HHS Public Access
Author manuscript
Nat Hum Behav. Author manuscript; available in PMC 2018 January 10.

Published in final edited form as:
Nat Hum Behav. 2017 ; 1: . doi:10.1038/s41562-017-0139.

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], **Matt Davidson**[5], **Werner Doyle**[2], **Ruben Kuzniecky**[2], **Orrin Devinsky**[2], **Eric Halgren**[3], and **Hakwan Lau**[1,7]

[1]Department of Psychology, University of California, Los Angeles, Los Angeles, California, USA

[2]Comprehensive Epilepsy Center, Department of Neurology, New York University Medical Center, New York, New York, USA

[3]Multimodal Imaging Laboratory, University of California, San Diego, La Jolla, California, USA

[4]Department of Physiology & Neuroscience, St. George's University, Grenada, West Indies

[5]Department of Psychology, Columbia University, New York, New York, USA

[6]Neuroscience Institute, New York University, New York, New York, USA

[7]Brain Research Institute, University of California, Los Angeles, Los Angeles, California, USA

Summary paragraph

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 behavioral 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 and 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 overly relies on evidence

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

Correspondence should be addressed to: Megan A. K. Peters, 1285 Franz Hall, Box 951563, University of California, Los Angeles, Los Angeles, CA 90095, (323) 596-1093, meganakpeters@ucla.edu.
[*]these authors contributed equally to this work

Author Contributions

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

Competing Interests

The authors declare no competing interests.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

supporting a decision (e.g., face evidence for a 'face' decision), even while decisions themselves reflect the optimal balance of all evidence (e.g., both face and house evidence). Thus, confidence may not reflect a readout of the probability of being correct; instead, observers may sacrifice optimality in favor of self-consistency[20] in the face of limited neural and computational resources. While seemingly suboptimal, this strategy may reflect the inference problem that perceptual systems are evolutionarily optimized to solve.

Results

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 judgments by pressing buttons on a keyboard (Figure 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 analyze 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 (Figure 1b), and had faster reaction times for high- versus low-confidence responses ($\mu_{high} = 1059ms$, $\mu_{low} = 1439ms$, t(5) = 4.32, p = .007) but not for high-versus low-contrast trials ($\mu_{high} = 1096ms$, $\mu_{low} = 1132ms$, t(5) = 1.96, p = .11). Subjects also showed little response bias to respond 'face' versus 'house' (Figure 1b).

Following previous work which 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. Indeed, the mean time-frequency spectrum averaged over all subjects, electrodes, and trials was most salient in this high-gamma range, centered around 250–400ms after stimulus onset (Supplementary Figure 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 behavioral 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 behavioral factors above chance at different time bins after stimulus onset (Figure 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 250ms, but Confidence decodability did not reach significance for half of subjects until 450ms (Supplementary Table 3).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Importantly, both factors were able to be decoded above chance well before any movement onset (mean RT = 1136ms), suggesting that decoding is not based on movement (finger movement preparation can typically be decoded only up to ~200ms before movement onset with ECoG[37]; see also Supplementary Results: Motor preparation and neuroanatomical localization results, below).

The above results suggest that Decision and Confidence behaviors may reflect different evidence at different time points. One could argue that this dissociation may be trivial, since 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 via 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.

Therefore, we also directly assessed the representation correlates' spatial separation[41–45]. We quantified each electrode's contribution to decodability by calculating a normalized *contribution index* (*C*) (Supplementary Methods: Equations S1 & S2), which we projected onto its MNI coordinates averaged across coarse 200ms time bins to reveal broad patterns (see Supplementary Methods: Neuroanatomical localization of representations) (Figure 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 (Figure 2c) (see also Supplementary Figures 6–8, and Supplementary Tables 4 & 5).

Occipital regions showed localized contributions to Decision starting at 200–400ms despite their sparsity in electrode numbers, but Confidence appears to be more neuroanatomically distributed (significant main effects of lobe for Decision ($F(3,870) = 7.748$, $p < .001$) but not Confidence ($F(3,870) = 1.896$, $p = .129$); Supplementary Results: Representational overlap) with marked contributions from parietal[6] and frontal areas[2,46–49] (Figures 2b & 2c). Note that the separability of Decision and Confidence representations does not mean that there is completely no overlap. In terms of simple response level (rather than decodability), there are individual electrodes that showed some sensitivity to both Decisions and Confidence judgments, although they did so in ways also congruent with our central hypotheses (Supplementary Results: Representational overlap; Supplementary Figure 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 due to 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 likely not 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 judgment will reflect the strength of the neural Evidence for ‘face’ but will be largely insensitive to the (lack of) Evidence for ‘house’. While this hypothesis has received some support from behavioral studies[14–19], it remains controversial, with a number of researchers arguing that confidence judgments 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 (i.e., electrode-frequency-timepoint; Supplementary Methods: Support vector machine decoding). We combined these weights with each Feature’s power to define ‘Evidence’ -- i.e., how much the neural code reflected both the stimulus’ ‘face-ness’ and ‘house-ness’ on each trial (Methods: Choice probability analysis, Equations 1 & 2) -- and categorized Evidence depending on the subject’s Decisions: Face Evidence is Decision-Congruent on trials when 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 the Balance-Of-Evidence versus Decision-Congruent-Only rules: on a trial-by-trial basis for each subject, we assigned Decisions and Confidence judgments as hits and false alarms according to standard Receiver Operating Characteristic (ROC) methods[51], and calculated the area under the curve 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 behavior (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) x 2 (Evidence: Balance, Decision-Congruent) repeated-measures ANOVA: no main effect for Predictor (F(1,5) = 4.538, p = .086), main effect for Evidence (F(1,5) = 9.665, p = .027), and significant interaction between Predictor and Evidence (F(1,5) = 6.961, p = .046)) (Figure 3a & 3b). 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 < .001; Confidence: t(5) = 0.6719, p = 0.531) (Figure 3a & 3b). This means that taking into account Decision-Incongruent Evidence does not help to better predict Confidence rating behavior 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 Figure 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[51,52] to build a normative forward model, and to formally assess the noise ceiling stipulated by the decodability of the data. Assuming subjects are Bayesian ideal observers, their Confidence should be monotonically related to Accuracy[4], i.e. it should optimally reflect the probability of a Decision's being correct on a trial-by-trial basis[7–12] (Figure 4a; Methods: Signal detection theoretic 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 (Figure 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, etc. 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 signal detection theory (Figure 4a). Based on the observed level of decoding noise ($\alpha_{decoding}$), we estimated the theoretically maximal expected decodability for both Accuracy and Confidence (Figure 4b; Methods: Signal detection theoretic forward model). We then compared this expected maximum to actual data (i.e., decodability of Accuracy and Confidence via the forward model, based on all available Features).

Indeed, statistical tests confirmed that Accuracy decodability achieved via the model was indistinguishable from the theoretical maximum given noise ($\alpha_{decoding}$), but Confidence decodability was significantly worse than the theoretical maximum (Figure 4c & 4d). 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-utilized in the computation of Confidence.

One may worry that the detection theoretic model failed because of the different timecourses of information flow for Decision (Type 1) and Confidence (Type 2) judgments[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'

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(especially after *t*). However, we saw no evidence for temporal dissociations that could have led to the model's failure (Supplementary Results: Lag in predicting from Decision to Confidence?; Supplementary Figure 11). This analysis demonstrates the informativeness of neural signals in evaluating the 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 formally address this concern, we capitalized on Bayesian generative model simulations to directly compare 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 & 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 percent 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' behavior 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).

Discussion

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 where changes in Accuracy are not accompanied by appropriate changes in Confidence[55], where inactivation of cortical or subcortical structures affects Confidence but not Accuracy[56,57], and where 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[7,59,60], Confidence increases while Accuracy stays constant or decreases. This occurs because increased fluctuation in neural Evidence favoring both stimulus alternatives is symmetric around a decision criterion (at zero; Figure 4a), but can only increase the average magnitude of Decision-Congruent Evidence (as it is by definition an absolute value; Figure 4a). Thus, Confidence rises even as Accuracy remains unchanged or even decreases. Our results provide an account of how these dissociations between behavioral Accuracy and Confidence may arise from differences in computations at the neural level.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

That Decision-Congruent Evidence magnitude directly influences Confidence has important implications for the possible neural substrates underlying probabilistic Confidence computations[12,61–66]. Specifically, why would the system elect to compute confidence in this seemingly suboptimal way? The answer may have to do with the types of tasks 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 (e.g., 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 vs. B, but to *identify* the stimulus, i.e. 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 due to the numerous possible alternative categories; the categories about which the observer has the most information are the face 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[67,68], 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 *if* there is something out there (simultaneous identification and detection), reliance on Decision-Congruent Evidence magnitude might very well lead to adaptive behavior.

The observation that Decision-Incongruent Evidence is discarded in certain types of post-Decision judgments is not unique to Confidence: several authors have reported biases in continuous stimulus estimation[69], especially following a categorical decision[70], that seem to follow a similar pattern[20,71]. In one study, once subjects had made a categorical motion direction discrimination, 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]. 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 Decision-Congruent Evidence Bayesian heuristic model[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 judgments. Additionally, despite reports that memory confidence appears to reflect the balance of evidence at the single neuron level[72], 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[73], as we did here.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

out from neural representations at the time of the confidence judgment, is used in the computation of the decision, and yet is discarded or ignored in the confidence computation. Specifically, this heuristical account provided 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 appears to be the general strategy employed 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 judgments appear optimal at the behavioral 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 further test whether self-consistency is truly a general principle contributing to an organism's evaluation of its own internal uncertainty. Since it has been speculated that this strategy may account for a wide range of high-level social phenomenon 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 behavioral methods, electrocorticography (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 support vector machine (SVM) analysis, the result of SVM 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

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

where

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 $e_s^*(t)$ (i.e. 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 judgments: the Balance of Evidence favoring the Decision versus that against the Decision (Balance-Of-Evidence), and the Evidence favoring the Decision alone (Decision-Congruent-Only). Behavioral Decisions and Confidence for each subject were assigned as hits and false alarms according to standard Receiver Operating Characteristic (ROC) methods[51], and the AUC was calculated as before to obtain 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, i.e. 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, leading 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 House Evidence alone (on both Face and House trials) (Figure 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 analyzed 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) x 2 (behavior) repeated-measures ANOVA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Signal detection theoretic forward model: In standard signal detection theory (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 (e.g. face/house, Figure 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, x's magnitude, 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 (Figure 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 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→ Accuracy and Decision→Confidence. 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 (i.e., 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 which dictate whether a subject said 'face' or 'house' (Figure 4b). It is important to estimate this decoding noise empirically in order 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 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, e.g. 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 behavioral data. Next, for each subject, using Monte Carlo simulations, we drew 1000 samples x, representing the internal decision values, from each of two Gaussian distributions

Author Manuscript

representing ‘face’ and ‘house’ centered 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 ($\text{AUC}_{\text{Decision}}$). Following the above discussion, we then computed $\text{AUC}_{\text{Accuracy}}$ via 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’ (Figure 4a), we swept through possible values for c from 0 to 5 in steps of .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 $\text{AUC}_{\text{Confidence}}$ on these $|x|$ values also according to the same methods as used for Accuracy.

By utilizing each subject’s behavioral 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 -- i.e., 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:

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

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:

$$x*=x+\varepsilon \qquad (4)$$

We fit $\alpha_{decoding}$ at each timepoint in the peri-stimulus window by minimizing the sum of squared error between $\text{AUC}_{\text{Decision}}$ calculated on $\hat{y}$ (i.e., the true decoding accuracy for the Decision estimator at that timepoint) and $\text{AUC}_{\text{Decision}}$ 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 $\text{AUC}_{\text{Accuracy}}$ and $\text{AUC}_{\text{Confidence}}$ given decoding noise, again at each timepoint in the peri-stimulus window for each subject.

It should be noted that the theoretical maxima for $\text{AUC}_{\text{Accuracy}}$ and $\text{AUC}_{\text{Confidence}}$ differ from one another due to the mathematical relationship among trial-by-trial Accuracy, trial-by-trial Confidence, and trial-by-trial decision values x. According to signal detection theory and other optimal models, Confidence is defined as the magnitude of the difference between

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

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 (Figure 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, though 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 (Figure 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 (i.e., 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 5-point moving average (window size 50ms).

Data availability

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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 Uri Maoz for useful discussion on some technical issues regarding analysis.

References

1. Charles L, King JR, Dehaene S. Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. J Neurosci. 2014; 34:1158–1170. [PubMed: 24453309]
2. Fleming SM, Huijgen J, Dolan RJ. Prefrontal contributions to metacognition in perceptual decision making. J Neurosci. 2012; 32:6117–6125. [PubMed: 22553018]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. Fleming SM, Weil RS, Nagy Z, Dolan R, Rees G. Relating Introspective Accuracy to Individual Differences in Brain Structure. Science. 2010; 329

4. Kepecs A, Uchida N, Zariwala HA, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. Nature. 2008; 455:227–231. [PubMed: 18690210]

5. Kiani R, Corthell L, Shadlen MN. Choice Certainty Is Informed by Both Evidence and Decision Time. Neuron. 2014; 84:1329–1342. [PubMed: 25521381]

6. Kiani R, Shadlen MN. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. Science. 2009; 324:759–764. [PubMed: 19423820]

7. Fetsch CR, Kiani R, Newsome WT, Shadlen MN. Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. Neuron. 2014:1–8.

8. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: distinct probabilistic quantities for different goals. Nat Neurosci. 2016; 19:366–374. [PubMed: 26906503]

9. Kepecs A, Mainen ZF. A computational framework for the study of confidence in humans and animals. Philos Trans R Soc Lond B Biol Sci. 2012; 367:1322–1337. [PubMed: 22492750]

10. Meyniel F, Schlunegger D, Dehaene S. The Sense of Confidence during Probabilistic Learning: A Normative Account. PLoS Comput Biol. 2015; 11:1–25.

11. Sanders JI, Hangya B, Kepecs A. Signatures of a Statistical Computation in the Human Sense of Confidence. Neuron. 2016; 90:499–506. [PubMed: 27151640]

12. Meyniel F, Sigman M, Mainen ZF. Confidence as Bayesian Probability: From Neural Origins to Behavior. Neuron. 2015; 88:78–92. [PubMed: 26447574]

13. Gherman S, Philiastides MG. Neural representation of choice confidence emerges from the process of decision formation during perceptual choices Task/Behaviour. Neuroimage. 2015; 106:121587–121587.

14. Berg R, van D, et al. A common mechanism underlies changes of mind about decisions and confidence 2. Elife. 2015:1–5.

15. Koizumi A, Maniscalco B, Lau H. Does perceptual confidence facilitate cognitive control? Atten Percept Psychophys. 2015; doi: 10.3758/s13414-015-0843-3

16. Maniscalco B, Peters MAK, Lau H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Atten Percept Psychophys. 2016; doi: 10.3758/s13414-016-1059-x

17. Samaha J, Barrett JJ, Sheldon AD, Larocque JJ, Postle BR. Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. Front Psychol. 2016; 7:1–8. [PubMed: 26858668]

18. Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. Front Integr Neurosci. 2012; 6:79–79. [PubMed: 23049504]

19. Aitchison L, Bang D, Bahrami B, Latham PE. Doubly Bayesian analysis of confidence in perceptual decision-making. PLoS Comput Biol. 2015; 11:e1004519–e1004519. [PubMed: 26517475]

20. Stocker AA, Simoncelli EP. A Bayesian Model of Conditioned Perception. Adv Neural Inf Process Syst. 2008; 20:1409–1416.

21. Ray S, Crone NE, Niebur E, Franaszczuk PJ, Hsiao SS. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. Journal of Neuroscience. 2008; 28:11526–11536. [PubMed: 18987189]

22. Ray S, Maunsell JHR. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol. 2011; 9

23. Winawer J, et al. Asynchronous broadband signals are the principal source of the bold response in human visual cortex. Curr Biol. 2013; 23:1145–1153. [PubMed: 23770184]

24. Mukamel R, et al. Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science. 2005; 309:951–954. [PubMed: 16081741]

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. 2013; 65:242–249. [PubMed: 23032488]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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. 2013; 66:457–468. [PubMed: 23138047]

27. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. Nature. 2001; 412:150–157. [PubMed: 11449264]

28. Crone, NE., Sinai, A., Korzeniewska, A. Research, C. N. A. W. K. B. T.-P. in B. Vol. 159. Elsevier; 2006. p. 275-295.

29. Crone NE, Boatman D, Gordon B, Hao L. Induced electrocorticographic gamma activity during auditory perception. Brazier Award-winning article, 2001. Clin Neurophysiol. 2001; 112:565–582. [PubMed: 11275528]

30. Hermes D, Miller KJ, Wandell BA, Winawer J. Stimulus Dependence of Gamma Oscillations in Human Visual Cortex. Cereb Cortex. 2015; 25:2951–2959. [PubMed: 24855114]

31. Hipp JF, Engel AK, Siegel M. Oscillatory synchronization in large-scale cortical networks predicts perception. Neuron. 2011; 69:387–396. [PubMed: 21262474]

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. 2012; 5:484–491. [PubMed: 21962982]

33. Davidesco I, et al. Exemplar selectivity reflects perceptual similarities in the human fusiform cortex. Cereb Cortex. 2014; 24:1879–1893. [PubMed: 23438448]

34. Privman E, et al. Antagonistic relationship between gamma power and visual evoked potentials revealed in human visual cortex. Cereb Cortex. 2011; 21:616–624. [PubMed: 20624838]

35. Shum J, et al. A brain area for visual numerals. J Neurosci. 2013; 33:6709–6715. [PubMed: 23595729]

36. Dastjerdi M, Ozker M, Foster BL, Rangarajan V, Parvizi J. Numerical processing in the human parietal cortex during experimental and natural conditions. Nat Commun. 2013; 4:2528. [PubMed: 24129341]

37. Kubánek J, Miller KJ, Ojemann JG, Wolpaw JR, Schalk G. Decoding flexion of individual fingers using electrocorticographic signals in humans. J Neural Eng. 2009; 6:066001–066001. [PubMed: 19794237]

38. Yu S, Pleskac TJ, Zeigenfuse MD. Dynamics of Postdecisional Processing of Confidence. J Exp Psychol Gen. 2015; 144:489–510. [PubMed: 25844627]

39. Pleskac TJ, Busemeyer JR. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychol Rev. 2010; 117:864–901. [PubMed: 20658856]

40. Maniscalco B, Lau H. The signal processing architecture underlying subjective reports of sensory awareness. Neuroscience of Consciousness. 2016:1–41.

41. Chen J, Feng T, Shi J, Liu L, Li H. Neural representation of decision confidence. Behav Brain Res. 2013; 245:50–57. [PubMed: 23415909]

42. Heereman J, Walter H, Heekeren HR. A task-independent neural representation of subjective certainty in visual perception. Front Hum Neurosci. 2015; 9:1–12. [PubMed: 25653611]

43. McCurdy LY, et al. Anatomical coupling between distinct metacognitive systems for memory and visual perception. J Neurosci. 2013; 33:1897–1906. [PubMed: 23365229]

44. Schwiedrzik CM, Singer W, Melloni L. Subjective and objective learning effects dissociate in space and in time. Proc Natl Acad Sci U S A. 2011; 108:4506–4511. [PubMed: 21368168]

45. Li Q, Hill Z, He BJ. Spatiotemporal dissociation of brain activity underlying subjective awareness, objective performance and confidence. J Neurosci. 2014; 34:4382–4395. [PubMed: 24647958]

46. Middlebrooks PG, Sommer MA. Neuronal Correlates of Metacognition in Primate Frontal Cortex. Neuron. 2012; 75:517–530. [PubMed: 22884334]

47. Fleming SM, Dolan RJ. The neural basis of metacognitive ability. Philos Trans R Soc Lond B Biol Sci. 2012; 367:1338–1349. [PubMed: 22492751]

48. Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci. 2010; 1:165–175. [PubMed: 24168333]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

49. Lau H, Passingham RE. Relative blindsight in normal observers and the neural correlate of visual consciousness. Proceedings of the National Academy of Sciences. 2006; 103:18763–18768.

50. Britten KH, Newsome WT. A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis Neurosci. 1996; 13:87–100. [PubMed: 8730992]

51. Green, DM., Swets, JA. Signal Detection Theory and Psychophysics. John Wiley & Sons, Inc; 1966.

52. Macmillan, NA., Creelman, CD. Detection Theory: A User's Guide. Taylor & Francis; 2004.

53. King JR, Dehaene S. Characterizing the dynamics of mental representations: The temporal generalization method. Trends Cogn Sci. 2014; 18:203–210. [PubMed: 24593982]

54. Peters MAK, Lau H. Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. Elife. 2015

55. Vlassova A, Donkin C, Pearson J. Unconscious information changes decision accuracy but not confidence. Proceedings of the National Academy of Sciences. 2014; 111:16214–16218.

56. Lak A, et al. Orbitofrontal cortex is required for optimal waiting based on decision confidence. Neuron. 2014; 84:190–201. [PubMed: 25242219]

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. 2013; 16:749–755. [PubMed: 23666179]

58. Zylberberg A, Roelfsema PR, Sigman M. Variance misperception explains illusions of confidence in simple perceptual decisions. Conscious Cogn. 2014; 27C:246–253.

59. Rahnev D, Maniscalco B, Luber B, Lau H, Lisanby SH. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. J Neurophysiol. 2012; 107:1556–1563. [PubMed: 22170965]

60. Rahnev D, et al. Attention induces conservative subjective biases in visual perception. Nat Neurosci. 2011; 14:1513–1515. [PubMed: 22019729]

61. Beck JM, Ma WJ, Latham PE, Pouget A. Probabilistic population codes and the exponential family of distributions. Prog Brain Res. 2007; 165:509–519. [PubMed: 17925267]

62. Beck JM, et al. Probabilistic Population Codes for Bayesian Decision Making. Neuron. 2008; 60:1142–1152. [PubMed: 19109917]

63. Ma WJ, Beck JM, Latham P, Pouget A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006; 9:1432–1438. [PubMed: 17057707]

64. Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci. 2010; 14:119–130. [PubMed: 20153683]

65. Ma WJ, Beck JM, Pouget A. Spiking networks for Bayesian inference and choice. Curr Opin Neurobiol. 2008; 18:217–222. [PubMed: 18678253]

66. Berkes P, Orban G, Lengyel M, Fiser J. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. Science. 2011; 331:83–87. [PubMed: 21212356]

67. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn. 2012; 21:422–430. [PubMed: 22071269]

68. Fleming SM, Lau H. How to measure metacognition. Front Hum Neurosci. 2014; 8:1–9. [PubMed: 24474914]

69. Wei X, Stocker A. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. Adv Neural Inf Process Syst. 2012:1–9.

70. Fleming SM, Maloney LT, Daw ND. The Irrationality of Categorical Perception. J Neurosci. 2013; 33:19060–19070. [PubMed: 24305804]

71. Jazayeri M, Movshon JA. A new perceptual illusion reveals mechanisms of sensory decoding. Nature. 2007; 446:912–915. [PubMed: 17410125]

72. Rutishauser U, et al. Representation of retrieval confidence by single neurons in the human medial temporal lobe. Nat Neurosci. 2015; 18

73. Zawadzka K, Higham PA, Hanczakowski M. Confidence in Forced-Choice Recognition: What Underlies the Ratings? J Exp Psychol Learn Mem Cogn. 2016; doi: 10.1037/xlm0000321

74. Vilares I, Körding KP. Bayesian models: the structure of the world, uncertainty, behavior, and the brain. Ann N Y Acad Sci. 2011; 1224:22–39. [PubMed: 21486294]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

75. Yu, AJ., Dayan, P. Advances in Neural Information Processing Systems 17. Saul, LK.Weiss, Y., Bottou, L., editors. MIT Press; 2005. p. 1577-1584.

76. James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning, with Applications in R. Springer; 2015.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

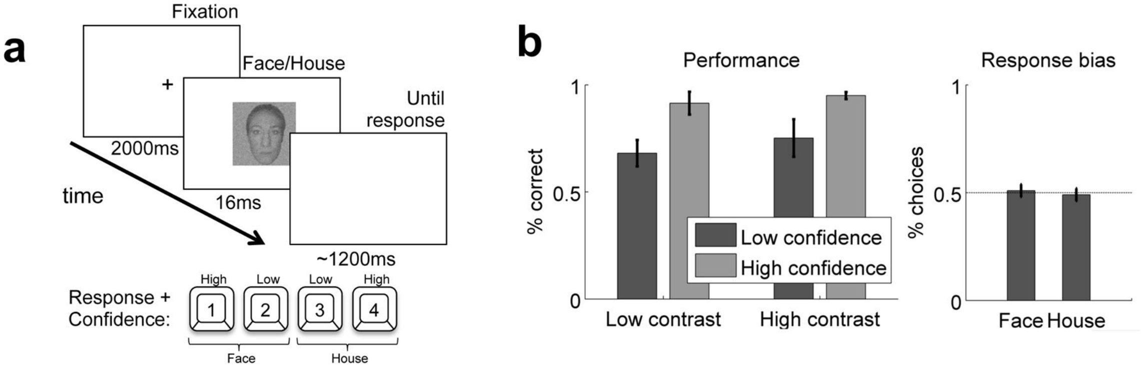

Figure 1.

Behavioral task and results. (a) Subjects discriminated noisy stimuli as faces/houses and indicated their confidence (high vs low) with a single button press; responses were all made with one hand. (b) As expected, subjects showed higher accuracy for high versus low contrast stimuli, and for high confidence versus low confidence responses (2 (contrast: high/low) x 2 (confidence: high/low) repeated measures ANOVA: $F(1,5)_{confidence} = 8.418$, $p = .034$; $F(1,5)_{contrast} = 1.783$, $p = .239$; $F(1,5)_{confidenceXcontrast} = 0.502$, $p = .10$), 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.

Author Manuscript

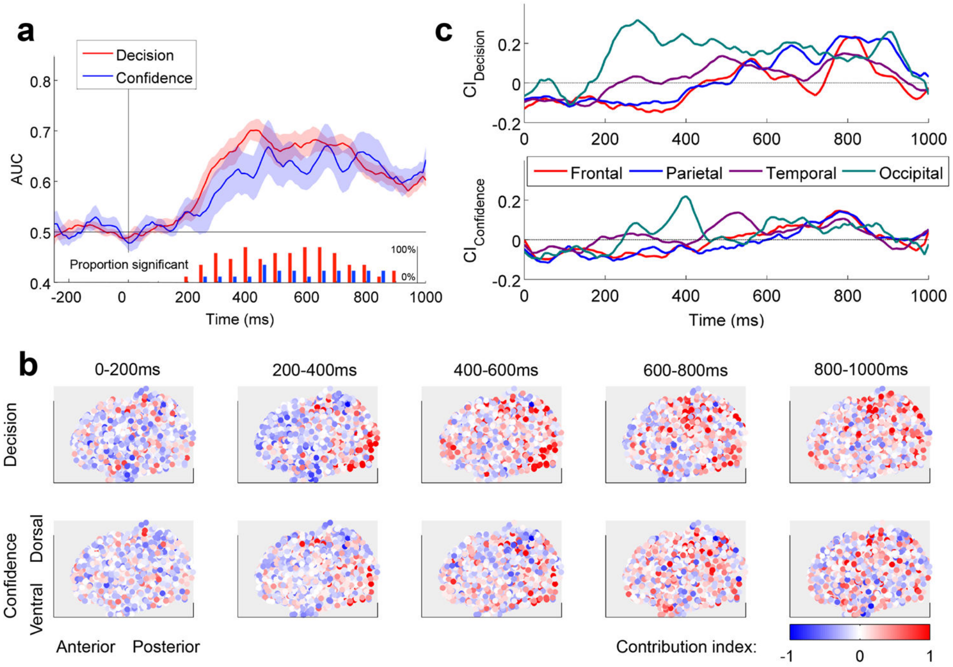

Figure 2.

Spatiotemporal dissociation between Decision and Confidence decoding. (a) Decoder accuracy for both estimators (Decision and Confidence) rises just around 200ms post stimulus onset. 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 50ms 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 200ms. $C < 0$ (dark blue) indicates the electrode contributed very little, whereas $C > 0$ (red) indicates the electrode contributes 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 200–700ms, while Confidence occupies a more distributed spatial representation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

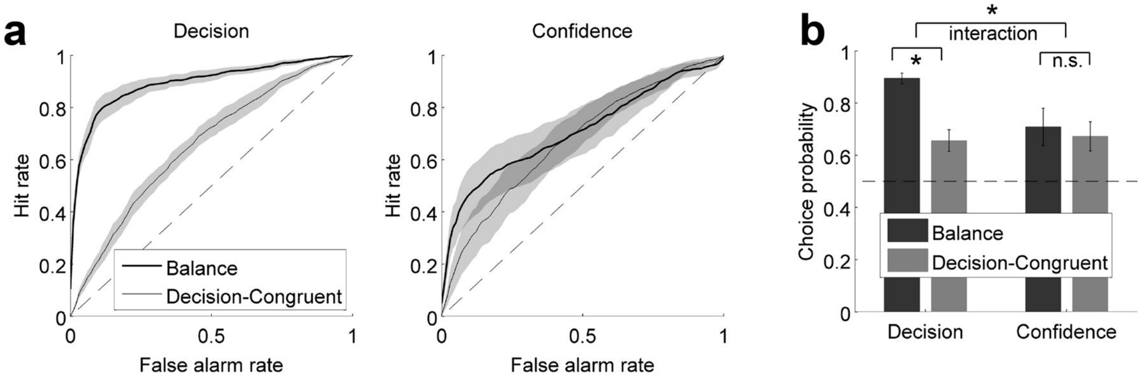

Figure 3.

Choice probability analyses show 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 utilize Decision-Incongruent Evidence.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

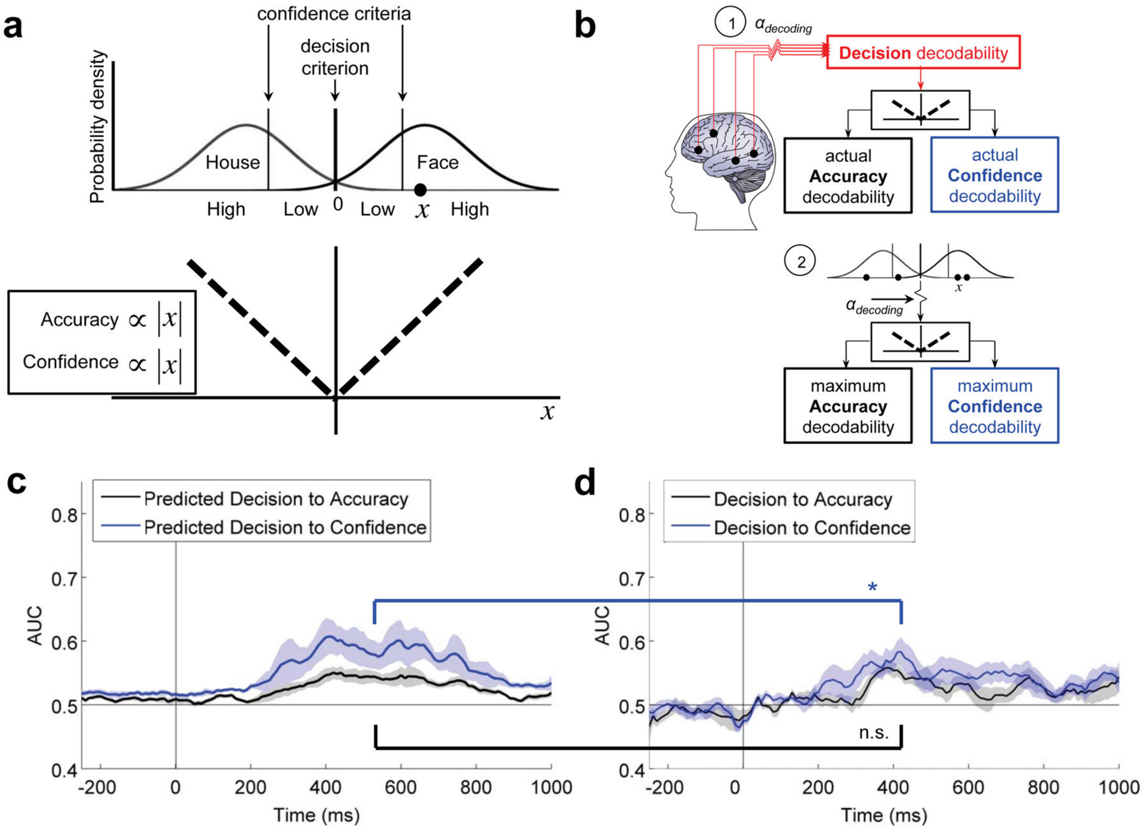

Figure 4.

Violations of the normative model for Confidence but not Accuracy. (a) In signal detection theory, 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, *x*’s magnitude, 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 signal detection theoretic model, $\alpha_{decoding}$, to each subject by degrading the predicted Decision decodability (based on subjects’ performance and the stimulus decoder; see Methods: Signal detection theoretic 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. (d) We compared the actual Accuracy and Confidence decodability achieved via the model to the theoretical maxima predicted by the model. While mean Accuracy decodability reached the theoretical maximum (t(5) = 1.58, p = .173), Confidence decodability was significantly worse (t(5) = 2.868, p = .035). This

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

indicates that Confidence cannot depend purely on the same internal information as Decision and Accuracy. Shaded regions indicate the standard error of the mean.