the subjective experience of object size. *Nat. Neurosci.* **14**, 28–30 (2011).
43. Smith, S. M. et al. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* **18**, 1565–1567 (2015).
44. Kanai, R., Bahrami, B., Roylance, R. & Rees, G. Online social network size is reflected in human brain structure. *Proc. Biol. Sci.* **279**, 1327–1334 (2012).
45. Weil, L. G. et al. The development of metacognitive ability in adolescence. *Conscious. Cogn.* **22**, 264–271 (2013).
46. Palmer, E. C., David, A. S. & Fleming, S. M. Effects of age on metacognitive efficiency. *Conscious. Cogn.* **28**, 151–160 (2014).
47. David, A. S., Bedford, N., Wiffen, B. & Gilleen, J. Failures of metacognition and lack of insight in neuropsychiatric disorders. *Phil. Trans. R. Soc. Lond. B* **367**, 1379–1390 (2012).

Acknowledgements

J.N. and B.B. were supported by the European Research Council StG (NEUROCODEC, no. 309865); C.H. was supported by a studentship from the Medical Research Council (UK); H.F. was supported by a Chevening scholarship; M.K. and P.E.L. were supported by the Gatsby Charitable Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.N. and B.B. designed the experiments. J.N., C.H. and H.F. conducted the experiments. J.N., M.K., P.E.L. and B.B. developed the analysis approach and computational models. J.N. analysed the data. J.N., P.E.L. and B.B. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-017-0215-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.N.

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

nature research

Corresponding author(s): Joaquin Navajas

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   The sample size used in Experiment 1 (n=30) is based on previous studies investigating individual differences in human confidence (ref.7-9). The sample size in Experiment 2 (n=14) is similar to the one used in previous studies testing the reliability of confidence ratings over time (ref. 6-7). The sample size in Experiment 3 (n=20 performing 40 sessions) was based on the inter-individual variability observed in Experiment 1.

2. Data exclusions

   Describe any data exclusions.

   No data were excluded from the analysis.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   Experiment 2 is a retest of the same participants of Experiment 1. The visual session of Experiment 3 replicated Experiment 1 with different participants.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   There was no comparison between different experimental groups or treatments. A within-subjects measure was obtained for each participant, and its reliability was tested over time and across tasks.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Experiment 2: Experimenters were blind to the results of Experiment 1 when re-testing participants in Experiment 2.
   Experiment 3: They were also blind to the results of the first session when testing the same participants in the second session.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	These data was analyzed using Matlab R2016a (Mathworks)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No Eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No Eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No Eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Participants were 60 healthy adults (aged 18-45, 43 right-handed, 31 female) with normal or corrected-to-normal vision and no history of neurological disease. All participants were recruited through advertisement at University College London.