nature human behaviour

LETTERS

PUBLISHED: 10 JULY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0139

Perceptual confidence neglects decision-incongruent evidence in the brain

Megan A. K. Peters[1*†], Thomas Thesen[2,3,4†], Yoshiaki D. Ko[5†], Brian Maniscalco[6], Chad Carlson[2,7], Matt Davidson[5], Werner Doyle[2], Ruben Kuzniecky[2], Orrin Devinsky[2], Eric Halgren[3] and Hakwan Lau[1,8]

Our perceptual experiences are accompanied by a subjective sense of certainty. These confidence judgements typically correlate meaningfully with the probability that the relevant decision is correct[1–6], bolstering prevailing opinion that both perceptual decisions and confidence optimally reflect the probability of having made a correct decision[6–13]. However, recent behavioural reports suggest that confidence computations overemphasize information supporting a decision, while selectively down-weighting evidence for other possible choices[14–19]. This view remains controversial, and supporting neurobiological evidence has been lacking. Here we use intracranial electrophysiological recordings in humans together with machine-learning techniques to demonstrate that perceptual decisions and confidence rely on spatiotemporally separable neural representations in a face/house discrimination task. We then use normative computational models to show that confidence relies excessively on evidence supporting a decision (for example, face evidence for a 'face' decision), even while decisions themselves reflect the optimal balance of all evidence (for example, both face and house evidence). Thus, confidence may not reflect a readout of the probability of being correct; instead, observers may sacrifice optimality in favour of self-consistency[20] in the face of limited neural and computational resources. Although seemingly suboptimal, this strategy may reflect the inference problem that perceptual systems are evolutionarily optimized to solve.

We recorded cortical electrophysiological signals (ECoG) from epilepsy patients with surgically implanted intracranial electrodes as they distinguished degraded faces from houses at two contrast levels and provided binary confidence judgements by pressing buttons on a keyboard (Fig. 1a). Subjects performed at an intermediate level of accuracy in their perceptual decisions (81.0% correct), as expected from performance thresholding procedures, which provided the opportunity to analyse and compare correct and incorrect decisions at different levels of subjective confidence.

Subjects rated confidence meaningfully, tracking their own decision accuracy rather than just stimulus contrast (Fig. 1b), and had faster reaction times for high-confidence versus low-confidence responses (mean reaction time $\mu_{\text{high-confidence}} = 1{,}059$ ms, $\mu_{\text{low-confidence}} = 1{,}439$ ms, Student's $t(5) = 4.32$, $P = 0.007$) but not for high-contrast versus low-contrast trials ($\mu_{\text{high-contrast}} = 1{,}096$ ms, $\mu_{\text{low-contrast}} = 1{,}132$ ms, $t(5) = 1.96$, $P = 0.11$). Subjects also showed little response bias to respond 'face' versus 'house' (Fig. 1b).

Following previous work that has shown that activity in the high-gamma frequency range (80–120 Hz) reflects the most relevant neuronal activity[21–27] specifically regarding perceptual processes[28–32], we focused further analyses on this frequency range. The mean time-frequency spectrum averaged over all subjects, electrodes and trials was indeed most salient in this high-gamma range, centred around 250–400 ms after stimulus onset (Supplementary Fig. 1), congruent with previous reports[33–36]. Because we confirmed that including a much wider range of frequency bands did not alter the qualitative pattern of the main results (and only very slightly altered them quantitatively; see Supplementary Results: Frequencies outside 80–120 Hz), this focus also helps to keep the computational demands for decoding analysis manageable and to avoid overfitting.

We used machine-learning classification (support vector machine; SVM) to decode two behavioural factors: perceptual 'decision' (face/house) and 'confidence' (high/low). 'Features' for SVM decoding were defined as each electrode's normalized power at a particular frequency band and particular timepoint in the peri-stimulus window (Supplementary Methods: Support vector machine decoding).

We were able to decode both behavioural factors above chance at different time bins after stimulus onset (Fig. 2a). Chance level was defined with permutation tests (see Supplementary Methods: Support vector machine decoding), and was found to be 0.5001, justifying our use of 0.5 as chance level decodability. 'Decision' decoding reached above-chance levels for at least half of subjects beginning at 250 ms, but 'confidence' decodability did not reach significance for half of subjects until 450 ms (Supplementary Table 3). Importantly, both factors were able to be decoded above chance well before any movement onset (mean reaction time = 1,136 ms), suggesting that decoding is not based on movement (finger movement preparation can typically be decoded only up to ~200 ms before movement onset with ECoG[37]; see also Supplementary Results: Motor preparation and neuroanatomical localization results).

The above results suggest that 'decision' and 'confidence' behaviours may reflect different evidence at different time points. One could argue that this dissociation may be trivial, as it is generally accepted that metacognitive representations arise later than those underlying perceptual decisions[38,39] and may decay over time[40]. Although, in our experiment, subjects made both the decision and confidence responses simultaneously by a single button press, one could argue that in their minds they might have done it sequentially because it would be natural to do so.

[1]Department of Psychology, University of California, Los Angeles, Los Angeles, California 90095, USA. [2]Comprehensive Epilepsy Center, Department of Neurology, New York University Medical Center, New York, New York 10016, USA. [3]Multimodal Imaging Laboratory, University of California, San Diego, La Jolla, California 92037, USA. [4]Department of Physiology & Neuroscience, St. George's University, Grenada, West Indies. [5]Department of Psychology, Columbia University, New York, New York 10027, USA. [6]Neuroscience Institute, New York University, New York, New York 10016, USA. [7]Department of Neurology, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. [8]Brain Research Institute, University of California, Los Angeles, Los Angeles, California 90095, USA. †These authors contributed equally to this work. *e-mail: meganakpeters@ucla.edu

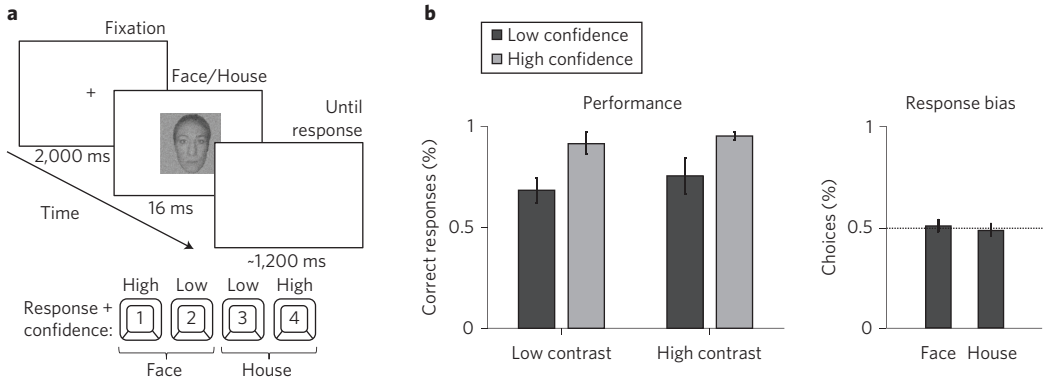

Figure 1 | Behavioural task and results. a, Subjects discriminated noisy stimuli as faces/houses and indicated their confidence (high versus low) with a single button press; responses were all made with one hand. **b**, As expected, subjects showed higher accuracy for high-contrast versus low-contrast stimuli, and for high-confidence versus low-confidence responses (2 (contrast: high/low) × 2 (confidence: high/low) repeated measures ANOVA: $F(1,5)_{confidence} = 8.418$, $P = 0.034$; $F(1,5)_{contrast} = 1.783$, $P = 0.239$; $F(1,5)_{confidence \times contrast} = 0.502$, $P = 0.10$), but showed negligible bias to respond 'face' more often than 'house' ($t(5) = 0.316$, $P = 0.765$). Error bars represent the standard error of the mean across subjects.

Therefore, we also directly assessed the spatial separation of the representations' neural correlates[41–45]. We quantified each electrode's contribution to decodability by calculating a normalized contribution index (C) (Supplementary Methods: equations (S1) and (S2)), which we projected onto its MNI coordinates averaged across coarse 200-ms time bins to reveal broad patterns (see Supplementary Methods: Neuroanatomical localization of representations) (Fig. 2b). We also averaged C across electrodes within the four neocortical lobes—frontal (36.0% of all electrodes), parietal (24.2%), temporal (33.8%) and occipital (6.0%)—and plotted C for each lobe as a function of time after stimulus onset (Fig. 2c) (see also Supplementary Figs 6–8, and Supplementary Tables 4 and 5).

Occipital regions showed localized contributions to decision starting at 200–400 ms despite their sparsity in electrode numbers, but confidence appears to be more neuroanatomically distributed (significant main effects of lobe for decision ($F(3,870) = 7.748$, $P < 0.001$) but not confidence ($F(3,870) = 1.896$, $P = 0.129$); Supplementary Results: Representational overlap) with marked contributions from parietal[6] and frontal areas[2,46–49] (Fig. 2b,c). Note that the separability of decision and confidence representations does not mean that there is no overlap at all. In terms of simple response level (rather than decodability), there are individual electrodes that showed some sensitivity to both decisions and confidence judgements, although they did so in ways also congruent with our central hypotheses (Supplementary Results: Representational overlap; Supplementary Fig. 9). Overall, this analysis of separable contributions to decision and confidence confirms that our measure of decoding contribution by lobe is not due to trivial overrepresentation of electrodes: if a lobe's decoding contribution were statistically biased because of electrode density, then denser regions (frontal and temporal) should have shown the highest decoding contributions and occipital the lowest. This analysis also provides additional evidence that decision and confidence decoding was unlikely to be due to trivial decoding of movement: if estimators decoded movement preparation only, one should not expect strong and early contributions of occipital electrodes.

The dissociations in spatial representation correlates and decodability timecourse for decisions and confidence suggest that confidence computations may not rely on the same internal evidence as decisions. One possible hypothesis[14–19] is that decisions are based on the 'balance of evidence' between decision-congruent and decision-incongruent evidence on each trial, but confidence relies on decision-congruent evidence only[14–19]. For example, if subjects indicate a 'face' decision, their confidence judgement will reflect the strength of the neural evidence for 'face' but will be largely insensitive to the (lack of) evidence for 'house'. Although this hypothesis has received some support from behavioural studies[14–19], it remains controversial, with some researchers arguing that confidence judgements reflect an optimal readout of the same information that led to the decision[1–13]. Moreover, whereas previous studies concerned whether subjects may ignore decision-incongruent evidence provided by the physical stimuli, here we addressed the intriguing possibility that such evidence may be available in the brain at the time of the confidence computation, and yet the relevant neural mechanisms fail to make use of such information.

To evaluate this hypothesis, we trained an additional neural decoder on the stimulus presented on each trial and extracted the 'weights' assigned to each feature (electrode–frequency–time-point; Supplementary Methods: Support vector machine decoding). We combined these weights with each feature's power to define 'evidence'—that is, how much the neural code reflected both the 'face-ness' and 'house-ness' of the stimulus on each trial (Methods: Choice probability analysis, equations (1) and (2))—and categorized evidence depending on the subject's decisions: face evidence is decision-congruent on trials in which subjects responded 'face' but decision-incongruent when they responded 'house', and vice versa for house evidence.

We then computed the choice probability (CP)[50] for 'balance-of-evidence' versus 'decision-congruent-only' rules: on a trial-by-trial basis for each subject, we assigned decisions and confidence judgements as hits and false alarms according to standard receiver-operating-characteristic (ROC) methods[51], and calculated the area under the curve (AUC) to obtain CP values. The degree to which CP > 0.5 therefore indicates how well a given rule (balance-of-evidence or decision-congruent-only) can be used to correctly predict the relevant behaviour (decision or confidence), based on the neural evidence (Methods: Choice probability analysis).

CP was significantly above chance for both decision and confidence (Supplementary Table 6) for both computation rules, but statistical tests also revealed an interaction between decision/confidence and computation rule (2 (predictor: decision, confidence) × 2 (evidence: balance, decision-congruent) repeated-measures ANOVA: no main effect for predictor ($F(1,5) = 4.538$, $P = 0.086$), main effect for evidence ($F(1,5) = 9.665$, $P = 0.027$), and significant interaction between predictor and evidence ($F(1,5) = 6.961$, $P = 0.046$)) (Fig. 3a,b). This interaction occurred because, as hypothesized, subjects used balance-of-evidence to compute decision, but balance-of-evidence and decision-congruent-only CPs were indistinguishable when computing confidence (paired two-tailed t-tests; decision: $t(5) = 17.7044$, $P < 0.001$; confidence: $t(5) = 0.6719$,

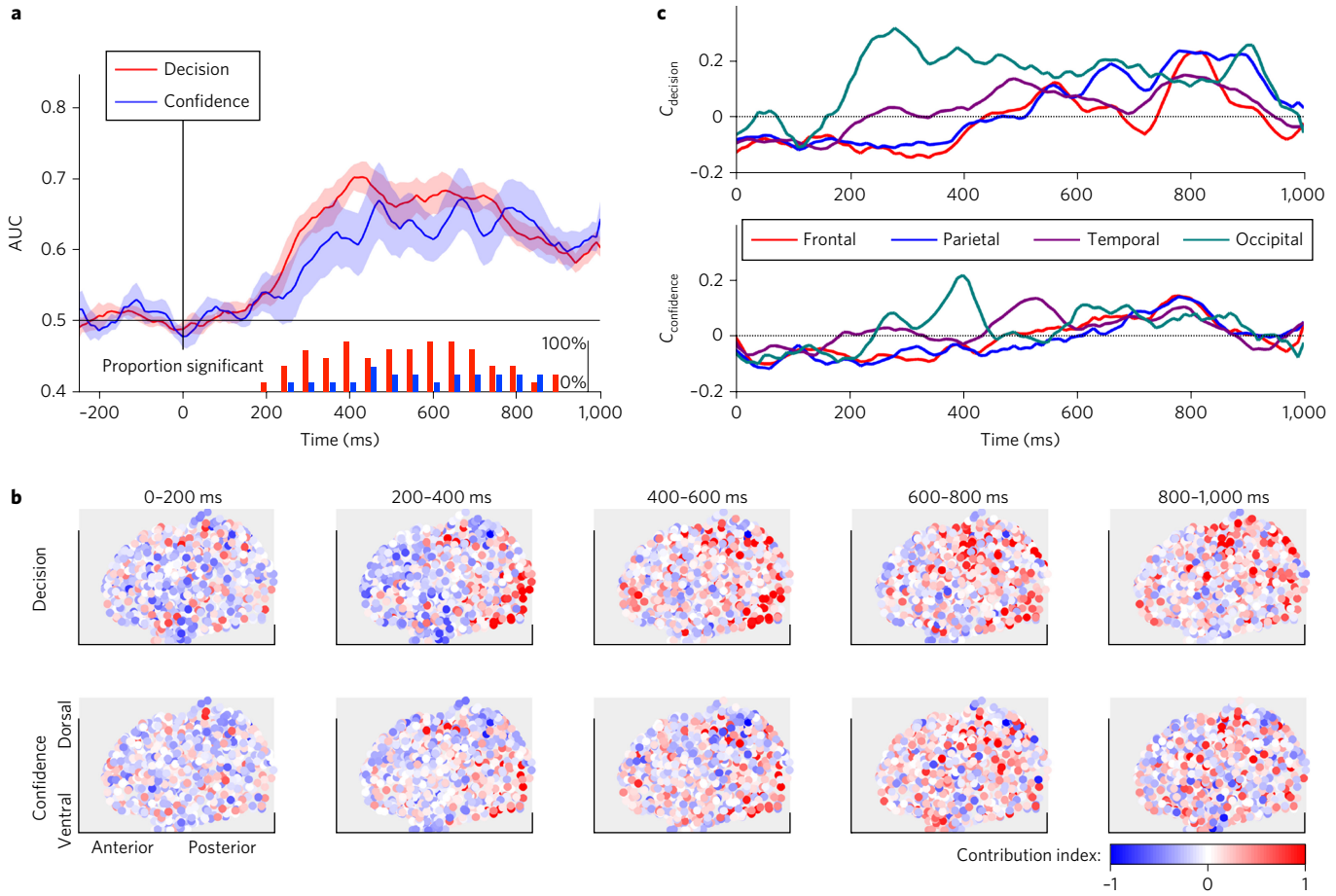

Figure 2 | Spatiotemporal dissociation between 'decision' and 'confidence' decoding. a, Decoder accuracy for both estimators (decision and confidence) rises just around 200 ms after onset of the stimulus. However, decodability for decision rises more quickly and peaks earlier than for confidence. Shaded regions indicate the standard error of the mean. Lower bars denote 50-ms post-stimulus time bins in which decodability was above chance for some proportion of participants. **b**, To localize factors contributing to decoding performance, we projected each electrode's contribution index *C* (see Supplementary Methods: Neuroanatomical localization of representations) onto its MNI coordinates across all subjects, averaged across coarse time bins of 200 ms. Contribution index $C < 0$ (blue) indicates that the electrode contributed very little, whereas $C > 0$ (red) indicates that the electrode contributed more to decoding. **c**, We calculated average *C* within four broadly defined regions of interests by lobe, and plotted it as a function of time after stimulus onset. Decision shows strong contributions from occipital electrodes around 200–700 ms, whereas confidence occupies a more distributed spatial representation.

$P = 0.531$) (Fig. 3a,b). This means that taking into account decision-incongruent evidence does not help to better predict confidence rating behaviour even though it had exactly this effect for decisions, as if subjects relied nearly exclusively on decision-congruent evidence alone when judging confidence even though they incorporated decision-incongruent evidence to calculate their decisions.

The CP analyses provide support for the hypothesis that confidence computations disproportionately ignore decision-incongruent evidence, in agreement with the finding that electrodes' simple response level also reflects confidence in a decision-congruent manner (Supplementary Fig. 9). However, one could argue that the lack of improvement in predicting confidence via including decision-incongruent evidence is essentially a null result. In principle, the significant interaction between computation rule and decision/confidence addresses this concern, but perhaps confidence is supported by a more complex process than decision, and therefore it is more difficult to achieve high CP given the noisiness of data; we might have reached the noise ceiling for confidence, which would lead to the false appearance of a lack of improvement when decision-incongruent evidence was also included.

To address this concern, we used the simple framework of signal detection theory (SDT)[51,52] to build a normative forward model, and to formally assess the noise ceiling stipulated by the decodability of the data. Assuming that subjects are Bayesian ideal observers, their confidence should be monotonically related to accuracy[4]: that is, it should optimally reflect the probability of a decision's being correct on a trial-by-trial basis[7–12] (Fig. 4a; see Methods: SDT forward model). Therefore, both trial-by-trial accuracy and confidence should depend on similar calculations; they can both be thought of as the distance of some internal decision variable x from a decision criterion (Fig. 4a). With this simple model, we can thus formally relate the decodability of the decision response, accuracy and confidence, and compare the observed data to the model.

The fact that we cannot decode decision at 100% accuracy means there must be noise inherent in the data, the measurement and decoding technique, and so on. We empirically assessed this noise level, $\alpha_{decoding}$, for each subject based on decision decodability, which would be 100% if $\alpha_{decoding} = 0$ according to SDT (Fig. 4a). Based on the observed level of decoding noise ($\alpha_{decoding}$), we estimated the theoretically maximal expected decodability for both accuracy and confidence (Fig. 4b; see Methods: SDT forward model). We then compared this expected maximum to actual data (that is, decodability of accuracy and confidence by the forward model, based on all available features).

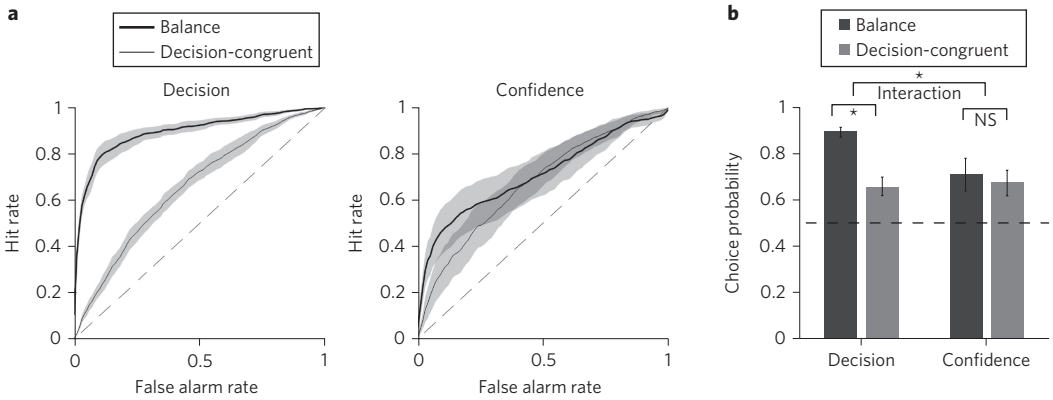

Figure 3 | Choice probability analyses show that confidence computations were insensitive to decision-incongruent evidence. a, Differences in 'decision' versus 'confidence' representations mapped onto differential use of decision-congruent evidence versus decision-incongruent evidence for decision and confidence computations. **b**, Decision and confidence were predicted differentially by the balance-of-evidence rule than by the decision-congruent-only rule: decision was significantly better predicted by balance-of-evidence, but confidence showed no difference between balance-of-evidence and decision-congruent-only computation rules. This indicates that the computation of confidence overly relied on the magnitude of decision-congruent evidence and did not appear to make use of decision-incongruent evidence NS, not significant.

Indeed, statistical tests confirmed that accuracy decodability achieved by the model was indistinguishable from the theoretical maximum given noise (α_{decoding}), but confidence decodability was significantly worse than the theoretical maximum (Fig. 4c,d). This finding indicates that the computation of confidence must differ in efficiency from the computation of the decision[8,11], and therefore cannot optimally reflect the probability of being correct (accuracy)[8–12]. Crucially, that there was no problem in predicting accuracy optimally given the observed noise level means the decision-incongruent evidence was available in the brain, and yet under-used in the computation of confidence.

One may worry that the detection theoretical model failed because of the different timecourses of information flow for decision (Type 1) and confidence (Type 2) judgements[38,39]. We addressed this concern by conducting temporal generalization analysis[53], which evaluates whether the decision estimator trained at time t can decode confidence at some other time t' (especially after t). However, we saw no evidence for temporal dissociations that could have led to the model's failure (see Supplementary Results: Lag in predicting from decision to confidence?; Supplementary Fig. 11). This analysis demonstrates the informativeness of neural signals in evaluating the SDT model; without neural information, it would have been difficult to ensure that the model's failure was not due to differences in processing timecourse between decision and confidence.

Finally, one might argue that although the decoding noise ceiling was reached, the CP analysis still failed to demonstrate that the decision-congruent-only rule can predict confidence better than the balance-of-evidence rule. To address this concern formally, we capitalized on Bayesian generative model simulations to compare directly how well a balance-of-evidence ideal observer[54] and decision-congruent-only heuristic observer[16] could predict subjects' confidence (Supplementary Methods: Generative Bayesian models). We fed the trial-by-trial evidence (equations (1) and (2)) as two-dimensional data points $x = [\text{evidence}_{\text{face}}, \text{evidence}_{\text{house}}]$ to two Bayesian observers, one implementing the balance-of-evidence rule and one implementing the decision-congruent-only rule for confidence. We then computed the percentage of cases in which the decision-congruent-only produced higher CP for confidence than the balance-of-evidence rule for each subjects, which gives the exceedance probability of the decision-congruent-only rule (the likelihood that it predicted subjects' behaviour better than the balance-of-evidence rule).

This direct model comparison revealed that the decision-congruent-only rule is not just equivalent but superior in predicting confidence, with exceedance probability of 72.8% (chance is 50%). This result demonstrates that confidence is in fact better predicted by decision-congruent evidence alone than by a balance-of-evidence rule (see also Supplementary Results: Generative Bayesian models).

Our results demonstrate not only that neural representations (correlates) and computations underlying decisions and confidence are dissociable, but also that confidence selectively reflects the magnitude of decision-congruent evidence. This interpretation helps to explain previous findings in the literature regarding dissociations between accuracy and confidence, including cases in which changes in accuracy are not accompanied by appropriate changes in confidence[55], cases in which inactivation of cortical or subcortical structures affects confidence but not accuracy[56,57], and cases in which confidence disproportionately tracks decision-congruent evidence magnitude even when this strategy reduces metacognitive sensitivity[16]. Our findings are also in keeping with previous studies showing that when noise is added to a stimulus[58] or observer's internal representation[7,59–61], confidence increases while accuracy stays constant or decreases. This occurs because increased fluctuation in neural evidence favouring both stimulus alternatives is symmetric around a decision criterion (at zero; Fig. 4a), but can only increase the average magnitude of decision-congruent evidence (as it is by definition an absolute value; Fig. 4a). Thus, confidence rises even as accuracy remains unchanged or even decreases. Our results provide an account of how these dissociations between behavioural accuracy and confidence may arise from differences in computations at the neural level.

That decision-congruent evidence magnitude directly influences confidence has important implications for the possible neural substrates underlying probabilistic confidence computations[12,62–67]. Specifically, why would the system elect to compute confidence in this seemingly suboptimal way? The answer may have to do with the types of task that the perceptual system must solve in the real world. Most laboratory tasks present an artificial scenario in which an observer must decide between two known categories (for example face/house, left/right): in the real world you would never know for sure that an object exists but not know what it is. In contrast, in an ecologically valid setting, the task is not to categorize a stimulus into category A versus B, but to identify the stimulus—that is, to ask, "Is there something there, and if so, what is it?" Once a categorical decision has been made, the observer may have very little decision-incongruent evidence owing to the numerous possible alternative categories; the categories about which the observer has the most

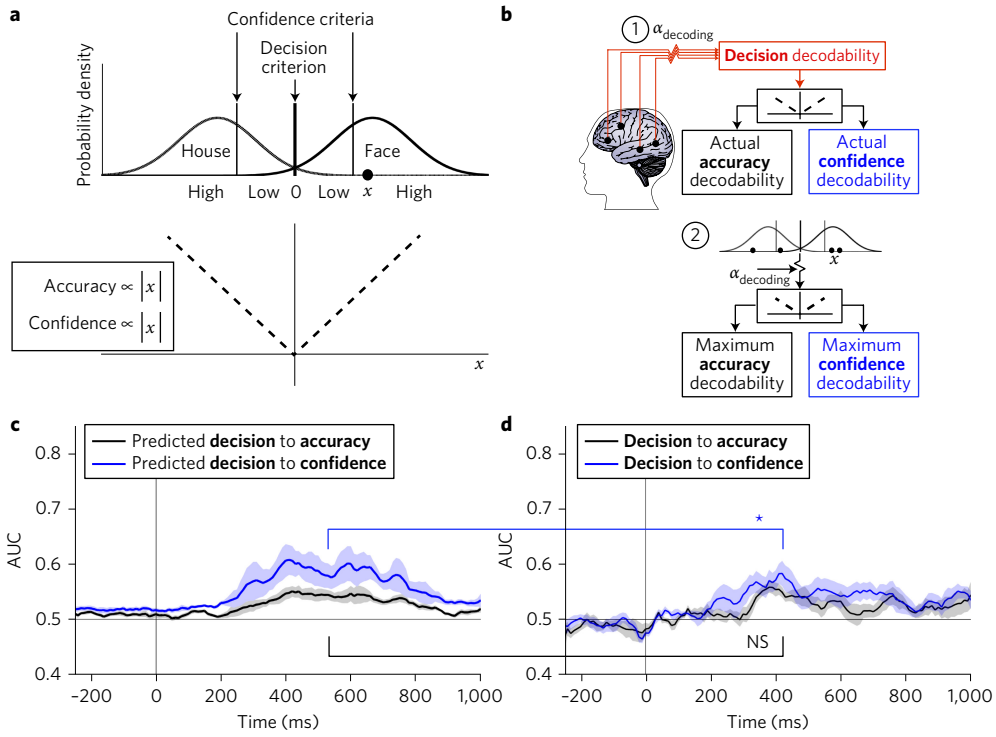

Figure 4 | Violations of the normative model for confidence but not accuracy. a, In SDT, on a given trial the internal evidence available to a system can be represented as *x*, a sample drawn from one of two distributions representing stimulus categories in a discrimination task. The sign of *x* dictates which category an unbiased observer will choose, such that positive *x* (above the decision criterion at zero) leads to a 'face' decision and negative *x* (below the decision criterion) to a 'house' decision. Likewise, the magnitude of *x*, or its distance from the decision criterion at zero, indicates how strongly it indicates a 'face' or 'house' choice: the farther *x* is from zero, the more likely observers are to be correct, and so the more confident they should be in their categorization choices. Thus, the absolute value of *x* predicts both the trial-by-trial accuracy (trial-by-trial correct choices/errors) and confidence in a decision. **b**, We fitted the assumed decoding noise in the SDT model, $\alpha_{decoding}$, to each subject by degrading the predicted decision decodability (based on subjects' performance and the stimulus decoder; see Methods: SDT forward model) to match the observed decision decodability. Incorporating this noise, we then used the model to predict the theoretical maximum for accuracy and confidence decodability for each subject. **c**, Given the presence of observed decoding noise, the model predicts that the theoretically expected maximal level of decodability for confidence will be above that for accuracy. (See Methods for explanation of the phrases 'decision to accuracy' and 'decision to confidence'. No ratio is implied.) **d**, We compared the actual decodability for accuracy and for confidence achieved via the model to the theoretical maxima predicted by the model. Whereas mean accuracy decodability reached the theoretical maximum ($t(5) = 1.58$, $P = 0.173$), confidence decodability was significantly worse ($t(5) = 2.868$, $P = 0.035$). This indicates that confidence cannot depend purely on the same internal information as decision and accuracy. Shaded regions indicate the standard error of the mean.

information are the chosen category itself, and some (presumably known) 'nothingness' category. Thus, perhaps the detectability of a stimulus itself is a primary contributor to confidence[16,54]. In other words, in the actual environment, objects that are more detectable are generally more discriminable: if you can see it well, you can probably tell what it is very well. This implies that the neural circuitry developed for stimulus detection may be recruited for confidence despite their conceptual differences[68,69], and perhaps even that the optimal solution to a laboratory-based discrimination task may not be the same as the optimal solution (or a heuristic-based approximation) in an ecologically valid setting. From an evolutionary perspective, this recruitment of detection circuitry seems reasonable: when an organism must judge both what is out there in the environment and whether there is something out there (simultaneous identification and detection), reliance on decision-congruent evidence magnitude might very well lead to adaptive behaviour.

The observation that decision-incongruent evidence is discarded in certain types of post-decision judgements is not unique to confidence: several authors have reported biases in continuous stimulus estimation[70], especially following a categorical decision[71], that seem to follow a similar pattern[20,72]. In one study, once subjects had made a categorical discrimination of motion direction, their subsequent estimations of motion direction indicated that they assumed any motion direction on the 'wrong' (unchosen) side of the reference criterion to be impossible[20,73] a similar effect has been reported in orientation discrimination (Luu & Stocker, 2016). Stocker and Simoncelli explain these biases as maximizing 'self-consistency' to maintain stable interpretations of the environment, and their Bayesian model is conceptually akin to our Bayesian heuristic model that relies on decision-congruent evidence[20]. Both models have the advantage of reducing costly storage and computation requirements in maintaining the full posterior probability distribution over many unchosen alternatives; in many real-life scenarios, this factor may overcome the need to minimize error in the expected estimation of motion direction, confidence, or other similar judgements. Additionally, despite reports that memory confidence appears to reflect the balance of evidence at the single neuron level[74], it has also been suggested that similar decision-congruent evidence dependence may underlie memory confidence in a task specifically designed to compare the two computational approaches[75], as we did here.

Here, motivated by previous studies[15–18], we tested the hypothesis that perceptual decisions and confidence judgements may involve dissociable mechanisms. Our findings go beyond previous behavioural results to reveal that decision-incongruent evidence

can indeed be read out from neural representations at the time of the confidence judgement, is used in the computation of the decision, and yet is discarded or ignored in the confidence computation. Specifically, this heuristical account provided a better fit to empirical data than a normative optimal model, as supported by our formal computational analysis. This over-emphasis on decision-congruent evidence is unlikely to be an *ad hoc* explanation, but rather seems to be the general strategy used by the brain in producing confidence reports in perceptual decisions. Future studies using similar neural decoding approaches may provide insight into use of neural evidence under other task conditions in which confidence judgements appear optimal at the behavioural level[54]. Also, it may be beneficial to apply this approach to other datasets with more comprehensive spatial coverage, as well as to directly assess the complex relationship between high gamma power, spiking activity, and lower frequency field potentials (see Supplementary Results: Frequencies outside 80–120 Hz, and Supplementary Notes). These may help to test further whether self-consistency is truly a general principle contributing to an organism's evaluation of its own internal uncertainty. As it has been speculated that this strategy may account for a wide range of high-level social phenomena including cognitive dissonance reduction[20], future investigation may be able to address the intriguing question of whether these mechanisms are common across species, or whether they might be uniquely human.

Methods

Details of the behavioural methods, ECoG data acquisition and preprocessing, support vector machine decoding, signal localization and generative Bayesian models can be found in the Supplementary Methods.

Choice probability analysis. *Definition of evidence.* In two-class linear SVM analysis, the result of training an estimator is a hyperplane that separates the two classes; one can take the dot product of the support vector coefficients (coefficients of the vector orthogonal to the hyperplane) and the support vectors themselves to determine the weights on each 'feature'. We then define whether a given feature provides evidence towards classifying the stimulus in a given trial as a face versus a house as the sign of its feature weight based on an SVM estimator trained on the trial-by-trial stimulus ('stimulus' estimator). Thus, mathematically, we define evidence for each timepoint t as

$$E_s(n,t) = \frac{1}{|e_s^*(t)|} \sum_{i \in e_s^*(t)} f_s(n,t,i) \quad (1)$$

where

$$f_s(n,t,i) = |w_i| \times g_{n,t,i} \times I_i \quad (2)$$

Here, $E_s(n, t)$ represents the overall evidence value for a given stimulus type s (face/house) and timepoint t in trial n, $e_s^*(t)$ represents the set of electrode-frequency features forming evidence for stimulus type s at timepoint t, $|e_s^*(t)|$ represents the cardinality of $e_s^*(t)$ (that is, the number of elements in the set), w_i represents the weight (described above) assigned to electrode-frequency feature i by the stimulus SVM estimator, $g_{n,t,i}$ represents the high-gamma power in trial n at time point t for electrode-frequency feature i, and I_i is an indicator function such that $I_i = 1$ if the sign of w_i matches the sign of the stimulus category s and 0 otherwise. Importantly, this definition of evidence maximizes the independence of 'face evidence' and 'house evidence', so their contributions to decisions and confidence can be independently evaluated.

Definition of balance-of-evidence and response-congruent-only rules. We evaluated two rules for predicting subjects' trial-by-trial decisions and confidence judgements: the balance of evidence favouring the decision versus that against the decision (balance-of-evidence), and the evidence favouring the decision alone (decision-congruent-only). Behavioural decisions and confidence for each subject were assigned as hits and false alarms according to standard ROC methods[51], and the AUC was calculated as before to obtain choice probability (CP) values for each subject for each rule. Conceptually, these hit and false alarm assignments were similar across both decision and confidence ROC analyses. Specifically, ROC methods sweep a criterion c through the decision value space, categorizing trials on the basis of whether their 'scores' (decision values; that is, the result of a particular classification rule) fall above or below c. For decisions, scores for the balance-of-evidence rule were defined as trial-by-trial face evidence minus house evidence. This leads to a 'hit' being defined as (face evidence) – (house evidence) > c ('face' response anticipated) and the subject responded 'face'; and a 'false alarm' being defined as (face evidence) – (house evidence) > c ('face' response anticipated) but the subject responded 'house'. The decision-congruent-only rule for decisions was defined as the average of the ROC curves and CPs for face evidence alone (on both face and house trials) and for house evidence alone (on both face and house trials) (Fig. 3a). For confidence, a balance-of-evidence 'hit' was defined as (response-congruent evidence) – (response-incongruent evidence) > c ('high confidence' anticipated) and the subject responded 'high confidence', and a 'false alarm' defined as (response-congruent evidence) – (response-incongruent evidence) > c ('high confidence' anticipated) but the subject responded 'low confidence'.

These CP values were used to assess the relative contribution of each type of evidence to decision and confidence over the analysed time period; note that a CP value of over 0.5 indicates that a given classifier is informative with regard to trial outcome (either decision or confidence), as this means that hits rise more rapidly than false alarms. We evaluated whether the CPs were significantly different from chance (CP = 0.5) using two-tailed t-tests, as well as inspecting differences in the CP performance of the balance-of-evidence versus decision-congruent-only rules for predicting decision and confidence using a 2 (rule) × 2 (behaviour) repeated-measures ANOVA.

SDT forward model. In standard SDT, on a given trial the internal evidence available to a system can be represented as x, a sample drawn from one of two distributions representing stimulus alternatives in a discrimination task (for example face/house; Fig. 4a). For an unbiased observer, the sign of x dictates which category the observer will choose, such that positive x leads to a 'face' decision and negative x to a 'house' decision. Likewise, the magnitude of x, or its distance from the decision criterion at zero, indicates how strongly it indicates a 'face' or 'house' choice, and thus dictates accuracy (probability of being correct). A normative observer should also rate confidence according to this same absolute magnitude: because the farther x is from zero the more likely a decision is to be correct, the more confident observers should be in their categorization choices (Fig. 4a).

Two-class linear SVM classification provides exactly such a 'sample' x in the form of the decision value (the trial-by-trial estimates $\hat{y}$; see Supplementary Methods) for each trial, such that positive $\hat{y}$ predict that the trial belongs to one group, and negative $\hat{y}$ the other (assuming no intercept bias). Following the normative framework, machine-learning methods such as SVM explicitly assume that the farther $\hat{y}$ is from the decision hyperplane, the more confident the classifier should be about its classification performance[76]. We therefore apply this forward model logic to the SVM decision values $\hat{y}$ to predict from decision to accuracy and confidence: we use the absolute value of the SVM $\hat{y}$ values for the decision estimator as inputs to the ROC analysis indexing classifier accuracy for accuracy and confidence on a trial-by-trial basis (see Supplementary Methods for more details). We tested this forward model's power to predict from 'decision to accuracy' and from 'decision to confidence' (Fig. 4c,d). All analyses and simulations were completed through custom-written software in MATLAB R2013a (MathWorks; Natuck, MA).

Evaluation of model. It would be unrealistic to assume that these SVM decision values $\hat{y}$ for the decision estimator represent a lossless readout of the internal decision variable x for each subject's face/house decision on each trial. If they represented a lossless readout, we would be able to decode all subjects' decisions (face/house button presses) with 100% accuracy with the SVM approach. Because decoding of decision does not reach this ceiling, we must instead assume that these $\hat{y}$ for the decision estimator are corrupted by some decoding noise with respect to the true internal decision variables x that dictate whether a subject said 'face' or 'house' (Fig. 4b). It is important to estimate this decoding noise empirically to validate the forward model. Essentially, this noise can be thought of as, "What is the signal degradation or noise that exists between the subject's access to his/her own neural representations, and our ability to access those neural representations through ECoG and an SVM decoder?" We estimated this decoding noise, $\alpha_{decoding}$, for each subject by building a simulated observer as follows. (Note that $\alpha_{decoding}$ will also therefore account for decoding noise due to subjects' errors, for example if a subject meant to indicate 'face' but erroneously pressed the 'house' button, as well as any degradation of signal due to limited spatial coverage with ECoG.)

Each subject's d' (objective performance capacity[51,52]) was first calculated from their behavioural data. Next, for each subject, using Monte Carlo simulations, we drew 1,000 samples x, representing the internal decision values, from each of two Gaussian distributions representing 'face' and 'house' centred at $\pm d'/2$ with standard deviation 1. Samples were classified according to the simple rule that $x > 0$ means 'face' and $x < 0$ means 'house' to provide the normative observer's decision (face/house), and subsequently classified as correct or incorrect according to the distribution that had generated them. We then used x to compute the decision ROC according to standard methods[51] to calculate the area under the curve ($AUC_{decision}$). Following the above discussion, we then computed $AUC_{accuracy}$ by the same method on $|x|$, the absolute value of x. To find the confidence criterion c used by each subject to separate confidence responses into 'high' versus 'low' (Fig. 4a), we swept through possible values for c from 0 to 5 in steps of 0.01, classifying $|x| > c$ as 'high' confidence and $|x| < c$ as 'low' confidence, to find the value of c that would provide a match to the proportion of 'high' and 'low' confidence responses given by

each subject. Finally, we computed $AUC_{confidence}$ on these $|x|$ values also according to the same methods as used for accuracy.

By using each subject's behavioural sensitivity and confidence criterion, this process provides a theoretical maximum for decodability of decision, accuracy and confidence. However, this theoretical maximum will in practice also be dictated by noise ($\alpha_{decoding}$) in the decoding process that corrupts our ability to access a subject's internal decision variable via an SVM decoder. To estimate $\alpha_{decoding}$ for each subject—that is, how 'bad' the SVM is at extracting the decision values that the subjects have access to in their own brains—we assume the following simple relationship between the SVM decision values $\hat{y}$ and the true internal decision variable x:

$$\hat{y} = \frac{\sigma_{\hat{y}}}{\sigma_x}(x + \varepsilon) \quad (3)$$

with $\varepsilon \sim N(0, \alpha_{decoding})$. Because ROC analyses do not depend on the actual values of x, only the shape of their distribution, we ignore the scaling factor $\frac{\sigma_{\hat{y}}}{\sigma_x}$ and define a proxy for $\hat{y}$ in simulation space:

$$x^* = x + \varepsilon \quad (4)$$

We fit $\alpha_{decoding}$ at each timepoint in the peri-stimulus window by minimizing the sum of squared error between $AUC_{decision}$ calculated on $\hat{y}$ (the true decoding accuracy for the decision estimator at that timepoint) and $AUC_{decision}$ calculated on x^* under increasing $\alpha_{decoding}$ noise at each timepoint in the peri-stimulus window for each subject. These best-fitting values for $\alpha_{decoding}$ were then used to predict the noisy theoretical maxima for $AUC_{accuracy}$ and $AUC_{confidence}$ given decoding noise, again at each timepoint in the peri-stimulus window for each subject.

It should be noted that the theoretical maxima for $AUC_{accuracy}$ and $AUC_{confidence}$ differ from one another because of the mathematical relationship among trial-by-trial accuracy, trial-by-trial confidence and trial-by-trial decision values x. According to SDT and other optimal models, 'confidence' is defined as the magnitude of the difference between the internal decision variable for 'decision' and the decision criterion[5,9,11]. As a result, confidence can be predicted almost perfectly from the internal decision variable for decision: the farther away it is from the decision criterion, the more confident one should be (Fig. 4a). On the other hand, for near-threshold psychophysics experiments such as the present one, predicting accuracy based on the magnitude of the internal decision variable is somewhat less trivial, although also mathematically clearly defined. Specifically, when the internal decision variable for decision is near the criterion, one does not always make errors; because of chance, one in fact makes a good portion of correct responses even in this range (Fig. 4a). Despite this, one should always be 'low confidence' in such near-criterion cases. As such, the theoretical bounds for how much one can decode confidence and accuracy are intrinsically different, with confidence theoretically easier to decode than accuracy from the magnitude of the internal decision variable under a given level of noise.

Therefore, if the forward model is true, and confidence is decoded from the same internal evidence as decision, then both accuracy and confidence decodability resulting from the rectified SVM decision values should reach these theoretical maxima. If, in contrast, confidence depends on information other than the magnitude of the internal decision variable for decision (that is, does not depend solely on the balance of evidence for face versus house), then accuracy decoding—defined by the trial-by-trial decision—should reach the theoretical maximum but confidence decoding should not. We tested whether the theoretical maximum for accuracy and confidence decoding had been reached via this forward model by using two paired t-tests to compare the mean decoding accuracy for accuracy and confidence from the SVM features to this theoretical maximum across the peri-stimulus time window. As before, to reveal global trends as a function of time, we smoothed the data using a five-point moving average (window size 50 ms).

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

Received 2 February 2017; accepted 6 June 2017; published 10 July 2017

References

1. Charles, L., King, J.-R. & Dehaene, S. Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. *J. Neurosci.* **34,** 1158–1170 (2014).
2. Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* **32,** 6117–6125 (2012).
3. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329,** 1541–1543 (2010).
4. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455,** 227–231 (2008).
5. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84,** 1329–1342 (2014).
6. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324,** 759–764 (2009).
7. Fetsch, C. R., Kiani, R., Newsome, W. T. & Shadlen, M. N. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* **83,** 797–804 (2014).
8. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19,** 366–374 (2016).
9. Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Phil. Trans. R. Soc. B* **367,** 1322–1337 (2012).
10. Meyniel, F., Schlunegger, D. & Dehaene, S. The sense of confidence during probabilistic learning: a normative account. *PLoS Comput. Biol.* **11,** e1004305 (2015).
11. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90,** 499–506 (2016).
12. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian probability: from neural origins to behavior. *Neuron* **88,** 78–92 (2015).
13. Gherman, S. & Philiastides, M. G. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *Neuroimage* **106,** 134–143 (2015).
14. van den Berg, R. *et al.* A common mechanism underlies changes of mind about decisions and confidence. *Elife* **5,** e12192 (2015).
15. Koizumi, A., Maniscalco, B. & Lau, H. Does perceptual confidence facilitate cognitive control? *Atten. Percept. Psychophys.* **77,** 1295–1306 (2015).
16. Maniscalco, B., Peters, M. A. K. & Lau, H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Atten. Percept. Psychophys.* **78,** 923–937 (2016).
17. Samaha, J., Barrett, J. J., Sheldon, A. D., Larocque, J. J. & Postle, B. R. Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Front. Psychol.* **7,** 851 (2016).
18. Zylberberg, A., Barttfeld, P. & Sigman, M. The construction of confidence in a perceptual decision. *Front. Integr. Neurosci.* **6,** 79–79 (2012).
19. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput. Biol.* **11,** e1004519 (2015).
20. Stocker, A. A. & Simoncelli, E. P. A Bayesian model of conditioned perception. *Adv. Neural Inf. Process. Syst.* **20,** 1409–1416 (2008).
21. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28,** 11526–11536 (2008).
22. Ray, S. & Maunsell, J. H. R. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9,** e1000610 (2011).
23. Winawer, J. *et al.* Asynchronous broadband signals are the principal source of the bold response in human visual cortex. *Curr. Biol.* **23,** 1145–1153 (2013).
24. Mukamel, R. *et al.* Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* **309,** 951–954 (2005).
25. Kunii, N., Kamada, K., Ota, T., Kawai, K. & Saito, N. Characteristic profiles of high gamma activity and blood oxygenation level-dependent responses in various language areas. *Neuroimage* **65,** 242–249 (2013).
26. Esposito, F. *et al.* Cortex-based inter-subject analysis of iEEG and fMRI data sets: application to sustained task-related BOLD and gamma responses. *Neuroimage* **66,** 457–468 (2013).
27. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412,** 150–157 (2001).
28. Crone, N. E., Sinai, A. & Korzeniewska, A. High-frequency gamma oscillations and human brain mapping with electrocorticography. *Prog. Brain Res.* **159,** 275–295 (2006).
29. Crone, N. E., Boatman, D., Gordon, B. & Hao, L. Induced electrocorticographic gamma activity during auditory perception. (Brazier Award-winning article, 2001). *Clin. Neurophysiol.* **112,** 565–582 (2001).
30. Hermes, D., Miller, K. J., Wandell, B. A. & Winawer, J. Stimulus dependence of gamma oscillations in human visual cortex. *Cereb. Cortex* **25,** 2951–2959 (2015).
31. Hipp, J. F., Engel, A. K. & Siegel, M. Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron* **69,** 387–396 (2011).
32. Laczó, B., Antal, A., Niebergall, R., Treue, S. & Paulus, W. Transcranial alternating stimulation in a high gamma frequency range applied over V1 improves contrast perception but does not modulate spatial attention. *Brain Stimul.* **5,** 484–491 (2012).
33. Davidesco, I. *et al.* Exemplar selectivity reflects perceptual similarities in the human fusiform cortex. *Cereb. Cortex* **24,** 1879–1893 (2014).
34. Privman, E. *et al.* Antagonistic relationship between gamma power and visual evoked potentials revealed in human visual cortex. *Cereb. Cortex* **21,** 616–624 (2011).

35. Shum, J. *et al.* A brain area for visual numerals. *J. Neurosci.* **33,** 6709–6715 (2013).
36. Dastjerdi, M., Ozker, M., Foster, B. L., Rangarajan, V. & Parvizi, J. Numerical processing in the human parietal cortex during experimental and natural conditions. *Nat. Commun.* **4,** 2528 (2013).
37. Kubánek, J., Miller, K. J., Ojemann, J. G., Wolpaw, J. R. & Schalk, G. Decoding flexion of individual fingers using electrocorticographic signals in humans. *J. Neural Eng.* **6,** 066001 (2009).
38. Yu, S., Pleskac, T. J. & Zeigenfuse, M. D. Dynamics of postdecisional processing of confidence. *J. Exp. Psychol. Gen.* **144,** 489–510 (2015).
39. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117,** 864–901 (2010).
40. Maniscalco, B. & Lau, H. The signal processing architecture underlying subjective reports of sensory awareness. *Neurosci. Conscious.* **2016,** niw002 (2016).
41. Chen, J., Feng, T., Shi, J., Liu, L. & Li, H. Neural representation of decision confidence. *Behav. Brain Res.* **245,** 50–57 (2013).
42. Heereman, J., Walter, H. & Heekeren, H. R. A task-independent neural representation of subjective certainty in visual perception. *Front. Hum. Neurosci.* **9,** 551 (2015).
43. McCurdy, L. Y. *et al.* Anatomical coupling between distinct metacognitive systems for memory and visual perception. *J. Neurosci.* **33,** 1897–1906 (2013).
44. Schwiedrzik, C. M., Singer, W. & Melloni, L. Subjective and objective learning effects dissociate in space and in time. *Proc. Natl Acad. Sci. USA* **108,** 4506–4511 (2011).
45. Li, Q., Hill, Z. & He, B. J. Spatiotemporal dissociation of brain activity underlying subjective awareness, objective performance and confidence. *J. Neurosci.* **34,** 4382–4395 (2014).
46. Middlebrooks, P. G. & Sommer, M. A. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* **75,** 517–530 (2012).
47. Fleming, S. M. & Dolan, R. J. The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B* **367,** 1338–1349 (2012).
48. Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E. & Lau, H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn. Neurosci.* **1,** 165–175 (2010).
49. Lau, H. & Passingham, R. E. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc. Natl Acad. Sci. USA* **103,** 18763–18768 (2006).
50. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13,** 87–100 (1996).
51. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (Wiley, 1966).
52. Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide* (Taylor & Francis, 2004).
53. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18,** 203–210 (2014).
54. Peters, M. A. K. & Lau, H. Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *Elife* **4,** e09651 (2015).
55. Vlassova, A., Donkin, C. & Pearson, J. Unconscious information changes decision accuracy but not confidence. *Proc. Natl Acad. Sci. USA* **111,** 16214–16218 (2014).
56. Lak, A. *et al.* Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84,** 190–201 (2014).
57. Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. & Miyamoto, A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* **16,** 749–755 (2013).
58. Zylberberg, A., Roelfsema, P. R. & Sigman, M. Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious. Cogn.* **27C,** 246–253 (2014).
59. Rahnev, D., Maniscalco, B., Luber, B., Lau, H. & Lisanby, S. H. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J. Neurophysiol.* **107,** 1556–1563 (2012).
60. Rahnev, D. *et al.* Attention induces conservative subjective biases in visual perception. *Nat. Neurosci.* **14,** 1513–1515 (2011).
61. Peters, M.A.K. *et al.* Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex* **93,** 119–132 (2017).
62. Beck, J. M., Ma, W. J., Latham, P. E. & Pouget, A. Probabilistic population codes and the exponential family of distributions. *Prog. Brain Res.* **165,** 509–519 (2007).
63. Beck, J. M. *et al.* Probabilistic population codes for Bayesian decision making. *Neuron* **60,** 1142–1152 (2008).
64. Ma, W. J., Beck, J. M., Latham, P. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9,** 1432–1438 (2006).
65. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14,** 119–130 (2010).
66. Ma, W. J., Beck, J. M. & Pouget, A. Spiking networks for Bayesian inference and choice. *Curr. Opin. Neurobiol.* **18,** 217–222 (2008).
67. Berkes, P., Orban, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **331,** 83–87 (2011).
68. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21,** 422–430 (2012).
69. Fleming, S. M. & Lau, H. How to measure metacognition. *Front. Hum. Neurosci.* **8,** 443 (2014).
70. Wei, X. & Stocker, A. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. *Adv. Neural Inf. Process. Syst.* **25,** 1313–1321 (2012).
71. Fleming, S. M., Maloney, L. T. & Daw, N. D. The irrationality of categorical perception. *J. Neurosci.* **33,** 19060–19070 (2013).
72. Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446,** 912–915 (2007).
73. Luu, L. & Stocker, A. A. Choice-induced biases in perception. Preprint at http://biorxiv.org/content/early/2016/04/01/043224 (2016).
74. Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat. Neurosci.* **18,** 1041–1050 (2015).
75. Zawadzka, K., Higham, P. A. & Hanczakowski, M. Confidence in forced-choice recognition: what underlies the ratings? *J. Exp. Psychol. Learn. Mem. Cogn.* **43,** 552–564 (2016).
76. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R* (Springer, 2015).

Acknowledgements

This work is supported by funding from the Templeton Foundation (grant 21569 to H.L.) and the US National Institute of Neurological Disorders and Stroke (NIH R01 NS088628 to H.L.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank U. Maoz for discussion on some technical issues regarding analysis.

Author contributions

M.A.K.P. and H.L. together developed the key theoretical ideas behind the project, analysed the data and wrote the paper. H.L., T.T., E.H. and M.D. designed the behavioral paradigm and initiated project planning. T.T. and M.D. were primarily responsible for data collection. B.M., Y.D.K. and M.D. contributed to data analysis. W.D., R.K. and O.D. contributed to data collection and overcoming logistical challenges. T.T. oversaw the logistical issues and planning involved in the entire project.

Additional information

Supplementary information is available for this paper.

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

Correspondence and requests for materials should be addressed to M.A.K.P.

How to cite this article: Peters, M. A. K. *et al.* Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* **1,** 0139 (2017).

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

Competing interests

The authors declare no competing interests.