Robust averaging during perceptual judgment

Vincent de Gardelle[1] and Christopher Summerfield

Department of Experimental Psychology, University of Oxford, Oxford OX13UD, United Kingdom

Edited by Edward E. Smith, Columbia University, New York, NY, and approved June 28, 2011 (received for review March 22, 2011)

An optimal agent will base judgments on the strength and reliability of decision-relevant evidence. However, previous investigations of the computational mechanisms of perceptual judgments have focused on integration of the evidence mean (i.e., strength), and overlooked the contribution of evidence variance (i.e., reliability). Here, using a multielement averaging task, we show that human observers process heterogeneous decision-relevant evidence more slowly and less accurately, even when signal strength, signal-to-noise ratio, category uncertainty, and low-level perceptual variability are controlled for. Moreover, observers tend to exclude or downweight extreme samples of perceptual evidence, as a statistician might exclude an outlying data point. These phenomena are captured by a probabilistic optimal model in which observers integrate the log odds of each choice option. Robust averaging may have evolved to mitigate the influence of untrustworthy evidence in perceptual judgments.

decision making | diffusion model | information integration

Perceptual judgments typically involve a deliberative process in which evidence concerning the current state of the external world is considered. Over recent years, the twin goals of characterizing the computational mechanisms and the neural representations underlying this deliberation have come to the fore (1, 2). Because sensory evidence coming from the external world is intrinsically noisy, decisions will benefit from repeated sampling and accumulation of the collected evidence (3–5). Mathematical modeling studies support the view that serial sampling is a basic principle of choice behavior (3, 6–9), and recent neurophysiological recordings identify the parietal cortex as a candidate site for evidence accumulation in psychophysical tasks (10–14). However, the precise computations by which a decision variable (DV) is constructed and updated during decision making remain controversial (2, 4).

One popular framework posits that integration is a simple summation process under which choices and their latencies depend linearly on the strength of sensory input (3, 15, 16). This mechanism is often illustrated by analogy to a court of law, where the jury tots up evidence for or against a guilty verdict (1). However, in a stochastic environment, committing to an action on the basis of evidence strength alone can be suboptimal, because evidence may strongly favor one option over another just by chance (17, 18). Rather, a statistically optimal policy is to base decisions on independent estimates of the strength (i.e., mean) and reliability (i.e., variance) of the currently extant sensory evidence, just as a researcher might compare two samples of data on the basis of an inferential statistic rather than merely calculating their central tendencies (19). To continue the courtroom analogy, a shrewd jury will consider not only how incriminating evidence is, but also the trustworthiness of the source of the evidence. These two factors are not necessarily correlated: For example, severely indicting evidence (e.g., an eyewitness to a crime) might originate from an untrustworthy source (e.g., a coconspirator with a vested interest), whereas mildly incriminating evidence (e.g., doubt cast on an alibi) might be offered by a highly authoritative source (e.g., official telephone records).

Although a rich psychophysical literature has sought to characterize the computations by which sensory evidence is transformed into action (20), most previous studies were poorly suited to estimating the independent contributions of evidence strength and reliability to this process, because the signal (mean) and noise (variability) are typically manipulated inversely as signal-to-noise ratio. For example, in the random dot motion paradigm, in which observers discriminate the net direction of motion of a cloud of moving dots, the average signal strength rises, and the signal variability falls, as the percentage of coherently moving dots tends toward the maximum (21, 22). The question thus arises of whether our perceptual judgments reflect only the mean of the evidence or also its reliability.

We addressed this question by considering three bounded accumulation models that compute the decision variable in distinct ways. Our first model simply accumulates the mean of the evidence on a given trial (the *simple averaging* model below). Our second model integrates the signal-to-noise ratio, scaling the mean evidence by its deviance (the *SNR model* below). Finally, the third model accumulates the log of the probability ratio between the two alternatives (the *LPR model* below), thereby converting the bounded accumulation process into a sequential probability ratio test (SPRT). The SPRT is optimal in the sense that it makes the fastest decisions for a given error rate (2, 9) and this model also proposes a DV that is known to scale with the firing rates of parietal neurons during categorization tasks (13). Comparing these models to human data allowed us to arbitrate between these different accounts of human perceptual judgments.

Here, we used a multielement averaging task (Fig. 1*A*) in which participants viewed arrays of eight simultaneously presented elements and discriminated the average value on a feature dimension (e.g., is the average color more red or blue? Is the shape more like a square or a circle?) (23, 24). This task allowed us to manipulate orthogonally the mean and the SD of the evidence presented to the subject on a given trial (Fig. S1). The mean was set either near (low mean evidence) or far (high mean evidence) from the boundary between the two response categories, and the array variance was varied at three levels (high, medium, or low evidence reliability). Critically, this paradigm also allowed us to assess the unique contribution of each element to the choice and thus to compare the contributions of *inlying* (i.e., close to the array mean) and *outlying* (i.e., far from the array mean) elements to the decision. We report two behavioral phenomena: (*i*) decision latencies are lengthened and errors increased, when arrays are more variable, even when array mean, signal-to-noise ratio, and across-trial category uncertainty are taken into account; and (*ii*) humans observers base their decision principally on inlying perceptual evidence, downweighting or excluding the outlying elements in their decision much as a statistician would discard outliers in a sample of data. Moreover, this *robust averaging* behavior is captured by a decision model in which observers accumulate not the array mean

Author contributions: V.d.G. and C.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: vincent.gardelle@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1104517108/-/DCSupplemental.

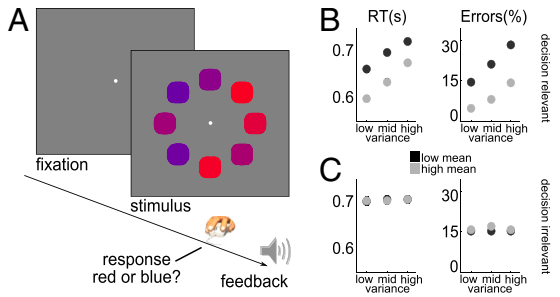

Fig. 1. (*A*) Participants discriminated the average color (red vs. blue, experiments 1, 2, and 3a) or shape (square vs. circle, experiment 3b) of eight elements arranged in a circle around fixation. Reponses were made with a button press with a deadline of 1,500 ms. Fully informative feedback was offered in every trial. (*B*) The eight elements were drawn from distributions with weak or strong evidence in favor of each option (low mean vs. high mean) and with low, medium, or high variance. Response times (*Left*) and errors (*Right*) were affected by the manipulation of mean and variance in the decision-relevant dimension. Error bars (SEM) are represented, but are typically smaller than the size of the data point, given the number of participants (total $n = 77$). (*C*) However, the same manipulation of mean and variance in the decision-irrelevant dimension did not produce the effects (data from experiment 3 only).

or the signal-to-noise ratio, but a quantity reflecting the logarithm of the probability ratio between the two possible choices.

Results

We describe four behavioral studies using different variants of the multielement averaging task. In all experiments, participants in each trial viewed an array of eight elements varying in shapes and colors. Experiments 1 and 2 required participants to judge the average color (red/blue) of ellipses with variable axes of elongation. In experiment 3, subjects judged the color (red/blue, experiment 3a) in one session and the shape (circle/square, experiment 3b) in the other session. Finally, in experiment 4, subjects judged the average color of stimuli occupying two different portions of the feature space (red/purple, purple/blue), a control manipulation allowing us to rule out low-level nonlinearities in the perceptual space as an explanation for our results.

In what follows we denote by x_k the evidence value for the element k, that is, the position of the element k relative to the boundary between the two possible choices, x_k taking positive or negative values (e.g., in the red/blue task x_k is positive if element k is more red and negative if it is bluer). We also denote by μ the mean of these evidence values over the eight elements of the array and by σ the SD of these evidence values. With these notations our manipulation of the mean strength is reflected in the absolute value of μ (noted $|\mu|$), i.e., the absolute distance between the mean evidence and the category boundary. Our manipulation of the reliability of the evidence is reflected in σ.

Effects of the Mean and the Variance in the Decision Space. Fig. 1*B* presents the effects of evidence strength ($|\mu|$) and reliability (σ) on choice accuracy and response times (RTs) on correct choices across all three experiments. As expected, in experiment 1 we found that increasing the mean evidence ($|\mu|$) led to lower error rates and shorter RTs [$F^{err}_{(1,30)} = 466.96$, $F^{RT}_{(1,30)} = 276.52$, both $P < 0.001$]. Crucially, we also observed independent effects of evidence variability σ on performance: Decisions about more variable arrays were both slower and less accurate [$F^{err}_{(2,60)} = 127.68$, $F^{RT}_{(2,60)} = 244.44$, both $P < 0.001$]. These effects were replicated in experiment 2 [effect of $|\mu|$, $F^{err}_{(1,13)} = 407.15$, $F^{RT}_{(1,13)} = 97.81$; effect of σ, $F^{err}_{(2,26)} = 127.76$, $F^{RT}_{(2,26)} = 50.37$, all $P < 0.001$] and in experiment 3 in which two different feature dimensions were used: color [effect of $|\mu|$, $F^{err}_{(1,15)} = 262.13$, $F^{RT}_{(1,15)} = 58.31$; effect of σ, $F^{err}_{(2,30)} = 78.59$, $F^{RT}_{(2,30)} = 52.35$; all $P < 0.001$] and shape [effect of $|\mu|$, $F^{err}_{(1,15)} = 402.91$, $F^{RT}_{(1,15)} = 141.1$; effect of σ, $F^{err}_{(2,30)} = 77.02$, $F^{RT}_{(2,30)} = 82.42$; all $P < 0.001$]. Importantly, trials were presented in a randomized order, so the uncertainty about category assignations (or, in signal detection theory terms, the variance of the probability density functions associated with each option) was fully equated between conditions and in particular was not confounded with our within-trial variance manipulation.

Having established the effect of within-trial variability on performance, we ran several analyses to demonstrate that this effect cannot simply be explained by (*i*) disruption of low-level perceptual processes, (*ii*) variations in signal-to-noise ratio, or (*iii*) participants adopting a strategy of counting the number of elements either side of the category boundary.

First, in experiment 3, the manipulation of variance on both the task-relevant and task-irrelevant dimensions allowed us to separate perceptual and decision-related effects of variability. Specifically, if the effect of variability is a low-level perceptual effect, we would expect for instance the color variability to affect behavior both during the color task and during the shape task. Conversely, if variability effects occur at the decision level, they should be confined to the task-relevant dimension. The results of experiment 3 argued strongly in favor of the latter view (Fig. 1*C*). Indeed, manipulating the color variance affected participants' performance during the color task [$F^{err}_{(2,30)} = 78.59$, $F^{RT}_{(2,30)} = 54.96$; both $P < 0.001$], but not during the shape task (both $P > 0.1$), leading to significant variability × task interactions [$F^{err}_{(2,30)} = 41.41$, $F^{RT}_{(2,30)} = 39.79$; both $P < 0.001$]. The same interactions were observed for the shape dimension [$F^{err}_{(2,30)} = 77.54$, $F^{RT}_{(2,30)} = 34.42$; both $P < 0.001$]: Shape variability impacted on behavior during the shape task [$F^{err}_{(2,30)} = 77.02$, $F^{RT}_{(2,30)} = 78.96$] but not during the color task (both $P > 0.43$). Together, these results confirm that the effects of σ are limited to decision-relevant evidence.

Second, using partial correlation analyses across all three experiments, we show that accuracy was influenced by variability even when signal-to-noise ratio (i.e., $|\mu|/\sigma$ for each combination of μ and σ) was accounted for. To do this, we conducted a partial regression analysis with an additional covariate encoding the signal-to-noise ratio (SNR) (i.e., $|\mu|/\sigma$ using our notation) for each trial. The main effect of variance remained reliable in the presence of this covariate ($P < 0.005$), indicating that evidence reliability influences choice even beyond the SNR. This result is also illustrated in Fig. 1*B*, where the "low variance, low mean" conditions (left black data point) and the "mid-variance, high mean" conditions (center gray data point), which have identical SNRs, give rise to significantly different levels of performance [$t_{(1,76)} > 4$, $P < 0.001$, for both the error rates and RTs].

Finally, we found similarly that choice latencies correlated with evidence variability even when the number of elements falling on the other side of the boundary was controlled for ($P < 0.005$). We conducted this analysis in a similar manner to that above, using a correlation analysis across all three experiments with a partial covariate that encoded the number of elements belonging to the category of the nonchosen option. This analysis allows us to rule out interpretation of the variability effect based on participants counting the elements on each side of the boundary.

Computational Simulations of Evidence Integration in the Multielement Task. How, then, do observers integrate information during perceptual judgments? To address this question, we adopted a well-described computational model (the drift–diffusion model) (3–5), in which choices result from noisy perceptual evidence being accumulated up to a bound:

$$\mathrm{DV}_t = \mathrm{DV}_{t-1} + A + N(0, c^2). \quad [1]$$

Specifically, the DV grows on each sample t by an increment composed of the momentary evidence A and Gaussian noise of mean 0 and variance c^2, until choices are made when the DV crossed one of two prespecified bounds Z and $-Z$ (*SI Methods*). Here, we simulated three variants of this process, in which the momentary evidence A was derived from the sensory inputs in three distinct ways.

In the *simple averaging* model, the momentary evidence was the average of the evidence carried by the n channels (here, $n = 8$). This model ignores the within-trial variance in the perceptual array. In a second model, the SNR model, the DV is updated by the ratio between the mean evidence and its SD, thus allowing within-trial variance to affect decisions. Thus, we have

$$A = \mu, \text{ for the simple averaging model} \quad [2]$$

$$A = \frac{\mu}{\sigma}, \text{ for the SNR model.} \quad [3]$$

Finally, we created a third model [the logarithm of the posterior ratio (LPR) model] that accumulated sensory input in a probabilistic space, by proposing that a stimulus value x drawn from the array is a probabilistic cue that the response category R_1 or R_2 will be reinforced and that observers should use this cue to maximize the accuracy of their decisions. In this situation, the optimal decision variable to consider is the logarithm of the likelihood ratio (LLR) (2, 9). In our simulations we used the LPR. This quantity is equal to the LLR when the two responses are of equal prior probabilities (by Bayes' rule), and it can be easily estimated from the data, from the relative frequencies of R_1 and R_2 being rewarded when x is presented (13) (for more details, see *Methods* and *SI Methods*). In this model, the decision increment is the log posterior ratio averaged over the elements presented on the current trial:

$$A = \frac{1}{n} \times \sum_{k=1}^{n} \mathrm{LPR}(x_k), \text{ for the LPR model,}$$
$$\text{where } \mathrm{LPR}(x_k) = \ln\left(\frac{p(R_1|x_k)}{p(R_2|x_k)}\right). \quad [4]$$

We fit each model to each individual participant's error rates, by letting the two free parameters (Z and a) vary over a large search space (minimizing the total mean square error between the six empirical and the six simulated error rates represented in Fig. 2). Then, we used the resulting parameters to predict RTs on correct choices (for more details, see *SI Methods*). Because experiments 1–3 yielded equivalent results, we present the simulations averaged across these experiments ($n = 77$, experiments 3a and 3b being treated as independent datasets).

The (fitted) error rates and the (predicted) RTs of the three models are shown in Fig. 2 *A–C* (model data, lines; human data, dots). As can be seen, the SNR and LPR models were able to capture the elevated decision times for the more variable condition, whereas the simple averaging model (Fig. 2*A*) was not. In fact, the simple averaging model incorrectly predicts shorter RTs when variability increases [$F^{\mathrm{RT}}_{(2,152)} = 28.55$, $P < 0.001$, ANOVA over the simulated subjects]. In situations where increasing the within-trial deviance σ also increases the variability of the mean across trials (which is true by default and true in experiments 1 and 2), these faster RTs are also predicted from previously described analytical solutions (e.g., equation 2.9 in ref. 2). We note that these analytical solutions also show that increasing the noise parameter of the diffusion reduces response latencies and thus cannot recreate the variability effect. In situations where the within-trial variability σ is decoupled from the between-trial variability of the mean (e.g., if the elements' values are pseudoselected to have identical precision of the mean across the different variance conditions, as in experiment 3) (*Methods*), then the model would predict almost constant RTs across variability conditions (in our simulations the difference between conditions was reduced to 6 ms), which still is incompatible with the empirical data. Together, these findings rule out an entire class of model in which simple summation is used to integrate incoming sensory evidence to form a perceptual decision.

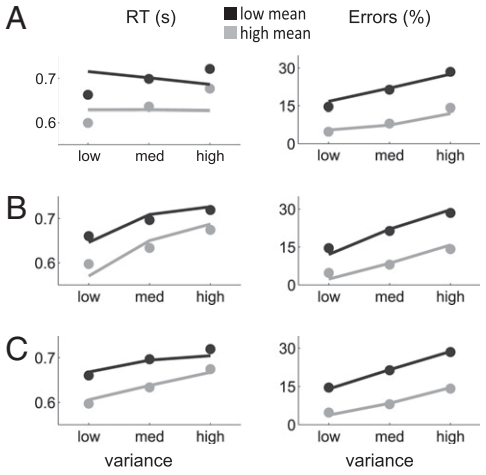

Fig. 2. Errors and reaction times for the three models. Dots indicate reaction times on correct choices (*Left*) and error rates (*Right*) for weak (low mean: black) vs. strong (high mean: gray) evidence and for low, medium, and high variability conditions (x axis). Lines show best fits to the data from (*A*) the simple averaging, (*B*) the SNR, and (*C*) the LPR models. In each case, best-fitting parameters were estimated from error rates and were used to predict RTs. Error bars (SEM) are typically smaller than the size of the dots, given the number of participants (total $n = 77$).

Robust Averaging Across Elements in the Decision Space. To distinguish between the SNR and LPR models, we turned to another aspect of the human observers' data, namely their tendency to base their decisions on inlying vs. outlying evidence within each stimulus array (evidence values being roughly Gaussian across elements within a trial). Because the SNR model scales the evidence from each element equally by the variance, it predicts that outlying elements (i.e., those that fall far from the array mean, e.g., the four extreme elements of the array according to rank) and inlying elements (i.e., those close to the array mean, e.g., the four centrally ranked elements of the array) should make equivalent contributions to the eventual choice. By contrast, the LPR model, in which evidence values are compressed at the extremes, predicts that the contribution of outlying elements should be compromised relative to inlying elements.

To arbitrate between these possibilities, we used logistic regression to estimate the weights associated with each element of the array ranked by its feature value. Across all three experiments, the decision tuning function across ranks exhibited an inverted-u shape [Fig. 3*A*: effect of rank on the beta: $F_{(7,76)} = 14.67$, $P < 0.001$], indicating that the contribution of outlying evidence was muted in the eventual decision. Crucially, this behavior was predicted by the LPR model (Fig. 3*B*), but neither by the simple averaging model nor by the SNR model (both $P > 0.25$; Fig. S2). The failure of these latter models to capture this effect not only allows us to distinguish between them as accounts of perceptual integration, but also provides reassurance that the observed shape

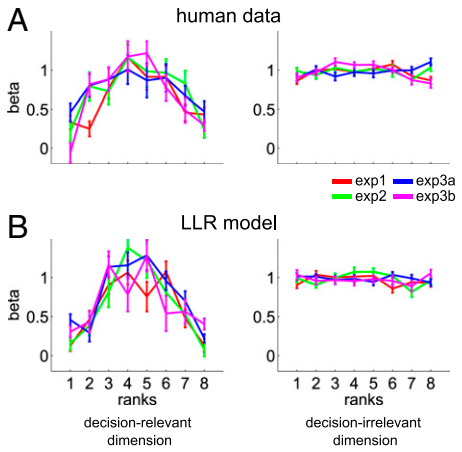

Fig. 3. Weighting of evidence across elements. (*A*) Weighting functions (normalized regression coefficients) across item rank for human data for each of experiments 1–3, when elements are sorted within each trial as a function of their value on the decision-relevant dimension (*Left*) or the decision-irrelevant dimension (*Right*). (*B*) Weighting functions for the LPR model.

of the decision tuning function is not a trivial function of our analysis strategy.

This downweighting of outlying elements was more prominent for the most variable trials, a feature that was also captured by the LPR model [Fig. 4*A*, bars; human data, lines, LPR model; interaction between inlying vs. outlying and variability, $F_{(2,152)} = 19.15$, $P < 0.001$]. The relation between downweighting and variability was also manifest across subjects: We observed a significant correlation between the difference in weights for inlying vs. outlying elements and the magnitude of the variability effect on RTs (Fig. 4*B*; correlation, $r = 0.30$, $P < 0.005$). A robust regression analysis of the downweighting effect on the variability effect revealed a positive slope ($P < 0.005$), confirming that this relation was not driven by outliers in the cohort. The downweighting effect was, however, not correlated with the effect of mean evidence on RTs ($r = 0.07$, $P > 0.5$), showing that downweighting outlying elements is specifically associated with observers being sensitive to the variability of the evidence.

Importantly, we ensured that the decision tuning functions reflected the elements' positions in the decision space, not in the stimulus space. Indeed, performing the logistic regressions in the different mean evidence conditions, we found that the weighting profile across ranks tracked the mean evidence of the *current* trial, rather than simply the mean across the entire experiment (Fig. 4*C* and Fig. S3), another feature of the data predicted by the LPR model (Fig. 4*D*). This result could not be the case if the weighting was due to the stimulus information only. Additionally, when ranked according to the decision-irrelevant dimension, the elements' decision-relevant values exhibited a flat weighting profile (Fig. 3*A*, *Right*), which further demonstrates that the contribution of an element to the choice is determined only by its position in the decision space and that the inlying vs. outlying elements are defined only in the decision space. Because we also show that our effects occur across two different, arbitrarily defined feature spaces, shape and color, it is very unlikely that they reflect low-level nonlinearities in perceptual processing.

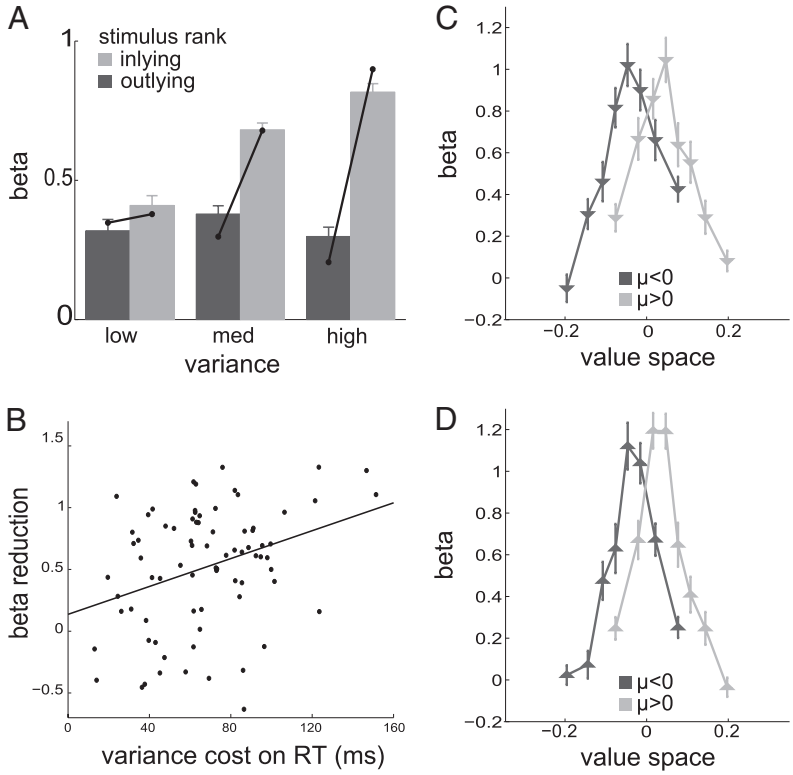

Fig. 4. (*A*) Weighting functions (normalized regression coefficients) across item rank, averaged for outlying (ranks 1, 2, 7, and 8) and inlying (ranks 3, 4, 5, and 6), for human data (bars), and the LPR model (lines), in the different variance conditions. Inlying elements received greater weights than outlying elements (main effect of inlying vs. outlying: $P < 0.001$) and even more so when variability was higher (interaction with variability: $P < 0.001$). (*B*) The variance cost on RTs (*x* axis) and the reduction in weights for outlying compared with inlying elements (*y* axis) were correlated across all subjects ($r = 0.30$, $P < 0.005$). (*C*) Weighting functions for human data for evidence favoring left ($\mu < 0$, black) or right ($\mu > 0$, gray) choices. The *x*-axis positions correspond to the average value of the items (in the eight ranks) in the decision space. (*D*) Weighting functions computed from the choices of the LPR model, for the same division of trials.

Nevertheless, to further demonstrate this point, we carried out an additional control (experiment 4) in which participants had to average information drawn from two distinct portions of the color space (i.e., red vs. purple, purple vs. blue) in different sessions. Computing decision tuning functions separately for each session allowed us to ask whether participants downweighted information with a particular value in perceptual space (e.g., they distrusted extreme values of red or blue) or whether they downweighted information at the extremes of the decision space (e.g., they distrusted the extremes of the decision-relevant color space). Specifically, we tested the hypothesis that perceptually identical elements would be assigned different decision weights as the task-relevant feature space was shifted. Applying the same logistic regression analyses separately on the two tasks, the decision tuning functions obtained in experiment 4 (Fig. S4) provided straightforward evidence that participants based their decision more on inlying elements than on outlying elements, irrespective of where the elements fall in our color space [inlying vs. outlying: $t_{(23)} > 4$, $P < 0.001$ for both tasks].

Discussion

Perceptual judgments result from a deliberative process in which an observer gathers and integrates evidence from the external world. Here we used a multielement averaging task to shed light on the integration rule used by human observers during perceptual decisions. Manipulating the mean strength and the reliability of the decision-relevant information, we found that decision accuracy was compromised both by reducing the strength of the currently available evidence (i.e., by manipulating distance to the categorical boundary) and by decreasing its reliability (i.e., by manipulating the variability of evidence within a single trial). Moreover, we found that observers based their choices more on *inlying* evidence (that falling close to the mean of the evidence on that trial) than on *outlying* evidence (that falling at the extremes). This result implies that observers are engaging in robust averaging during perceptual judgment, i.e., they are excluding or downweighting the less trustworthy elements in the array.

Measuring choice latencies as well as choices allowed us to arbitrate among three competing accounts of how observers were performing the task, each embodying a distinct hypothesis about the computational rule by which perceptual decisions pool multiple sources of evidence. We found that the model that best describes the human data, and the only model that produced the robust averaging behavior, is one in which the accumulated evidence (the DV) scales with the LPR of the two perceptual options. This is interesting for several reasons. First, previous work in the psychology of categorization has demonstrated participants can learn to assign appropriate weights to discrete dimensions of a category exemplar when these contribute in unequal measure to the decision (25–27). Our results complement this finding by pointing to distortions that arise in the representation of continuous evidence along a single dimension (e.g., color). Second, our work contributes also to a literature that suggests that perceptual systems encode the summary statistics associated with a visual array or scene (25, 26). Our data suggest that participants actively use these summary statistics in deciding what is present in the external world. Third, neurophysiological recordings during a sequential integration task have shown that the responses of neurons in the parietal cortex scale with the log-likelihood ratio during perceptual decisions (13).

In the LPR model used here, the variability of the current perceptual evidence is not explicitly represented. Rather, robust averaging is a natural consequence of the learning mechanisms occurring in the stimulus–action–outcome frame of reference, i.e., the associative mechanisms linking portions of the decision space to each of the available options. Compression of the evidence values at the extremes of the decision space is a natural consequence of the roughly sigmoid shape of the posterior probability function. It is interesting that an optimal framework for integrating evidence from multiple sources can be approximated with a simple, neurobiologically plausible mechanism such as this. Nevertheless, the work described here does not rule out accounts in which observers would explicitly represent the variance in the external world. For example, a fully Bayesian model that attempts to recover the generative mean and variance of the current array is likely to capture many of the behavioral effects described here, albeit in a computationally more demanding fashion, and arguably with less neurobiological plausibility. Indeed, a recent brain imaging study showed that in reward-guided decisions, the anterior cingulate cortex may track the reliability of recent outcomes and adjust learning accordingly (18). An important question for further research is whether distinct neural correlates of the mean and variance of the evidence can be identified using functional neuroimaging. For instance, neurons in the macaque parietal cortex exhibit firing rates that scale with the expected value of a stimulus (14), including in tasks where successive stimuli contribute to a decision (13). In humans, successive evidence accumulation has been associated with the basal ganglia (27–30) or parieto-central loci (31). However, the precise determinants of these neural responses remain unclear.

To conclude, we have found that when human observers are required to consider multiple sources of evidence to form a perceptual judgment, they discard outlying decision-relevant values when evidence is variable. Control analyses demonstrate this robust averaging phenomenon is specific to the decision variable and not driven by low-level perceptual characteristics of the stimuli or local nonlinearities in the perceptual space (e.g., categorical perception of color). This downweighting strategy might seem somewhat counterintuitive in a task requiring the observer to make a judgment about the mean. However, from a statistical perspective this behavior might be more reliable than pure averaging: When sensory signals from a single source are corrupted by variability, extreme values arise that poorly indicate the true underlying source. Thus, compression of outlying values during integration of evidence may have evolved for our decision making to be resistant to grossly erroneous or irrelevant information. Indeed, an agent who bases decisions on the mean evidence alone, with no regard for its reliability, will be prone to impulsive choices based on noise mistaken for signal (17). Together, the present findings reveal that human perceptual decision making is *robust*, with outlying evidence being downweighted, much as a researcher might choose to exclude an observation that differs radically from the rest of the sample (32).

Methods

Participants. All four experiments took place in the Department of Experimental Psychology at University of Oxford. Subjects (number of participants: $N_1 = 31$, $N_2 = 14$, $N_3 = 16$, $N_4 = 24$) were students recruited from the University of Oxford (age range: 18–25). They reported normal or corrected vision and no history of neurologic or psychiatric illness. Experiment 3 was divided into two sessions (3a and 3b) occurring on different days (counterbalanced order). Experiment 4 was divided into two sessions (4a and 4b), run one after the other (counterbalanced order) on the same day. Participants provided written consent before the experiment and were reimbursed (£10/h) for their participation. All experiments were approved by the local ethics committee.

Stimuli. The stimulus array was constituted by eight elements positioned on a circle around fixation (radius ~3° visual arc) at regularly spaced angular positions (Fig. 1). Stimuli were generated using the PsychToolBox (www.psychtoolbox.org) for MATLAB (Mathworks) and presented on a 17″ CRT screen (resolution: 1,024 × 768) viewed from a distance of 60 cm. Each element was defined by two parameters, a "color value" C and a "shape value" S, manipulated independently and taking values between 0 and 1. For colors, the parameter C defined the color of the element in Red-Green-Blue (RGB) values between red ([1, 0, 0]) and blue ([0, 0, 1]) by following a linear transition in RGB space ([C, 0, 1 − C]). In experiments 3 and 4, we also applied gamma correction to the obtained values. For shapes, in experiments 1 and 2

the shapes were ellipses with vertical and horizontal main axes, elongated either vertically or horizontally, by using the parameter S to transform an original circle (width = height = 50 pixels) by adding the quantity $50 \times (S - 0.5)$ in pixels to its width and subtracting it from its height. In experiments 3 and 4, we used superelliptic shapes (a.k.a. squircles), which provide a parameterization between squares and circles. The parameter S was converted into a value n (simply, $n = S + 1$) corresponding to the curvature of a superellipse with equal semidiameters (a = 25 pixels), whose contours (x, y) are mathematically defined by the following equations:

$$\forall\theta \in [0, 2\pi],\ x(\theta) = a \times \text{sign}(\cos(\theta)) \times |\cos(\theta)|^n \text{ and}$$
$$y(\theta) = a \times \text{sign}(\sin(\theta)) \times |\sin(\theta)|^n.$$

Varying n in [1, 2] allowed us to create shapes varying gradually between a square and a circle (see *SI Methods* for an illustration). We further ensured that all elements occupied the same area on the screen by correcting the semidiameter a using the following relation:

$$\text{Area} = 4 \times a^2 \times \left(\Gamma\left(1 + \frac{1}{n}\right)\right)^2 \times \left(\Gamma\left(1 + \frac{2}{n}\right)\right)^{-1},$$

where Γ is the gamma function.

Task and Design. The participants' task was to judge whether the average color of the eight elements currently on the screen was more red/blue (experiments 1, 2, and 3a), red/purple (experiment 4a), or purple/blue (experiment 4b) or whether the average shape of the elements was more square/circle (experiment 3b). Participants indicated their response using a two-buttons mouse with their preferred hand and received auditory feedback: two ascending tones (400–800 Hz, 100 ms each) and descending tones (800–400 Hz, 100 ms each) followed correct responses and errors, respectively. The correct response on the current trial was defined by the generative mean of the eight values for the task-relevant parameter. Each participant ran a first training block (100 trials), followed by 6–10 experimental blocks, in which all conditions were presented in a randomized order. All sessions lasted ~1 h and corresponded to ~1,000 trials.

Color and shape parameters were defined in the interval [0, 1] and generated by trimming random samples from normal distributions with prespecified means (mc, ms) and SDs (σc, σs). In experiments 3 and 4 we ensured that the resulting means and SD matched the predefined ones (tolerance 0.1%), by resampling when necessary. Typically, the SDs were varied in three levels (low, medium, and high), and the means were varied in four levels, corresponding to low vs. high mean evidence for the two response categories (e.g., for the red/blue task: really red vs. slightly red vs. slightly blue vs. really blue). Numerical values for the SDs were in the [0.05, 0.15] range, and for the means they varied in the [−0.1, 0.1] range around the value of the category boundary. For experiments 1–3 (red/blue task or square/circle task), the boundary between the response categories was in the middle of the parameter space (i.e., 0.5), but for experiment 4 we used a boundary at 0.25 (purple/blue task) or 0.75 (purple/red task). Specific values for all parameters in all experiments are given in *SI Methods*.

LPR Model. This model is an adaptation of the LLR model, already described in previous studies (e.g., ref. 13), to our case of multiple elements. To simulate this model, we first converted the elements' individual values into log posterior ratios. To do so, we considered all elements presented to the subject as associated with either R_1 or R_2 being the correct response (given the feedback). Then, we computed the posterior probability of R_1 on 100 bins regularly spaced on the parameter space (after discarding 5% of extreme values for which the probability estimation involves very few data points): In each bin, we computed the frequency of R_1 being the correct association for the elements falling in that bin. We fitted a sigmoid function of the bins to the resulting probabilities (*SI Methods* and Figs. S5 and S6), which we then applied to each element. Finally, we converted the obtained posterior probabilities in log odds [i.e., transforming $P \rightarrow$ into $\log(P/(1 - P))$] to get the LPR. This procedure was carried out for each subject separately.

The sum of the LLR is the optimal decision variable to consider for a bounded accumulation model (2, 9). Also, because in our models the noise, bounds, and increments scale together (see ref. 2), taking the sum is equivalent to taking the average over the eight elements. We favored the average to express all three models in similar forms. Importantly, the optimality depends on the elements being sampled independently. Thus, the LPR model represents the optimal process for an observer considering (inaccurately) the elements as if they were independent (*SI Methods*).

1. Gold JI, Shadlen MN (2007) The neural basis of decision making. *Annu Rev Neurosci* 30:535–574.
2. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* 113:700–765.
3. Ratcliff R (1978) A theory of memory retrieval. *Psychol Rev* 85:59–108.
4. Ratcliff R, McKoon G (2008) The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput* 20:873–922.
5. Smith PL, Ratcliff R (2004) Psychology and neurobiology of simple decisions. *Trends Neurosci* 27:161–168.
6. Link SW, Heath RA (1975) The relative judgement theory of two choice response times. *J Math Psychol* 12:114–135.
7. Audley RJ, Pike AR (1965) Some alternative stochastic models of choice. *Br J Math Stat Psychol* 18(2):207–278.
8. Stone M (1960) Models for choice reaction time. *Psychometrika* 25:251–260.
9. Wald A, Wolfowitz J (1949) Bayes solutions of sequential decision problems. *Proc Natl Acad Sci USA* 35:99–102.
10. Kiani R, Hanks TD, Shadlen MN (2008) Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *J Neurosci* 28:3017–3029.
11. Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J Neurosci* 22:9475–9489.
12. Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J Neurophysiol* 86:1916–1936.
13. Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. *Nature* 447: 1075–1080.
14. Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400:233–238.
15. Carpenter RH, Williams ML (1995) Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377:59–62.
16. Brown S, Heathcote A (2005) A ballistic model of choice response time. *Psychol Rev* 112:117–128.
17. Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. *Neuron* 46: 681–692.
18. Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214–1221.
19. Cox RT (1946) Probability, frequency and reasonable expectaion. *Am J Phys* 14:1–13.
20. Parker AJ, Newsome WT (1998) Sense and the single neuron: Probing the physiology of perception. *Annu Rev Neurosci* 21:227–277.
21. Newsome WT, Britten KH, Salzman CD, Movshon JA (1990) Neuronal mechanisms of motion perception. *Cold Spring Harbor Symp Quant Biol* 55:697–705.
22. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J Neurosci* 12: 4745–4765.
23. Ariely D (2001) Seeing sets: Representation by statistical properties. *Psychol Sci* 12: 157–162.
24. Chong SC, Treisman A (2003) Representation of statistical properties. *Vision Res* 43: 393–404.
25. Alvarez GA (2011) Representing multiple objects as an ensemble enhances visual cognition. *Trends Cogn Sci* 15:122–131.
26. Alvarez GA, Oliva A (2009) Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proc Natl Acad Sci USA* 106:7345–7350.
27. Bogacz R, Wagenmakers EJ, Forstmann BU, Nieuwenhuis S (2010) The neural basis of the speed-accuracy tradeoff. *Trends Neurosci* 33:10–16.
28. Forstmann BU, et al. (2008) Striatum and pre-SMA facilitate decision-making under time pressure. *Proc Natl Acad Sci USA* 105:17538–17542.
29. Gluck MA, Poldrack RA, Kéri S (2008) The cognitive neuroscience of category learning. *Neurosci Biobehav Rev* 32:193–196.
30. Poldrack RA, et al. (2001) Interactive memory systems in the human brain. *Nature* 414: 546–550.
31. de Lange FP, Jensen O, Dehaene S (2010) Accumulation of evidence during sequential decision making: The importance of top-down factors. *J Neurosci* 30:731–738.
32. Huber PJ (1981) *Robust Statistics* (Wiley, New York).