Title:

Human noise blindness drives suboptimal cognitive inference

Authors:

Santiago Herce Castañón[1,*], Dan Bang[2,3], Rani Moran[2,4], Jacqueline Ding[1], Tobias Egner[5,6], Christopher Summerfield[1]

Affiliations:

[1]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK.
[2]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, UK.
[3]Interacting Minds Centre, Aarhus University, 8000 Aarhus, Denmark.
[4]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, UK
[5]Center for Cognitive Neuroscience, Duke University, Durham, North Carolina 27710, USA
[6]Department of Psychology and Neuroscience, Duke University, Durham, North Carolina 27708-0086, USA
[*]Correspondence: santiagohc@gmail.com

Abstract

Humans typically make near-optimal sensorimotor judgements but show systematic biases when making more cognitive judgements. Here we test the hypothesis that, while humans are sensitive to the noise present during early sensory processing, the "optimality gap" arises because they are blind to noise introduced by later cognitive integration of variable or discordant information. In six psychophysical experiments, human observers judged the average orientation of an array of contrast gratings. We varied the stimulus contrast (encoding noise) and orientation variability (integration noise) of the array. Participants adapted optimally to changes in encoding noise, but, under increased integration noise, they displayed a range of suboptimal behaviours: they ignored stimulus base rates, reported excessive confidence in their choices, and refrained from opting out of objectively difficult trials. These overconfident behaviours were captured by a Bayesian model which is blind to integration noise. Our study provides a computationally grounded explanation of suboptimal cognitive inferences.

The question of whether humans make optimal choices has received considerable attention in the neural, cognitive and behavioural sciences. On one hand, the general consensus in sensory psychophysics and sensorimotor neuroscience is that choices are optimal. For example, humans have been shown to combine different sources of stimulus information in a statistically optimal manner, weighting each source by its reliability (Ernst & Banks, 2002; Knill, Kersten, & Yuille, 1996; Körding & Wolpert, 2006; Ma, Beck, Latham, & Pouget, 2006; Mamassian, Landy, & Maloney, 2002; Trommershäuser, Maloney, & Landy, 2008). Humans have also been shown to optimally utilise knowledge about stimulus base rates to resolve stimulus ambiguity (Kersten, Mamassian, & Yuille, 2004; Körding & Wolpert, 2004; O'Reilly, Jbabdi, Rushworth, & Behrens, 2013; Sun & Perona, 1998; Vilares, Howard, Fernandes, Gottfried, & Kording, 2012).

On the other hand, psychologists and behavioural economists, studying more cognitive judgements, have argued that human choices are suboptimal (Tversky & Kahneman, 1974). For example, when required to guess a person's occupation, humans neglect the base rate of different professions and solely rely on the person's description provided by the experimenter (Kahneman & Tversky, 1973). Such suboptimality has been attributed to insufficient past experience (Hertwig & Erev, 2009), limited stakes in laboratory settings (Levitt & List, 2007), the format in which problems are posed (Jarvstad, Hahn, Rushton, & Warren, 2013), distortions in representations of values and probabilities (Ackermann & Landy, 2014), and/or a reluctance to employ costly cognitive resources (Gershman, Horvitz, & Tenenbaum, 2015; Kahneman, 2011). However, an account of human decision-making that can explain both perceptual optimality and cognitive suboptimality has yet to emerge (Summerfield & Tsetsos, 2015).

Here we propose that resolving this apparent paradox requires recognizing that perceptual and cognitive choices often are corrupted by different sources of noise. More specifically, choices in perceptual and cognitive tasks tend to be corrupted by noise which arises at different stages of the information processing leading up to a choice (Faisal & Wolpert, 2009; Hunt, 2014; Juslin & Olsson, 1997; Ma & Jazayeri, 2014). In perceptual tasks, experimenters typically manipulate early sources of noise. For example, they may vary the contrast of a grating, or the net motion energy in a random dot kinematogram, which affects the signal-to-noise ratio of the encoded stimulus. Conversely, in cognitive tasks, which often involve written materials or clearly perceptible stimuli, experimenters typically seek to manipulate noise arising after stimulus encoding. For example, they may vary the discrepancy between different pieces of information bearing on a choice, such as the relative costs and benefits of a consumer product (Kahneman, 2011). Such judgements are difficult because they require integration of multiple, sometimes highly discordant, pieces of information within a limited-capacity system (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Eriksen & Eriksen, 1974; MacLeod, 1991).

Here we test the hypothesis that, while humans are sensitive to noise arising during early sensory encoding, they are blind to the additional noise introduced by their own cognitive system when integrating variable or discordant pieces of information. Such blindness to "integration noise" may explain a range of suboptimal behaviours found in more cognitive tasks such as base-rate neglect and overconfidence. We tested this explanation of the gap in optimality between perceptual and cognitive judgements using a psychophysical task which required observers to judge the average tilt of an array of gratings and which allowed us to independently manipulate early and late sources of noise. First, we manipulated encoding noise (i.e. the perceptual difficulty of encoding an individual piece of information) by varying the contrast of individual gratings, with decisions being harder for low-contrast stimuli. Second, we manipulated integration noise (i.e. the

cognitive difficulty of integrating multiple pieces of information) by varying the discrepancy of the orientations of individual gratings, with decisions being harder for high-variability stimuli. To pre-empt our results, we show that, while observers adapt optimally to increases in encoding noise, they fail to adapt to increases in integration noise. We argue that such "noise blindness" is a major driver of human suboptimal inference and may explain the gap in optimality between perceptual and cognitive judgements.

Results

Experimental dissociation of encoding noise and integration noise

All six experiments were based on the same psychophysical task. On each trial, participants were presented with eight tilted gratings organized in a circular array. Participants were required to categorise the average orientation of the stimulus array as oriented clockwise (CW) or counter-clockwise (CCW) from the horizontal axis (**Fig. 1A-B**). After having made a response, participants received categorical feedback about choice accuracy, before continuing to the next trial. We manipulated two features of the stimulus array to dissociate encoding noise and integration noise: the *contrast* of the gratings (root mean square contrast, *rmc*: {0.15, 0.6}), which affects encoding noise, and the *variability* of the gratings' orientation (standard deviation of orientations, *std*: {0°, 4°, 10°}), which affects integration noise.

In Experiments 1 ($n = 20$) and 2 ($n = 20$), we assessed the effects of contrast and variability on choice accuracy and evaluated participants' awareness of these effects. In both experiments, at the beginning of each trial, we provided a "prior" cue which, on half of the trials, signalled the correct stimulus category with 75% probability (henceforth "biased" trials), and, on the other half of trials, provided no information about the stimulus category most likely to be correct (henceforth "neutral" trials) (**Fig. 1B**). The neutral trials provided us with a baseline measure of participants' choice accuracy in the different conditions of our factorial design, and the biased trials allowed us to assess the degree to which – if at all – participants compensated for reduced choice accuracy in a given experimental condition by relying more on the prior cue. To provide additional insight into participants' awareness of their own performance, in Experiment 2, we asked participants to report their confidence in each choice (i.e. the probability that the choice is correct; **Fig. 1C**).

Matched performance for different levels of encoding and integration noise

We first used the neutral trials to benchmark the effects of contrast and variability on choice accuracy. As intended, choice accuracy decreased with lower contrast (Exp1: $F(1,19) = 15.54$, $p < .001$; Exp2: $F(1,19) = 41.08$, $p < .001$; collapsed: $F(1,39) = 49.3$, $p < .001$) and with higher variability (Exp1: $F(1.3,24.7) = 8.51$, $p < .001$; Exp2: $F(1.6,32.2) = 26.0$, $p < .001$; collapsed: $F(1.4,57.3) = 30.61$, $p < .001$). Our factorial design contained three critical conditions which allowed us to compare participants' behaviour under distinct sources of noise: (i) "baseline", (ii) "low-c" and (iii) "high-v". In the baseline condition, the total amount of noise is lowest (high contrast, .6; zero variability, 0º). In the low-c condition (low contrast, .15; zero variability, 0º), encoding noise is high but integration noise is low. Conversely, in the high-v condition, integration noise is high but encoding noise is low (high contrast, 0.6; high variability, 10º). As expected, choice accuracy was reduced both in the low-c and in the high-v conditions (about 12%) compared to the baseline condition (baseline>low-c: $t(39) = 9.24$, $p < .001$; baseline>high-v: $t(39) = 9.70$, $p < .001$). Critically, choice accuracy was at statistically similar levels in the low-c and the high-v

conditions, (Exp1, high-v>low-c: $t(19) = 0.36$, $p > 0.7$; Exp2, high-v>low-c: $t(19) = 0.11$, $p > 0.9$; collapsed, high-v>low-c: $t(39) = 0.34$, $p > 0.7$; **Fig. 2A**).

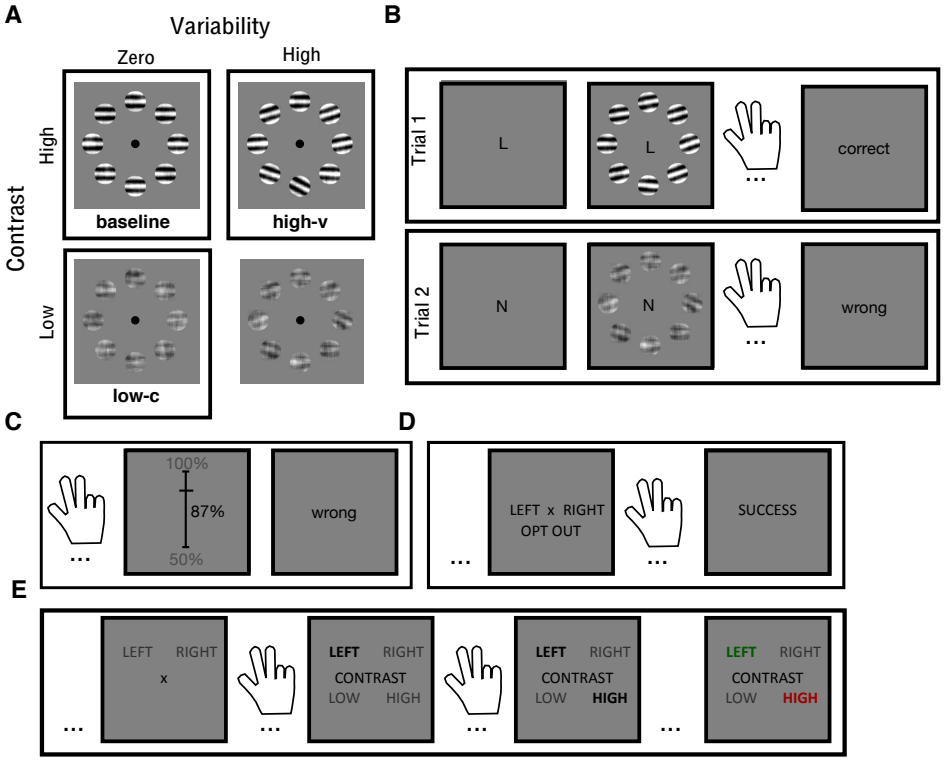

Fig. 1. Experimental paradigm. (**A**) We manipulated the stimulus contrast and the orientation variability of an array of eight gratings in a factorial manner. (**B**) Participants categorized the average orientation of the gratings as clockwise (CW, "left") or counter-clockwise (CCW, "right") relative to horizontal. A cue, which was shown at the start of each trial and remained on the screen until a response had been made, indicated the prior probability of occurrence of each stimulus category (L: 25% CW, 75% CCW; N: 50% CW, 50% CCW; R: 75% CW, 25% CCW). Participants received categorical feedback about choice accuracy, before continuing to the next trial. Feedback was based on the average orientation of the gratings shown. (**C**) In Experiment 2, after having made a choice, participants were required to estimate the probability that the choice was correct by sliding a marker along a scale (50% to 100% in increments of 1%). (**D**) In Experiment 3, participants could opt out of making a choice and receive the same reward as a correct choice with a 75% probability. (**E**) In Experiment 4, after having made a choice, participants were required to categorise (low versus high) either the contrast or the variability of the stimulus array.

Do people utilise the prior cue to compensate for increased errors?

We next leveraged the biased trials to assess the degree to which participants adapted to the changes in choice accuracy induced by our factorial design. Given the pattern of performance described above, we would expect participants to rely more on the prior cue in the low-c and high-v conditions than in the baseline condition. To test this prediction, we applied Signal Detection Theory (Macmillan & Creelman, 2004; Stanislaw & Todorov, 1999) to finely measure the degree to which participants shifted their decision criterion in accordance with the prior cue (**Fig. 2C-D**). Greater values of this "bias index" indicate that the prior cue has a greater influence on choices. Consistent with an ideal observer framework, participants used the prior cue more in the low-c than in the baseline condition ($t(39) = 4.89$, $p < .001$). However, contrary to an ideal observer

framework, participants used the prior cue less in the high-v than in the baseline condition ($t(39) = 2.85$, $p < .01$; compare psychometric curves in **Fig. 2B**). In line with these results, a full factorial analysis of the bias index identified a positive main effect of contrast ($F(1,39) = 24.02$, $p < .001$) and a negative main effect of variability ($F(1.9,37.1) = 9.9$, $p < .001$; **Fig. 2D**). Finally, including both neutral and biased trials, we used trial-by-trial logistic regression to test how contrast (c) and variability (v) mediated the influence of the prior cue and sensory evidence (μθ) on choices (μθ, cue, μθ*c, μθ*v, cue*c, cue*v; **Fig. 2E**). The prior cue had a larger influence on choices on low-contrast compared to high-contrast trials ($t(39) = 4.05$, $p < .001$) and on low-variability compared to high-variability trials ($t(39) = 5.21$, $p < .001$). Overall, these results show that participants failed to adapt to noise arising during integration of discordant pieces of information.

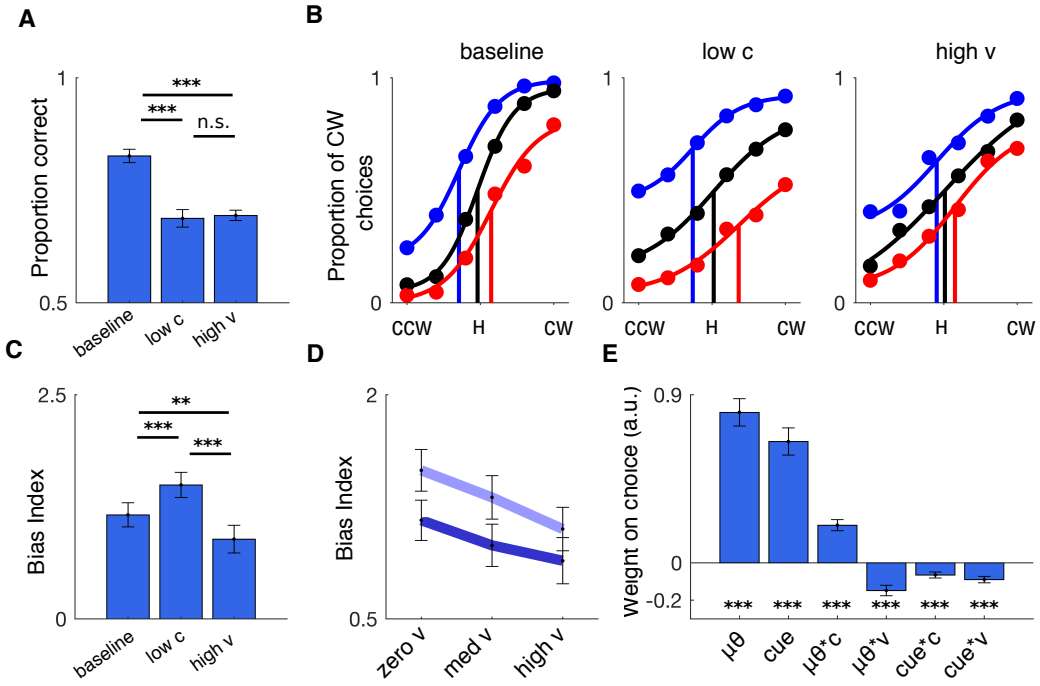

Fig. 2. Effects of contrast and variability on choice behaviour. (**A**) Choice accuracy for the baseline, reduced contrast (low-c) and increased variability (high-v) conditions. (**B**) Psychometric curves are shallower in the low-c and high-v conditions compared to baseline. Choices shift towards the cued category on biased trials (blue: 75% CW; red: 25% CW) compared to neutral trials (black) but least so in the high-v condition. Vertical lines mark the inflection points of psychometric functions fitted to the data. See Methods for construction of psychometric functions. (**C**) Bias index, a measure of cue usage, is higher in the low-c condition but lower in the high-v condition compared to baseline. See Methods for construction of the bias index. (**D**) Factorial analysis of the effects of contrast and variability on the bias index shows an increase with contrast (dark blue: high contrast; pale blue: low contrast) but a decrease with variability. (**E**) Trial-by-trial influence of prior cue on choices as measured through logistic regression. c: contrast; v: variability; μθ: signed mean orientation; cue: signed prior cue. (**A-E**) Data is represented as group mean ± SEM. * $p < .05$, ** $p < .01$, *** $p < .001$. For panel A, only neutral trials were used. For panel B and E, both neutral and biased trials. For panels C and D, only biased trials were used.

Are people blind to integration noise?

To test whether participants failed to adapt because they were "blind" to integration noise, we analysed the confidence reports elicited in Experiment 2 (**Fig. 1C**). We implemented a strictly

proper scoring rule such that it was in participants' best interest (i) to make as many accurate choices as possible and (ii) to report confidence as accurately as possible (Sonnemans & Theo Offerman, 2001). In support of our hypothesis, analysis of the full factorial design showed that while confidence varied with contrast ($F(1,19) = 32.97$, $p < .001$), confidence did not vary with variability ($F(1.2,22.5) = 0.73$, $p > 0.4$). In addition, direct comparison between the low-c and high-v conditions showed that participants were more confident in the high-v condition ($t(19) = 3.98$, $p < .001$), with participants overestimating their performance (comparison between mean confidence and accuracy; $t(19) = 2.66$, $p < .05$). While participants reported slightly lower confidence in the high-v condition compared to baseline (**Fig. 3A**), this decrease was due to participants utilising response times as a cue to confidence (Zakay & Tuvia, 1998). A trial-by-trial regression analysis revealed that confidence decreased with longer response times (RTs) and was unaffected by variability once RTs had been accounted for (v: $t(19) = 0.38$, $p > 0.7$; all other t-values > 4, all $p < .001$; **Fig. 3B**). In other words, participants were overconfident under integration noise, as if they were "blind" to the impact of integration noise on their performance.

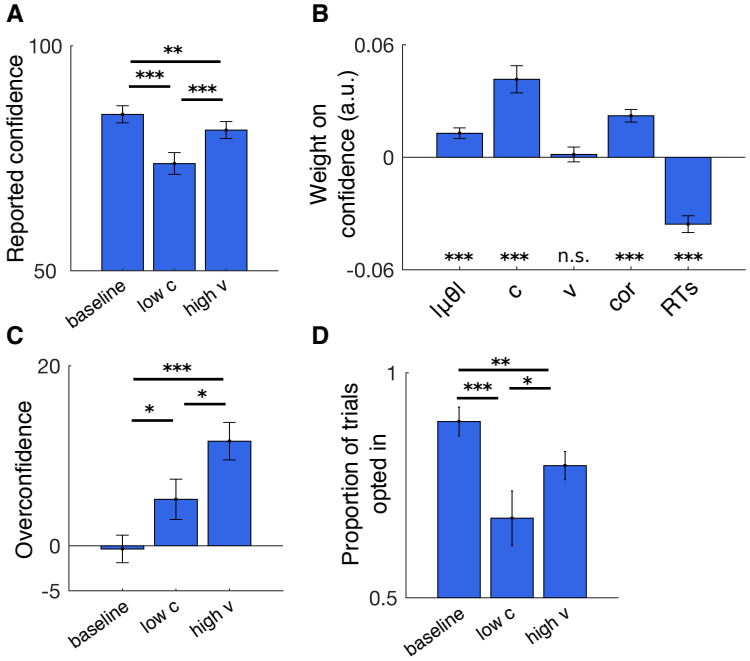

Fig. 3. Effects of contrast and variability on explicit and implicit markers of confidence. (**A**) Mean confidence in the baseline, reduced contrast (low-c) and increased variability (high-v) conditions. (**B**) Trial-by-trial confidence is not influenced by variability (v) but is influenced by the deviation of the average orientation from horizontal (|μθ|), contrast (c), choice accuracy (cor), and response times (RTs). (**C**) Overconfidence, the difference between mean confidence and mean choice accuracy, is highest in the high-v condition. (**D**) Higher probability of making a choice (and thus not opting out) in the high-v condition compared to the low-c condition. **(A-D)** Data are represented as group mean ± SEM. For panel B, only biased trials were used. For all other panels, only neutral trials were used.

Since confidence reports can be idiosyncratic (Aitchison, Bang, Bahrami, & Latham, 2015; Bang et al., 2017), in Experiment 3 ($n = 18$), we obtained an implicit, but arguably more direct measure, of confidence (Hampton, 2001; Kepecs & Mainen, 2012; Kiani & Shadlen, 2009). More specifically, on half of the trials ("optional trials"), we introduced an additional choice option, an

opt-out option, which yielded the same reward as a correct choice with 75% probability. On the other half of trials ("forced trials"), participants had to make an orientation judgement. Under this design, to maximise reward, participants should choose the opt-out option whenever they thought they were less than 75% likely to make a correct choice. Despite matched levels of choice accuracy in the low-c and the high-v conditions (forced trials, $t(17) = 0.24$, $p > 0.8$), participants decided to make an orientation judgement more often on high-v than on low-c trials (optional trials, $t(17) = 2.32$, $p < .05$; **Fig. 3D**), again suggesting overconfidence in the face of integration noise. A full factorial analysis verified that the proportion of such opt-in trials varied with contrast ($F(1,17) = 21.2$, $p < .001$) but not with variability ($F(1.4,23.9) = 3.6$, $p > 0.05$). Similarly, a trial-by-trial logistic regression showed that the probability of opting in varied with contrast ($t(17) = 6.93$, $p < .001$) but not with variability ($t(17) = 1.6$, $p > 0.1$). Overall, while participants opted out more often when encoding noise was high, they did not do so when integration noise was high, despite making a comparable amount of errors in the two conditions.

Computational model of noise blindness

To provide a mechanistic explanation of the observed data, we compared a set of models derived from the ideal observer framework, a common approach in psychophysics and sensory neuroscience. The guiding principle behind this framework is that an agent has an internal model of how the stimulus generates sensory data – a generative model – and inverts this model to compute choices and confidence. We first describe the features that are common to the models, and then explain how they make different assumptions about an agent's generative model.

In our task, the average orientation of a stimulus array was sampled from a distribution of orientations common across experimental conditions (**Fig. 4A**). We modelled the agent's sensory data as a random (noisy) sample from a Gaussian distribution centred on the average orientation of the stimulus array (**Fig. 4B**), with the variance of this distribution determined by both encoding noise and integration noise. We used participants' data from the neutral trials to parameterise the levels of encoding noise and integration noise in each experimental condition. Our factorial design allowed us to disentangle the contribution of the two types of noise. We assumed that the agent uses Bayesian inference to compute the probability that a stimulus is clockwise (CW) and counter-clockwise (CCW) given the sensory data. Specifically, we assumed that an agent has, for each experimental condition, a pair of functions that describe the probability density over sensory data given a CW and a CCW stimulus. We assumed that the agent knows the current experimental condition (e.g., high-contrast and high-variability trial) and uses the relevant density functions to compute the probability of the observed sensory data given a CCW and a CW category (**Fig. 4C**). On neutral trials, each category is, a priori, equally likely, and the agent computes the probability that a stimulus is CW and CCW directly from the density functions. However, on biased trials, the categories have different prior probabilities, and the agent scales the two density functions by the prior probability of each category as indicated by the prior cue (**Fig. 4D**). After having computed the probability that a stimulus is CW and CCW, the agent can generate a choice (i.e. chose the category with the higher posterior probability given the sensory data) and confidence in this choice (i.e. the probability that the choice is correct).

We considered three models which take into account encoding noise but which make different assumptions about an agent's awareness of and/or sensitivity to integration noise. First, we considered an *optimal* model which takes into account both encoding noise and integration noise and has a pair of density functions for each experimental condition. Second, we considered

a *variability-blind* model which takes into account both encoding noise and integration noise but, because she cannot identify the different experimental conditions of orientation variability, mixes her density functions across these conditions. As a result, compared to the optimal model, the density functions are wider on low-variability trials but narrower on high-variability trials. Finally, we considered a *noise-blind* model which takes into account encoding noise but entirely neglects integration noise. As for the variability-blind model, the noise-blind model uses a common pair of density functions across the different levels of orientation variability, but, critically, these density functions do not reflect the extra noise induced by high-variability stimuli.

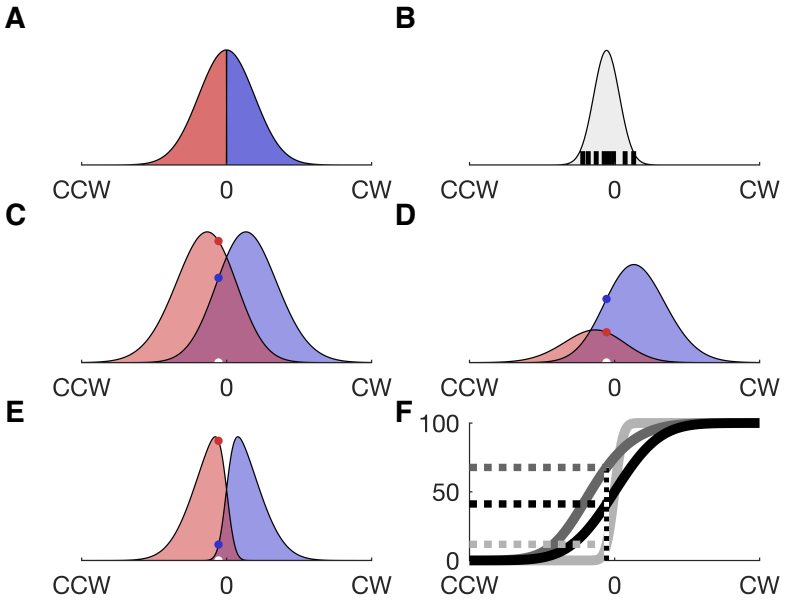

Fig. 4. Computational model. (**A**) Distribution of average orientations conditioned on CCW (red) and CW (blue). (**B**) An agent's sensory data was modelled as a sample from a Gaussian distribution centred on the average orientation of the eight gratings in a stimulus array (black vertical lines), with the variance of this distribution determined by both encoding noise and integration noise. (**C**) An agent has a pair of category-conditioned probability density functions over sensory data for each experimental condition (i.e. contrast x variability level; here a single condition is shown). The agent uses the relevant pair of density functions to compute the probability of the observed sensory data (white dot) given each category (red and blue dots). See an example of a full set of density functions for an experiment in **Fig. S1**. (**D**) Density functions from panel C after scaling by the prior cue (here 75% CW). The sensory data is now more likely to come from a CW stimulus than a CCW stimulus. (**E**) The noise-blind model only takes into account encoding noise and her density functions overlap less than in panel C. (**F**) Posterior belief that the stimulus is CW as a function of sensory data for the models shown in panels C (black, optimal model on neutral trials), D (dark grey, optimal model when prior cue is 75% CW) and E (light grey, noise-blind model on neutral trials).

While the three models make the same predictions about choice on neutral trials, they differ in terms of their subjective certainty and therefore, on biased trials, make different predictions about how participants utilise the prior cue. Similarly, the models make different predictions about confidence (implicit or explicit) on both neutral and biased trials. In support of our hypothesis, the noise-blind model provided the best fit to our data. First, the noise-blind model, and not the optimal model, predicted three key features of participants' behaviour: (i) overconfidence on high-variability trials within participants, (ii) no correlation between mean accuracy and mean

confidence across participants (**Fig. 5A**) and (iii) a diminished influence of the prior cue on high-variability trials, as seen by both the analysis of the bias index (**Fig. 5C**) and the trial-by-trial regression predicting confidence (**Fig. 5D**). In addition, quantitative comparison yielded "very strong evidence" (Kass & Raftery, 1995) for the noise-blind model over the optimal model, with an average ΔBIC across participants of -18.5 (**Fig. 5B**). Similarly, analyses of the patterns of overconfidence in the critical conditions of our factorial design favoured the noise-blind over the variability-blind model (**Fig. S2**), and quantitative comparison yielded "strong evidence" for the noise-blind over the variability-blind model (ΔBIC = -9.2). The modelling indicates that participants were not unable to assign integration noise to its appropriate cause (as in the variability-blind model) but that they instead neglected integration noise altogether (as in noise-blind model).

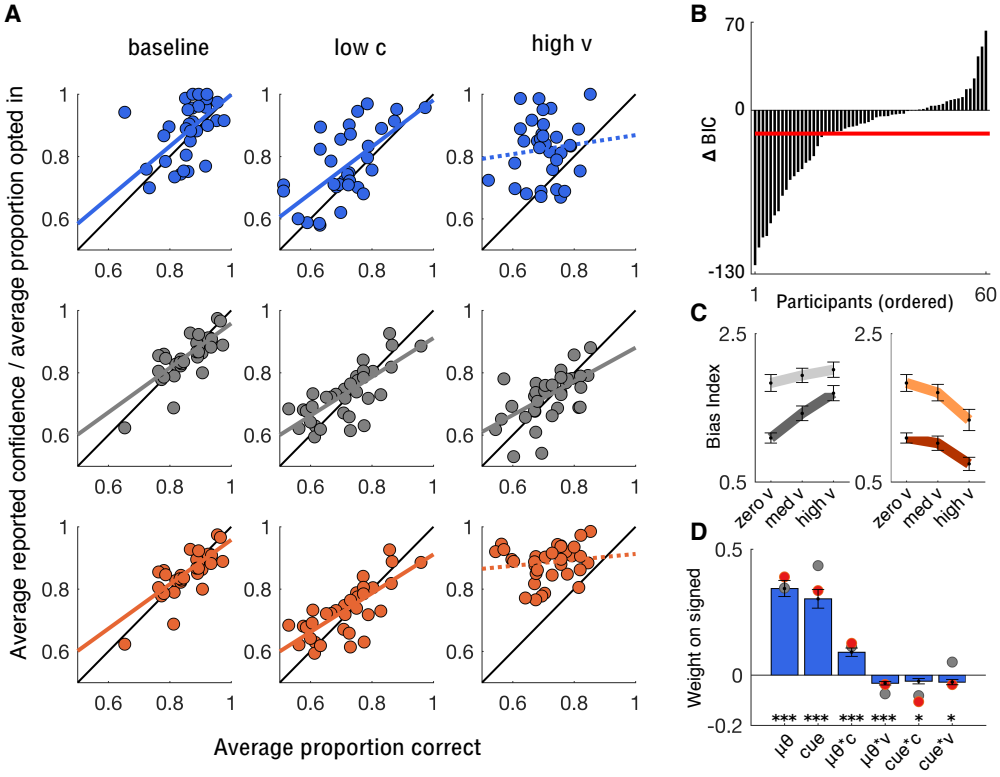

Fig. 5. Comparison of model and human behaviour. (**A**) Correspondence between mean accuracy and mean confidence (explicit estimates or opt-in responses) for participants (blue, data from Exp1-3), optimal observers (grey) and noise-blind observers (orange) in the baseline, reduced contrast (low-c) and increased variability (high-v) conditions. Coloured lines indicate best-fitting slope of a linear regression analysis: solid for $p < .001$, dotted for $p > 0.4$. (**B**) Model comparison (Exp1-3) suggests strong evidence in favour of the noise-blind model over the optimal model (average ΔBIC = -18.5). (**C**) Optimal model (left) and noise-blind model (right) make opposite predictions about the influence of the prior cue on choices (Exp1-2) as variability increases (positive versus negative slopes) but similar predictions as contrast decreases (lighter lines above darker ones). (**D**) Trial-by-trial analysis of signed confidence (Exp2; negative for CCW and positive for CW) for participants (blue), optimal model (grey) and noise-blind model (orange). **(C-D).** Data are represented as group mean ± SEM. For panel A only neutral and optional trials were used, for panels B and C only biased trials were used, for panel D both neutral and biased trials were used. See Methods for details on computational models.

Noise blind and not variability blind

To further rule out that participants were unable to track orientation variability as proposed by the variability-blind model, we ran Experiment 4 (n = 24) in which participants, after having made an orientation judgment, were asked to categorise either the contrast of the stimulus array (rmc, high: .6 vs. low: .15) or the variability of the stimulus array (std, high: 10º vs. low: 0º) (**Fig. 1E**). Again, choice accuracy on the neutral trials of the low-c and the high-v conditions were statistically indistinguishable ($t(23) = 1.16$, $p > 0.2$), allowing us to use the corresponding biased trials to compare cue usage when the variability categorisation was correct and incorrect (75.71% ± 2.26% of variability categorisations were correct). In line with our our hypothesis, participants were more biased on low-c than high-v trials regardless of whether the variability categorisation was correct ($t(23) = 3.21$, $p < .01$) or incorrect ($t(23) = 4.05$, $p < .001$; see **Fig. 6A-B**). In other words, participants showed noise blindness even when they could correctly report the variability of the array, ruling out variability blindness as an explanation of our findings.

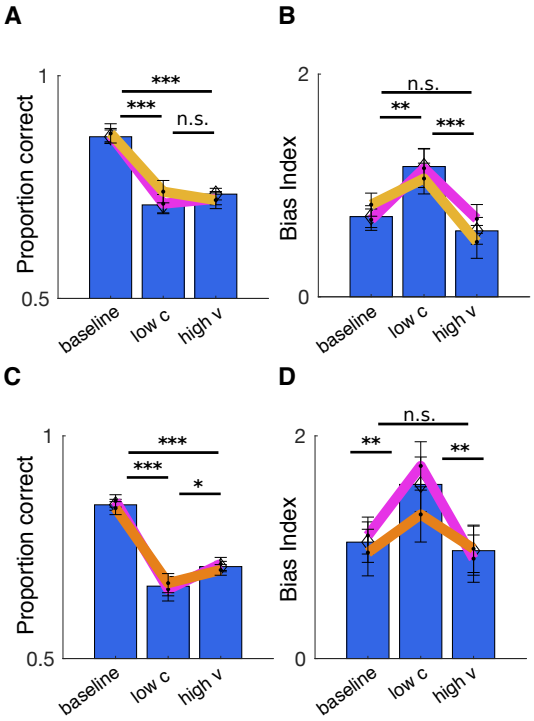

Fig. 6. Experimental evidence against variability-blindness. (**A**) Choice accuracy for the baseline, reduced contrast (low-c) and increased variability (high-v) conditions for Experiment 4. (**B**) For Experiment 4, the influence of the prior cue is highest when encoding noise is high (low-c) and lowest when integration noise is high (high-v). (**C**) Same as panel A, but for Experiment 5. (**D**) Same as panel B, but for Experiment 5. (**A-B**) Coloured lines indicate trials where the subsequent categorisation of contrast was correct (pink lines) and where the subsequent categorisation of variability was correct (yellow lines). (**C-D**) coloured lines indicate trials where the contrast level was blocked (pink lines) or when the variability level was blocked (brown lines). (**A-D**) Data are represented as group mean ± SEM. For panels A and C neutral trials were used, for panels B and D biased trials were used.

In Experiments 1-4, the experimental conditions were interleaved across trials, which may have made it too difficult for participants to maintain a separate pair of density functions for each experimental condition. To test the generality of our findings, we ran Experiment 5 (n = 24) in

which either the contrast or variability level were kept constant across a block of trials (**Fig. 6C-D**). Even then, and despite receiving trial-by-trial feedback, participants were, compared to the baseline condition, not more influenced by the prior cue when variability was high (biased trials, $t(23) = 0.32$, $p > 0.7$), only when contrast was low (biased trials, $t(23) = 3.31$, $p < .01$). The results show that, even when given ample opportunity to learn that integration noise corrupts performance, participants failed to do so.

Sequential sampling account of noise blindness

A recent study, using random dot kinematograms, investigated how the volatility of motion coherence within a trial affects choices and confidence (Zylberberg, Fetsch, & Shadlen, 2016). Participants were found to make faster responses and report higher confidence when the volatility of motion coherence was high. These results were explained by a sequential sampling model which assumes that observers are unaware of the volatility of motion coherence and therefore, unlike an optimal model, adopt a common choice threshold across trials. In the Supplementary Information, we show, using empirical and computational analyses, that this model cannot explain our results (see **Fig. S3**). In short, the model predicts *faster* RTs on high-variability trials, a prediction which is at odds with our observation of *slower* participants' RTs on high-variability trials.

Noise blindness cannot be explained by subsampling

We have proposed that orientation variability impairs performance because of additional noise inherent to cognitive integration of variable or discordant pieces of information. However, another possible explanation of this relationship is that participants based their choices on a subset of gratings rather than the full array. Under this *subsampling* account, choice accuracy on high-variability trials is lower because of a larger mismatch between the average orientation of the full array of gratings and the average orientation of the sampled subset. To test this hypothesis, we conducted Experiment 6 ($n = 20$) in which we manipulated the set-size of the stimulus array (4 or 8 gratings, with orientation variability equated across set-sizes). Under the subsampling account, we would expect a decrease in choice accuracy on high-variability trials for eight compared to four gratings because, on the former set of trials, there would be a larger mismatch between the average orientation of the full array and the average orientation of a sampled subset. While we observed effects of contrast ($F(1,20) = 40.9$, $p < .001$) and variability ($F(1,20) = 30.50$, $p < .001$) on choice accuracy, there was no effect of set-size ($F(1,20) = 0.006$, $p > 0.9$; **Fig. S4A**). Further, in the Supplementary Information, we show, using simulations, that a subsampling model provides a poor account of our behavioural data (see **Fig. S4B**). Taken together, these results support that the observed results were due to blindness to integration noise and not subsampling.

Discussion

Here we propose a new explanation for the previously reported gap in optimality between perceptual and cognitive decisions. Using a novel task, we showed that humans are sensitive to noise present during sensory encoding, in keeping with previous perceptual tasks (Ernst & Banks, 2002; Körding & Wolpert, 2004), but blind to noise arising when having to integrate variable or discordant pieces of information, a common requirement in cognitive tasks. This noise blindness gave rise to two signatures of suboptimality often found in cognitive tasks: base-rate neglect and overconfidence.

We provided several lines of evidence that participants were blind to integration noise. When stimulus variability was high, participants were overconfident, as indicated by cue usage, confidence reports and opt-in responses, even though they received trial-by-trial feedback, and even when stimulus variability was salient (Exp1-3), accurately categorised (Exp4) or constant across a block of trials (Exp5). These findings indicate that, while participants were able to identify stimulus variability, they neglected the additional noise associated with high-variability stimuli. In addition, control analyses and a control experiment (Exp6) ruled out that such noise blindness was due to attention or subsampling (i.e. basing choices on a subset of gratings), or inconsistency in the strategy adopted by participants across experimental conditions.

There is a vast literature on the different types of noise which affect human choices (Beck, Ma, Pitkow, Latham, & Pouget, 2012; Hunt, 2014; Juslin & Olsson, 1997). Our account is related to a previous distinction between noise which originates inside the brain, such as intrinsic stochasticity in sensory transduction (Thurstone, 1927), and noise which arises outside the brain, such as a probabilistic relationship between a cue and a reward (Brunswik, 1956). Specifically, our account classifies noise according to when it arises during the information processing that precedes a choice. Encoding noise refers to noise accumulated up to the point at which a stimulus is encoded. As such, encoding noise includes both externally generated noise (e.g., a weak correspondence between a retinal image in dim lighting and the object that caused the image) and internally generated noise (e.g., intrinsic stochasticity in sensory transduction). On the other hand, by integration noise we refer to strictly internal noise which arises at later stages of information processing, such as when integrating variable pieces of information within a limited-capacity system. Choices may be affected by other types of noise than those considered here. For example, cognitive decisions often involve memories, sometimes distant in the past, and additional factors such as risk and ambiguity (Bach & Dolan, 2012; Payzan-LeNestour & Bossaerts, 2011). However, recent work has shown that noisy cognitive inference, related to our notion of integration noise, is a primary source of behavioural variability and suboptimal choices (Beck et al., 2012; Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016).

We do not know why humans are blind to integration noise. One possibility is that basing decision strategies on all sources of noise would prolong deliberation and thus reduce reward rates, or that recognising one's own cognitive deficiencies requires a much longer timeframe. However, a well-known cognitive illusion may help understand why blindness to one's own cognitive deficiencies may not be catastrophic: even though failures to detect salient visual change suggests that cognitive processing is highly limited (Simons & Levin, 1997), humans enjoy rich, vivid visual experiences of cluttered natural scenes. Human information processing is sharply limited by capacity, but as agents we may not be fully aware of the extent of these limitation.

Acknowledgments

This work was supported by a Wellcome 4-year-PhD grant to S.H.C. (0099741/Z/12/Z) and an ERC starter grant to C.S. (281628). The Wellcome Centre for Human Neuroimaging is supported by core funding from Wellcome (203147/Z/16/Z).

Author contributions

S.H.C., D.B., T.E. and C.S. conceived the study. S.H.C., D.B., J.D. and C.S. designed the experiments. S.H.C. programmed the experiments. S.H.C, D.B. and J.D. performed the experiments. S.H.C., D.B. and R.M. developed the models. S.H.C. and D.B. analysed the data and performed the simulations. S.H.C., D.B., J.D., T.E. and C.S. interpreted the results. S.H.C. drafted the manuscript. S.H.C, D.B. and C.S. wrote the manuscript.

Competing interests

The authors declare no financial of non-financial competing interests.

Methods

Participants

One hundred and five healthy human participants with normal or corrected-to-normal vision were recruited to participate in six experiments (72 females, 8 left-handed, mean age ± SD: 25.02 ± 4.25; Exp1: n = 20; Exp2: n = *20*; Exp3: n = 20; Exp4: n = 24; Exp5: n = 24; Exp6: n = 20). Participants were reimbursed for their time and could earn an additional performance-based bonus (see below). All participants provided written informed. The experiments were conducted in accordance with local ethical guidelines.

Experimental paradigm

All six experiments were based on the same psychophysical task. On each trial, participants had to judge whether the average orientation of a circular array of gratings (Gabor patches; see **Fig. 1**) was tilted clockwise (CW) or counter-clockwise (CCW) relative to horizontal. The average orientation of the gratings in each trial was randomly selected from a mixture of two Gaussian distributions (centred at 3º either side of the horizontal axis, respectively, and with 8º of standard deviation). We manipulated encoding noise and integration noise by varying two features of the array in a factorial way manner: the root mean square contrast (*rmc*) of the individual gratings, which affected the difficulty of encoding each grating, and the variability of the orientations of the individual gratings, which affected the difficulty of estimating the average orientation of the array. The number of contrast and orientation variability conditions varied across different experiments: in Experiments 1-3, three contrast levels (rmc = {0, .16, .6}) and three variability levels (SD = {0º, 4º, 10º}); in Experiments 4-6, two contrast levels (rmc = {.15, .6}) and two variability levels (SD = {0º, 10º}). The stimulus array was presented for 150 ms and was followed by a 3000 ms choice period. Participants indicated their choice by pressing the right (CW) or the left (CCW) arrow-key on a QWERTY keyboard. They received feedback about choice accuracy, before continuing to the next trial. If no response was registered within the choice period, the word "LATE" appeared at the centre of the screen, and the next trial was started. Experiments 1, 2 and 3 consisted of 1296 trials, divided into 36 blocks of 36 trials each. Experiments 4, 5 and 6 consisted of 1200 trials, divided into 32 blocks of 40 trials each.

In Experiments 1 and 2, participants were presented with a cue to the prior probability of each stimulus category. The cue was presented 700 ms before the onset of the stimulus array and remained on the screen until a response was registered. An "N" indicated that the two stimulus categories were equally likely, an "R" indicated a 75% probability of a CW stimulus (and thus a 25% probability of a CCW stimulus) and an "L" indicated a 75% probability of a CCW stimulus. Half of the blocks contained neutral trials ("N") and the other half contained biased trials ("R" or "L"). The blocks were randomised across an experiment. In 2 Experiment, after having made a choice, participants were required to indicate the probability that the choice is correct by moving a sliding marker along a scale (50% to 100% in increments of 1%). In Experiment 3, on half of the blocks, participants could decide to opt out of making a choice and receive the same reward as a correct choice with 75% probability. There was no prior cue. In Experiment 4, after having made a choice, participants had to categorize (high vs. low) either the contrast or the variability of the stimulus array (counterbalanced across trials). Feedback about the accuracy of each judgement was presented at the end of the trial. In Experiment 5, for each block of trials, we kept contrast or variability constant while varying the other feature of the stimulus array (e.g., if contrast was fixed

at low then variability could be low or high). In Experiment 6, on half of the blocks, the stimulus array consisted of eight gratings (as in all other experiments) and, on the other half of blocks, the stimulus array consisted of four gratings. Further details about stimulus presentation, trial timings and so forth are provided in the Supplementary Information.

Statistical analyses

All statistics are reported at the group level. Catch trials with zero contrast (rmc = 0) were included in Experiments 1-3 to allow for estimation of participants' baseline cue usage, but these trials were excluded from most analyses. We performed analyses of variance (ANOVAs), with participants as a random variable, to test the effects of contrast and variability on choice accuracy, response times, cue usage, confidence (Exp2) and opt-in behaviour (Exp3). We performed most analyses of choice accuracy and confidence using neutral trials; analyses of cue usage were naturally based on biased trials. We used multiple linear regression and multiple logistic regression to isolate the effect of variability on confidence and on opt-in responses, respectively. For the analyses in Fig. 5A, seven participants were excluded because of excessive opt-out (n = 33), but result were almost identical when including them. All p-values lower than .001 are reported as "$p < .001$", p-values higher or equal than .001 but lower than .01 are reported as "$p < .01$", p-values higher or equal to .01 but lower than .05 are reported as "$p < .05$". All p-values greater or equal to .05 are reported as higher than the closest lower decimal (e.g., a p-value of .175 would be reported as "$p > 0.1$"), with exception of p-values between .05 and .1 which are reported as "$p > .05$". The degrees of freedom for the ANOVAs are specified using non-integer numbers when a Greenhouse-Geisser correction has been used to correct for violations of the sphericity assumption.

Computational modelling

We compared three models of the processes underlying participants' behaviour. We first detail the *optimal* model, which takes into account both encoding noise and integration. We then detail the *variability-blind* model, which takes into account integration noise but cannot distinguish the different experimental conditions of orientation variability, and the *noise-blind* model, which entirely neglects integration noise.

We modelled all agent's noisy estimate, x, of the true average orientation, μ, as a random sample from a Gaussian distribution with mean μ, and variance σ^2:

$$x = \epsilon(\mu, \sigma^2) \qquad \text{(eq. 1)}$$

where σ is the total level of noise (encoding noise plus integration noise; see below for estimation of noise levels).

We assumed that an optimal agent would, for each experimental condition, have a unique pair of category-conditioned probability density functions (PDFs) over sensory data, which reflect her total level of noise and the probability distribution over true average orientations, and which allows her to compute the probability of the observed sensory data given each category:

$$PDF_{cat} = \mathrm{P}(x|cat) \qquad \text{(eq. 2)}$$

where cat is the CW or the CCW category. We constructed these PDFs, for each experimental condition, by convolving the true probability distribution over average orientations with a zero-centred Gaussian distribution with variance σ^2. An example pair of PDFs is shown in **Fig. 4C.**

We assumed that an agent would compute the probability of each category using Bayes' theorem:

$$\mathrm{P}(cat|\ cue, x) = \frac{\mathrm{P}(x|cat) \cdot\ \mathrm{p}(cat)}{(\mathrm{P}(x|cat) \cdot\ \mathrm{p}(cat) + \mathrm{P}(x|cat_{alt}) \cdot\ \mathrm{p}(cat_{alt}))}$$

(eq. 3)

where $\mathrm{P}(x|cat)$ is computed using the PDFs described above and $\mathrm{p}(cat)$ is the prior probability of the category in question as indicated by the prior cue. If the category in question is CW then the alternative category, cat_{alt} is CCW, and vice versa. On neutral trials, the prior probability of each category is 50%. On biased trials, the prior probability of one category is 75% and the prior probability of the other category is 25%. The computation detailed in eq. 3 can be thought of as scaling the PDFs described above by the prior probabilities of each category. An example of such scaled PDFs is shown in **Fig. 4D**.

Finally, we assumed that an agent makes a decision, d, by simply selecting the category with the higher posterior probability and computes confidence in this decision as:

$$p(d = cat|\ cue, x)$$

(eq. 4)

which in our task is directly given by the posterior probability of the chosen category.

As mentioned above, we assumed that an optimal agent has a unique pair of PDFs for each experimental condition. As such, an optimal agent would have six pairs of PDFs in Experiments 1-3 and four pairs of PDFs in Experiments 4-6. Further, we assumed that these pairs of PDFs, for an optimal agent, incorporate the total amount of encoding noise and integration noise in each experimental condition. For example, a pair of PDFs will overlap the least when contrast is highest and variability is lowest, and overlap the most when contrast is lowest and variability is highest. Finally, we assumed that an optimal agent can correctly identify the current experimental condition (an assumption which we verified in Experiments 4-5). For these reasons, an optimal agent would (i) assign an appropriate weight to the prior cue, (ii) accurately estimate the probability of having made a correct choice, and (iii) opt out of trials when objectively being less than 75% likely to be correct. We now describe two models which relax these assumptions.

We first considered a variability-blind agent who is sensitive to integration noise but unable to distinguish the different levels of orientation variability. As such, a variability-blind agent is equivalent to an optimal agent whose PDFs have been mixed across variability levels. As a result, when orientation variability is low, the PDFs are more overlapping than under the optimal model. Conversely, when orientation variability is high, the PDFs are less overlapping than under the optimal model. For these reasons, a variability-blind model would display a mixture of under- and overconfidence.

We next considered a noise-blind agent who is entirely unaware of integration noise. Like in the case of a variability-blind agent, a noise-blind agent only has a pair of PDFs for each contrast level but, unlike in the case of a variability-blind agent, these PDFs only take into account encoding noise. As a result, when orientation variability is higher than zero, the PDFs are less overlapping

than under either of the two other models (**Fig. 4E**). A noise-blind agent would therefore, when orientation variability is higher than zero, hold stronger posterior beliefs and (i) assign too low weight to the prior cue, (ii) overestimate the probability of having made a correct choice, and (iii) not opt out of trials when objectively being less than 75% likely to be correct.

We note that the three models presented make the same predictions about choice on neutral trials but are distinguishable when focusing on (i) biased trials and (ii) confidence and opt-in behaviour on both neutral and biased trials. Our modelling approach allowed us to derive a choice probability for each model on each trial. For model analyses requiring a categorical choice (e.g., logistic regression), we sampled choices according to the aforementioned choice probabilities.

Noise estimation

Each condition was assumed to be affected by Gaussian noise with a specific standard deviation, σ_{cond}.We estimated the total level of noise in each experimental condition using a procedure which was entirely independent of the modelling detailed above. In particular, four parameters (three for Experiments 4-6) allowed us to simultaneously specify the total level of noise in each and every experimental condition. Two parameters characterised the level of encoding noise for each contrast level: one for low contrast (nC_{low}) and one for high contrast (nC_{high}). Two extra parameters (only one for Experiments 4-6) characterised the level of integration noise for each variability level: one for medium variability (nV_{med}, only for Experiments 1-3) and one for high variability (nV_{high}).. For a given experimental condition, the total level of noise (the standard deviation of the Gaussian noise distribution), σ_{cond}, is thus given by:

$$\sigma_{cond} = \sqrt{(\varepsilon\sigma_{cond}^2) + (\iota\sigma_{cond}^2)} \qquad \text{(eq. 5)}$$

where $\varepsilon\sigma_{cond}$ and $\iota\sigma_{cond}$ specify the contribution of encoding noise and integration noise, respectively. For instance, for the low-contrast high-variability condition the total amount of noise would be given by substituting nC_{low}, for $\varepsilon\sigma_{cond}$, and nV_{high}, for $i\sigma_{cond}$, in Equation 5. We fitted the four noise estimators for each participant by maximizing the likelihood of the participant's choice using neutral trials only (we used a genetic algorithm with a population size of 100 individuals and a maximum generation time of 1000 generations). Our factorial manipulation of contrast and variability meant that, for each experimental condition, the amount of total noise required to explain participants' choice probabilities was related to both contrast and variability. However, it also meant that there was a clear relationship between a given contrast or variability level and contributed noise (e.g., the noise contribution of low contrast is the same on low-contrast, low-variability trials and on low-contrast, high-variability trials). Once knowing the amount of encoding noise and integration noise in each experimental condition, these estimates could be used to build the PDFs of the models described above. We highlight that our procedure for noise estimation uses choices on neutral trials only, and that the model predictions are evaluated using independent features of the data: (i) confidence on neutral trial choices, (ii) choices (and choice probabilities) on biased trials, and (iii) probability of opting out.

Psychometric fits

For visualization purposes only, we fitted psychometric curves to the average proportion of clockwise choices in **Fig. 2B**, using four free parameters, according to the following logistic function:

$$P = \frac{A_1 - A_2}{1 + e^{(x-x_0)/dx}} + A_2$$

(eq. 6)

where P is the proportion of CW choices, A_1 is the right asymptote, A_2 is the left asymptote, x_0 is the inflection point and $1/dx$ is the steepness, and x is the average stimulus orientation at which the proportion of CW choices is evaluated.

Bias index

We used Signal Detection Theory to calculate the decision criteria, *c*, separately for trials on which the prior cue favoured CW and trials on which the prior favoured CCW; the decision criterion provides a signed estimate of the degree to which the prior cue biases a participants' choices which is independent of their sensitivity to average orientation (Macmillan & Creelman, 2004; Stanislaw & Todorov, 1999). We then used the difference in the values of *c* when cued CW (C_{CW}) and when cued CCW (C_{CCW}) as our measure of cue usage: bias index = C_{CW}-C_{CCW}. Values close to zero indicate low cue usage, whereas greater values indicate higher cue usage. We obtained a separate bias index for each participant and each experimental condition.

References

Ackermann, J. F., & Landy, M. S. (2014). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Attention, Perception, & Psychophysics*, *77*(2), 638–658.

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10), e1004519.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572.

Bang, D., Aitchison, L., Moran, R., Herce Castañón, S., Rafiee, B., Mahmoodi, A., … Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), s41562–017.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624.

Brunswik, E. (1956). *Perception and the representative design of experiments.* CA: Berkeley: University of California Press.

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between belief and performance in perceptual decision making. *PloS One*, *9*(5), e96511.

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *92*(6), 1398–1411.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Faisal, A. A., & Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *Journal of Neurophysiology*, *101*(4), 1901–1912.

Frydman, C., & Camerer, C. F. (2016). The Psychology and Neuroscience of Financial Decision Making. *Trends in Cognitive Sciences*, *20*(9), 661–675.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences*, *98*(9), 5359–5362.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523.

Hunt, L. T. (2014). What are the neural origins of choice variability? *Trends in Cognitive Sciences*, *18*(5), 222–224.

Jarvstad, A., Hahn, U., Rushton, S. K., & Warren, P. A. (2013). Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proceedings of the National Academy of Sciences*, *110*(40), 16271–16276.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: a sampling model of confidence in sensory discrimination. *Psychological Review*, *104*(2), 344.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1594), 1322–1337.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764.

Knill, D. C., Kersten, D., & Yuille, A. (1996). Introduction: A Bayesian formulation of visual perception. *Perception as Bayesian Inference*, 1–21.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Levitt, S. D., & List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *The Journal of Economic Perspectives*, *21*(2), 153–174.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Mamassian, P., Landy, M. S., & Maloney, L. T. (2002). Bayesian modelling of visual perception. *Probabilistic Models of the Brain: Perception and Neural Function*, 13–36.

O'Reilly, J. X., Jbabdi, S., Rushworth, M. F., & Behrens, T. E. (2013). Brain systems for probabilistic and dynamic prediction: computational specificity and integration. *PLoS Biology*, *11*(9).

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*(1), e1001048.

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*, *16*(1), 138.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267.

Sonnemans, J., & Theo Offerman, T. (2001). Is the quadratic scoring rule behaviorally incentive compatible?

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences*, *19*(1), 27–34.

Sun, J., & Perona, P. (1998). Where is the sun? *Nature Neuroscience*, *1*(3), 183–184.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*(8), 291–297.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, *22*(18), 1641–1648.

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*(9), 3593–3598.

Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, *98*(1), 103–115.

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *ELife*, *5*, e17688.

Supplementary Information

Experimental details

In all experiments, participants had to judge the average orientation of an array of gratings as clockwise (CW) or counter-clockwise (CCW) from horizontal. We first describe trial events, trial timings and stimulus construction for Experiment 1 and then explain the additional steps taken for Experiments 2-6.

In Experiment 1, a fixation dot first appeared at the centre of the screen for 300 ms to announce the start of a trial. The fixation dot was replaced by a cue which appeared 700 ms before the stimulus array and which remained on the screen until a response was registered. The cue determined the prior probability of each stimulus category ("L": prior probability of CCW is 75%; "N": CCW and CW equally likely; "R": prior probability of CW is 75%). The stimulus array was shown for 150 ms and was followed by an up-to 3000 ms long response window. Participants responded by pressing the left (CCW) or right (CW) arrow-key on a QWERTY keyboard using their right hand. Categorical feedback about choice accuracy ("CORRECT" or "WRONG") appeared once a response had been registered and remained on the screen for 500 ms, before the onset of the next trial. If no response was registered within the response period, the word "LATE" appeared at the centre of the screen for 3000 ms, and the next trial was automatically started.

The stimulus was composed of eight gratings displayed within a circular array. We manipulated two features of the stimulus array in a factorial manner: the contrast of the gratings and the variability of the gratings' orientation.

The centre of each grating was located at a distance of ~4.3 degrees of visual angle (400 pixels) from the centre of the screen. Each grating was a Gabor patch constructed using the following parameter values: diameter of ~1.07 degrees of visual angle (100 pixels); spatial frequency of ~5 cycles per degree of visual angle (0.05 cycles per pixel); random phase. All gratings had the same root mean square contrast (rmc). The rmc of a trial (henceforth contrast) was either 0 (no signal), .15 (low contrast) or .60 (high contrast). We added low-level random noise to the gratings (Wyart, Nobre, & Summerfield, 2012). For each grating, we convolved a unique patch of white noise with a two-dimensional zero-mean Gaussian with a standard deviation of ~0.21 degrees of visual angle (20 pixels). The amplitude of the noise was 10% of the maximum possible. We then added the noise to the grating. Finally, we convolved the grating with a 2-dimensional Gaussian envelope peaking at the centre of the grating and decaying with a standard deviation of ~0.21 degrees of visual angle (20 pixels).

The average orientation of gratings on a trial (henceforth trial mean) was randomly drawn from a Gaussian distribution with a mean of -/+ 3º and a standard deviation of 8º. The variability in the orientations of gratings on a trial (henceforth variability) was randomly drawn from a Gaussian distribution with a mean 0º and a standard deviation of either 0º (zero variability), 4º (medium variability) or 10º (high variability). To ensure the trial mean remained unchanged after the variability manipulation, we subtracted the mean deviation from 0 from the gratings' orientations. Together, these steps allowed us to independently manipulate trial mean, contrast and variability. We emphasise that feedback was determined by the average orientation of the gratings presented and not by the distribution from which they were drawn.

The experiment consisted of 1296 trials, distributed into 36 blocks of 36 trials each. On half of the blocks, the prior cue was "N" (neutral trials). On the other half of blocks, the cue varied

between “L” or “R” in a trial-by-trial manner (biased trials). Block order was randomised across an experiment and across participants.

In Experiment 2, we introduced an explicit measure of confidence. Participants indicated their choice by pressing “Z” (CCW) or “X” (CW) using their left hand. After having made a choice, participants had to indicate the probability that the choice is correct. Participants indicated their confidence by sliding a marker along a vertical scale (50% to 100% in increments of 1%) using a standard computer mouse with their right hand. The probability associated with the marker’s current position was updated in real-time and shown at the centre of the screen. Participants confirmed their response by clicking the left button of the mouse. There was no time limit for the confidence judgment. Feedback about choice accuracy appeared 300 ms after a response had been confirmed. Trial numbers, block types and structure were the same as for Experiment 1.

In Experiment 3, we introduced an implicit measure of confidence. On half of the blocks, participants could choose to opt out of making a choice and receive the same reward as a correct choice with a 75% probability. To remind participants about the choice options on a trial, the words “LEFT” (CCW) and “RIGHT” (CW) appeared to the left and the right of the fixation cross after the stimulus disappeared and, when the opt-out option was available, the words “OPT OUT” appeared below the fixation cross. The opt-out option was selected by pressing the downwards arrow key. For feedback, “SUCCESS” was shown after a correct choice and if an opt-out response was successful, whereas “FAILURE” was shown after an incorrect choice and if an opt-out response was unsuccessful. The experiment consisted of 1296 trials, distributed into 36 blocks of 36 trials each. On half of the blocks, the opt-out option was not available. On the other half of the blocks, the opt-out option was available. Block order was randomised across an experiment and across participants. There was no prior cue.

In Experiment 4, we asked participants to categorise either the contrast (rmc = {.15, .60}) or the variability (std = {0º, 10º}) of the stimulus array. In particular, after having made a choice (by pressing the “X” and “Z” buttons using their left hand, with the chosen category highlighted in bold), participants were then required to judge whether the contrast of the stimulus array was high or low or whether the variability of the stimulus array was high or low. The relevant stimulus dimension for the second judgment (indicating by displaying “CONTRAST” or “VARIABILITY at the centre of the screen), was determined randomly and was only revealed after an orientation discrimination had been made. Participants made the second judgment by pressing the left (low) or the right (high) arrow key (the options “LOW” and “HIGH” appeared equidistantly to the left and the right of the fixation point). Once participants had made their response, they received feedback about the accuracy of each judgement, indicated by changing the colours of the selected options to red (incorrect) or green (correct). The experiment consisted of 1200 trials, distributed into 32 blocks of 40 trials each. On half of the blocks, the prior cue was “N” (neutral trials). On the other half of blocks, the cue varied between “L” or “R” in a trial-by-trial manner (biased trials). Block order was randomised across an experiment and across participants.

In Experiment 5, we fixed either contrast or variability across blocks of trials. Specifically, within a block, one dimension was fixed (low or high), while the other dimension varied randomly (low or high), for instance, contrast fixed at low while variability varied (high or low) across trials within the block. As in Experiment 4, there were only two levels of contrast and two levels of variability. There were thus eight blocks for each trial type. On half of the blocks for a trial type, the prior cue was “N” (neutral trials). On the other half of blocks for a trial type, the prior cue varied between “L” or “R” in a trial-by-trial manner (biased trials). The experiment consisted of

1200 trials, distributed into 32 blocks of 40 trials each. Block order was randomised across an experiment and across participants.

In Experiment 6, we varied the set-size of the stimulus array. In particular, the stimulus array was composed of either four or eight gratings. As in Experiments 4 and 5, there were only two levels of contrast and two levels of variability. For arrays with only four gratings, the location of the gratings was fixed within a block but randomised across blocks, by sampling a random set of four contiguous locations from the full array with eight gratings. The set-size was varied in a trial-by-trial manner. Half of the trials had a set-size of four gratings and the other half of trials had a set-size of eight gratings. The experiment consisted of 1200 trials, distributed into 32 blocks of 40 trials each. On half of the blocks, the prior cue was "N" (neutral trials). On the other half of blocks, the cue varied between "L" or "R" in a trial-by-trial manner (biased trials). Block order was randomised across an experiment and across participants.

All participants were reimbursed for their participation and had the opportunity to earn an additional performance-based bonus. In all experiments except Experiment 2, participants received a flat rate of £10 and could earn an additional £1 for every 2% increase in choice accuracy relative to 60%. In Experiment 2, participants received a flat rate of £5 and could earn an additional bonus depending on the accuracy of their confidence judgements. We submitted participant' responses to a strictly proper scoring rule under which it was in participants' best interest to make as many correct decisions as possible and to estimate the probability that their choice is correct as accurately as possible (Sonnemans & Theo Offerman, 2001). The average bonus accrued was ~£12.

Participants received a 10-minute introduction to their corresponding task, including the stimulus, sources of choice difficulty, prior cue and prior probabilities, response contingencies, and the rules behind the performance-based bonus. Participants also completed a short practice session (two blocks) of the task before starting the experiment proper.

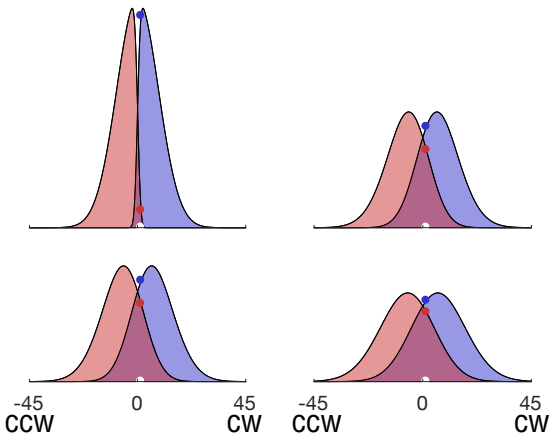

Fig. S1. Category-conditioned probability density functions. Four pairs of PDFs, one for each of the four conditions in Experiment 3. Top-left : high-contrast and low-variability trials (lowest noise). Top-right: high-contrast and high-variability trials (intermediate noise). Bottom-left: low-contrast and low-variability trials (intermediate noise). Bottom-right: low-contrast and high-variability trials (highest noise). The white dots on the x-axes denote an agents sensory data (same across examples). The blue and red dots indicate the probability that the sensory data came from a CW or a CCW category, respectively.

Category-conditioned probability density functions

As detailed in the main text, we assumed that an optimal agent has a unique pair of probability density functions over sensory data (PDFs) for each experimental condition. An example of a full set of such PDFs are shown in **Fig. S1**. Note that the PDFs look more skewed in experimental conditions with low noise (top-left in **Fig. S1**) because the PDFs more closely resemble the true distribution of average orientations (**Fig. 4A**).

Overconfidence in choices

Strong evidence in favour of the noise-blind over the variability-blind model is the pattern of *overconfidence*, defined as the difference between mean confidence and mean accuracy, across the critical conditions of our factorial design. Near-zero values indicate good calibration, whereas non-zero values indicate bad calibration (negative: underconfidence; positive: overconfidence). As shown below in **Fig. S2**, participants were well-calibrated in the baseline condition but highly overconfident in the high-variability condition. This pattern is captured by the noise-blind model. The variability-blind model, however, shows underconfidence in the baseline condition (where integration noise is lower than on average) and overconfidence in the high-variability condition (where integration noise is higher than on average).

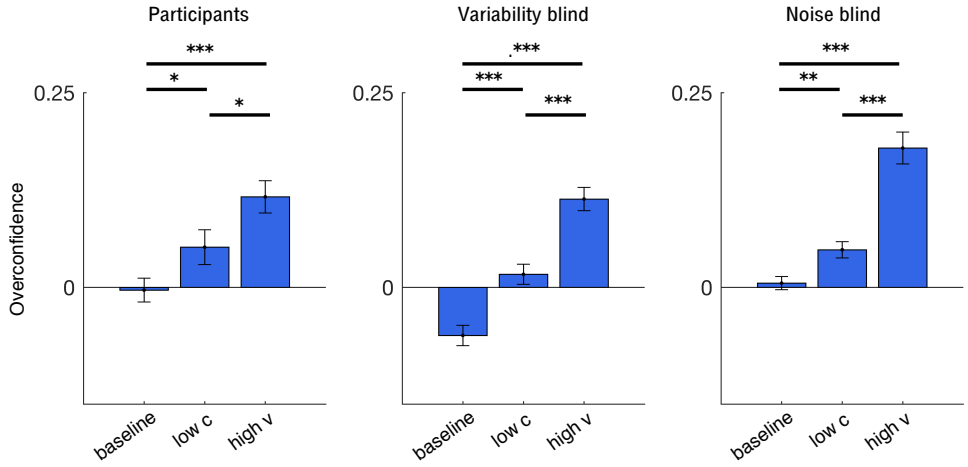

Fig. S2. Noise-blind model only accounts for participants' overconfidence. Participants (left) are well-calibrated in the baseline condition but overconfident in the other conditions. The variability-blind model (middle) shows under-confidence in the baseline condition but overconfidence in the high-variability condition. The noise-blind model shows good calibration in the baseline condition and overconfidence in the high-variability condition. Data are from neutral trials of Experiment 2 and represented as group men ± SEM.

Sequential sampling model of participants' choice behaviour

To test the sequential sampling account of suboptimal behaviour proposed by Zylberberg and colleagues (2016), we fitted a Drift Diffusion Model (DDM) to the data from Experiments 1-2. The DDM models two-choice decision-making as a process of accumulating noisy evidence over time with a certain speed, or drift-rate, until one of two choice thresholds is crossed and the associated response is executed. We assumed, as in the previous study, that the choice thresholds were fixed across experimental conditions, but that lower contrast led to a lower mean of the drift-

rate and that higher variability led to higher variance of the drift-rate. We implemented these mechanisms using three parameters. The first parameter is a scaling factor for the drift-rate and which depends on the contrast level. The second parameter specifies the baseline variance of the drift-rate. Finally, the third parameter scales the baseline variance of the drift-rate according to the variability level. To find the best parameters for each participant, we minimized the sum of squared errors between empirical and predicted choice accuracy across experimental conditions (we used a genetic algorithm with a population size of 1000 individuals and a maximum generation time of 1000 generations). Comparison between empirical data and model predictions are shown in **Fig. S3** using a similar set of analyses as in Zylberberg and colleagues (2016).

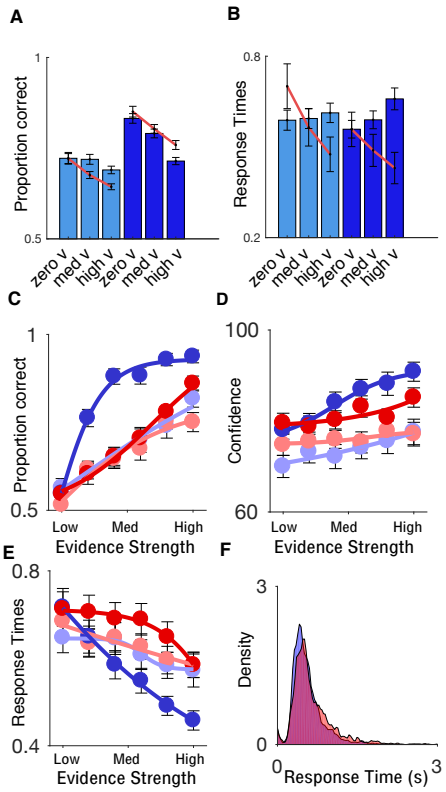

Fig. S3. Common choice threshold across experimental conditions cannot explain our data. (**A**) Choice accuracy for participants (blue bars) and DDMs (red lines) is lower when contrast is low (compare pale blue and dark blue bars) and when variability is high (negative slopes as condition changes from zero-v to high-v). (**B**) Response times for participants and DDMs show opposite effects for increases in variability (positive slopes for participants but negative ones for DDMs). (**C**) Participants' choice accuracy for different levels of evidence strength (blue: low variability, red: high variability; faint colours: low contrast; dark colours: high contrast). Note that the two critical conditions, high-variability and high-contrast trials (dark red curve) and low-contrast and zero-variability trials (faint blue curve), have similar slopes. (**D**) Participants' confidence for different levels of evidence strength (same colour scheme as in panel c). (**E**) Participants' response times for different levels of evidence strength (same colour scheme as in panel c). (**F**) Collapsing response times across participants for high-contrast and zero-variability trials (blue) and high-contrast and high-variability trials (red) demonstrate that high variability is associated with slower response times (i.e. red distribution has a longer tail). Data are from neutral trials of Experiments 1-2 and represented as group men ± SEM.

Subsampling

We used simulations to estimate the minimum number of gratings that participants must have considered to achieve the observed levels of choice accuracy. In particular, we simulated choices for *subsampling* agents who have same levels of encoding noise as participants but who have no integration noise and perfectly average the gratings in a sampled subset. First, we used the observed choice accuracy on high-variability and high-contrast trials to specify a performance benchmark (minimum choice accuracy) which a subsampling agent had to achieve to account for the data. We chose these trials because high variability maximises the cost of subsampling and high-contrast minimises the influence of other sources of choice variability. Second, for each participant, we constructed a subsampling agent with the same level of encoding noise as the participant as inferred from choice on high-contrast and zero-variability trials. We then simulated the agent's choice accuracy on our benchmark trials. For a given trial and hypothesised set-size (1-8 gratings), we computed the mean of a random sample of gratings from the full array and added Gaussian noise (encoding noise) to the average of this subset. We assumed that the agents' choice would be CW if the resulting value was above zero and CCW otherwise. The simulations support that participants could not have achieved the observed levels of choice accuracy if they sampled a subset of gratings rather than the full array (see **Fig. S4B**).

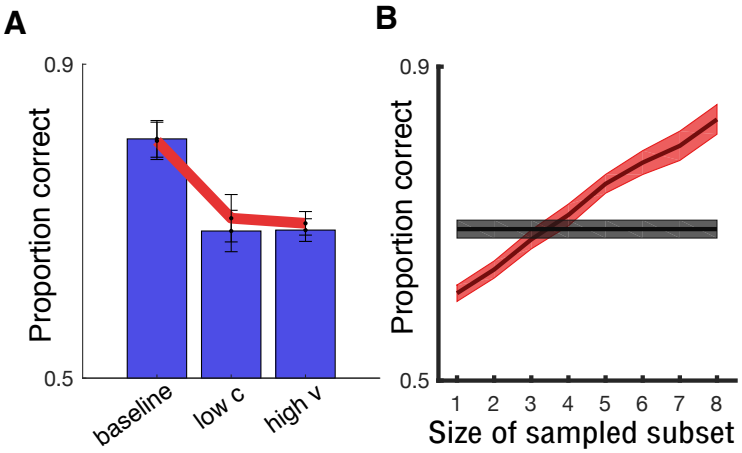

Fig. S4. Noise blind and not subsampling. (A) Participants in Experiment 6 achieved similar levels of performance when required to integrate four items (blue bars) and eight items (red line). **(B)** Choice accuracy for participants (grey shade) and a subsampling model (red shade). The subsampling model, which perfectly averages sampled gratings, would need to sample 3-4 gratings before reaching the same level of choice accuracy as participants, even in the absence of further noise sources. **(A-B)** Data are represented as group mean ± SEM.