PNAS RESEARCH ARTICLE | NEUROSCIENCE

Sensory population activity reveals downstream confidence computations in the primate visual system

Zoe M. Boundy-Singer[a,1], Corey M. Ziemba[a,2], and Robbe L. T. Goris[a,3]

Edited by Martin Banks, University of California, Berkeley, CA; received December 17, 2024; accepted May 22, 2025

Perception is fallible. Humans know this, and so do some nonhuman animals like macaque monkeys. When monkeys report more confidence in a perceptual decision, that decision is more likely to be correct. It is not known how neural circuits in the primate brain assess the quality of perceptual decisions. Here, we test two hypotheses. First, that decision confidence is related to the structure of population activity in the sensory cortex. And second, that this relation differs from the one between sensory activity and decision content. We trained macaque monkeys to judge the orientation of ambiguous stimuli and additionally report their confidence in these judgments. We recorded population activity in the primary visual cortex and used decoders to expose the relationship between this activity and the choice-confidence reports. Our analysis validated both hypotheses and suggests that perceptual decisions arise from a neural computation downstream of visual cortex that estimates the most likely interpretation of a sensory response, while decision confidence instead reflects a computation that evaluates whether this sensory response will produce a reliable decision. Our work establishes a direct link between neural population activity in the sensory cortex and the metacognitive ability to introspect about the quality of perceptual decisions.

neural coding | visual cortex | sensory uncertainty | population representation | metacognitive

Significance

It has long been known that our sense of confidence is indicative of the quality of our decisions and actions. For example, when we feel more confident in a perceptual decision, that decision is more likely to be correct. Yet, the neural processes that endow us with this "metacognitive" ability to evaluate the quality of our own brain processes have remained largely elusive. Here, we developed a perceptual confidence task in which monkeys directly report a perceptual decision and decision confidence. This task allowed us to show that biological brains acquire metacognitive abilities through simple deterministic transformations of sensory population activity.

Perception is fallible (1–3). Humans know this (4–6), and so do some nonhuman animals like macaque monkeys (7–14). Indeed, perceptual interpretations of the environment are automatically accompanied by a sense of confidence in this interpretation. For example, when soccer fans in a football stadium see a striker score a goal, they may hold their breath and ask other fans whether the ball really went in. Judging the trajectory of fast moving objects is difficult, and we know this. The "metacognitive" ability to evaluate the quality of perceptual interpretations helps us to plan future actions (15), learn from mistakes (16, 17), and optimize group decision-making (18). How does the brain assess the quality of perceptual decisions? A prominent hypothesis is that early areas in the sensory cortex provide raw sensory measurements which are used by downstream circuits in the association cortex to guide perceptual decisions (19–23) and assign confidence in these decisions (7, 10, 12, 13). This hypothesis pertains to the neural coding of information and specifically applies to population representations. In other words, there may exist a systematic relationship between neural population activity in the sensory cortex and confidence in perceptual decisions. Consistent with this hypothesis, a recent fMRI study showed that variability in neural activity in the human visual cortex was linked to variability in behavioral confidence reports (24). And two prior studies of macaque monkeys showed that electrical and optogenetic stimulation of visual area MT influenced the monkeys' opt-out choice behavior in a motion discrimination postdecision wagering task, suggesting that neural activity in the sensory cortex is causally related to confidence in perceptual decisions (10, 25). Here, we set out to document the basic characteristics of this relationship.

To examine the relationship between sensory activity and confidence, we developed a task that invites subjects to jointly report a perceptual decision and their confidence in this decision (also see ref. 27). We trained two macaque monkeys (F and Z) to judge whether a visual stimulus presented near the central visual field was oriented clockwise or counterclockwise from vertical. The monkeys communicated their judgment with a saccade to one of four choice targets, organized in a rectangular pattern around the fixation mark (Fig. 1*A*). Horizontal saccade direction indicated the perceptual judgment, vertical saccade direction indicated the confidence in the decision. Choices were rewarded in a manner that incentivizes observers to introspect about decision quality on a trial-by-trial basis. Specifically, high confidence judgments resulted in a larger immediate reward when correct, but in a loss of potential future reward when incorrect (*Materials and Methods*). While the animals performed this task, we recorded extracellular responses

Author affiliations: [a]Center for Perceptual Systems, University of Texas at Austin, Austin, TX 78712

Author contributions: Z.M.B.-S., C.M.Z., and R.L.T.G. designed research; Z.M.B.-S., C.M.Z., and R.L.T.G. performed research; Z.M.B.-S. analyzed data; and Z.M.B.-S., C.M.Z., and R.L.T.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]Present address: Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MD 02139.

[2]Present address: Laboratory of Sensorimotor Research, National Eye Institute, NIH, Bethesda, MD 20892.

[3]To whom correspondence may be addressed. Email: robbe.goris@utexas.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2426441122/-/DCSupplemental.

Published June 25, 2025.

from neural populations in the primary visual cortex (V1), the first sensory area in the primate visual system where individual neurons signal the task relevant feature, stimulus orientation (28).

We found that confidence in perceptual decisions can be predicted from V1 population activity. The relationship between sensory activity and decision confidence appears as strong as the relationship between sensory activity and decision content. This assessment is based on the analysis of one hidden-layer neural networks trained to either predict the perceptual choice or the confidence report from V1 population activity. In both cases, the networks captured behavioral effects of stimulus manipulations (variations in stimulus orientation and stimulus contrast) as well as behavioral variability under repeated presentations of the same stimulus. As predicted by theoretical models of perceptual confidence, the relation between sensory activity and decision confidence fundamentally differs from the one between sensory activity and decision content. It involves an additional nonlinearity and consideration of sensory uncertainty. Together, these results reveal how an essential metacognitive ability arises from downstream transformations of neural population activity in the sensory cortex.

Results

Behavior and Computational Hypothesis. Both monkeys learned to report confidence in a fine orientation discrimination task. Their perceptual choices lawfully depended on stimulus orientation, and they made few errors in the easiest stimulus conditions (monkey F = ± 18.2°, median performance, 100% correct; monkey Z = ± 15.0°, median performance, 100% correct). Consider the choice behavior for an example recording session. Choices reported with high confidence are shown in green, choices reported with low confidence in red, and symbol size is proportional to the number of trials (Fig. 1*B*). As is evident from the raw data, for every stimulus condition, high confidence choices tended to be more accurate than low confidence choices (Fig. 1*B*, green vs. red symbols). As a consequence, high confidence choices exhibited a steeper overall relationship with stimulus orientation (Fig. 1*B*, green vs. red curve). We quantified this effect by estimating the slope of both psychometric functions (operationalized as the inverse of the SD of a cumulative Gaussian function fit to the choice data) and computing the slope ratio per session (*Materials and Methods*). For the vast majority of

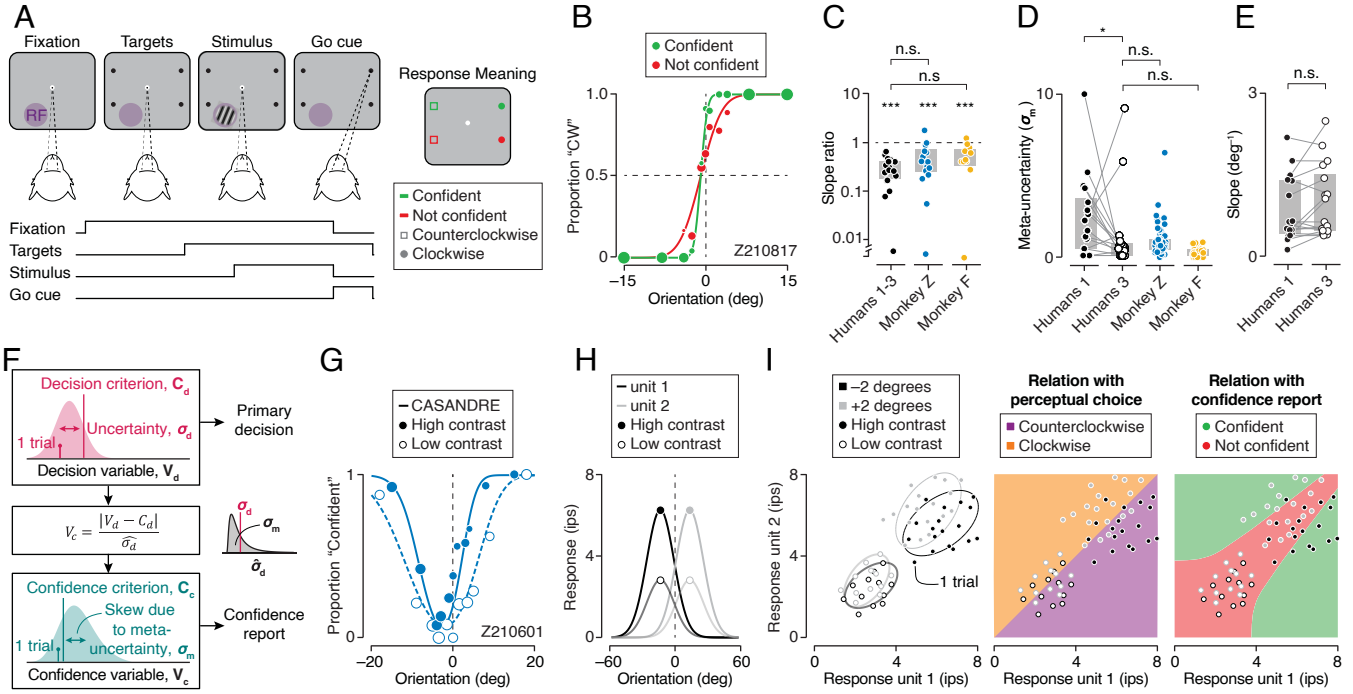

Fig. 1. Perceptual confidence task: behavior and computational hypothesis. (*A*) Orientation discrimination task sequence. After the observer fixates for at least 500 ms, four choice targets appear, followed by the stimulus. The stimulus is placed in the neurons' visual receptive field (RF). The observer judges whether the stimulus is rotated clockwise or counterclockwise relative to vertical. They jointly communicate this orientation judgment and their decision confidence with a saccade toward one of four choice targets. Horizontal saccade direction indicates the perceptual judgment, vertical saccade direction the confidence report. Correct decisions are followed by a juice reward (*Materials and Methods*). (*B*) Psychophysical performance for monkey Z in an example recording session. Proportion of clockwise (CW) choices for high-contrast stimuli is shown as a function of stimulus orientation, conditioned on the observer's confidence report. Symbol size reflects the number of trials (total 527 trials, slope ratio = 0.57). The curves are fits of a behavioral model (*Materials and Methods*). (*C*) The ratio of the slope of high contrast psychometric functions. Sensible confidence judgments yield values smaller than one. For the humans, each symbol represents the slope ratio of one subject across all three blocks of trials. For the monkeys, each symbol represents the slope ratio in one recording session (humans: $n = 19$; monkey Z: $n = 17$; monkey F: $n = 12$). Gray bars indicate interquartile range. (*D*) Meta-uncertainty for a group of human subjects and two monkeys. For the humans, each symbol represents metacognitive performance of one subject in one block of trials. For the monkeys, each symbol represent metacognitive performance in one behavioral session (humans: $n = 17$; monkey Z: $n = 60$; monkey F: $n = 58$). Gray bars indicate interquartile range. (*E*) Slope of high contrast psychometric functions for the same group of humans as (*D*). Each symbol represents the slope for one subject in one block of trials. (*F*) Schematic of a process model for decision confidence (26). (*G*) Proportion high confidence judgments as a function of stimulus orientation for high and low contrast stimuli (filled vs. open symbols) in an example recording session. Symbol size reflects the number of trials (total 744 trials). Solid lines are fits of the process model shown in panel (*F*). (*H*) Average firing rate as a function of stimulus orientation for two model neurons (black vs. gray) and two stimulus contrasts (open vs. closed symbols). (*I*) (*Left*) Joint responses of a pair of model neurons to repeated presentations of four stimuli that differ in orientation and contrast. (*Middle*) Illustration of a mapping rule that converts the pairwise activity into a perceptual decision. (*Right*) Illustration of a mapping rule that converts the same responses into a confidence report. Decision confidence is high when the sensory response is strong (toward *Upper Right* corner) and nonambiguous (away from the line of unity). n.s. not significant, *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$.

recording sessions, high confidence choices were associated with a steeper psychometric function than low confidence choices (Fig. 1*C*; monkey F: median slope ratio = 0.42, $P < 0.001$, Wilcoxon signed rank test against 1; monkey Z = 0.40, $P < 0.001$). This effect mirrored the choice behavior of a group of human subjects, naive to the purpose of our study, who performed a similar orientation discrimination task (Fig. 1*C*; $n = 19$, median slope ratio = 0.27, $P < 0.001$; *Materials and Methods*). These results suggest that the monkeys introspected about the quality of each perceptual decision and relied on a confidence assignment process that is qualitatively similar to the one used by humans (see *SI Appendix*, Fig. S1 for further comparison).

We wondered whether the monkeys' ability to assess the quality of perceptual decisions quantitatively resembles that of humans. The statistic we have considered thus far is inadequate to answer this question. The association between confidence and the slope of the psychometric function is a robust signature of sensible confidence assessments, but the slope ratio does not only depend on the quality of confidence assessments. It also depends on the subject's perceptual sensitivity and their proclivity to report high confidence (26, 29, 30). To quantify metacognitive ability, we need a statistic that is isolated from these other nuisance factors (26, 31, 32). Here, we estimated a statistic called meta-uncertainty from choice-confidence data in each session (σ_m; explained further below and in *Materials and Methods*). Meta-uncertainty expresses how well a decision maker can discriminate reliable from unreliable choices, and higher values indicate lower metacognitive ability(26). Surprisingly, we found that the monkeys outperformed the human subjects during this group's first visit to the lab (Fig. 1*D*, humans block 1 vs. monkeys; humans performed 1,100 trials in block 1, median $\sigma_m = 2.25$, median σ_m for monkeys = 0.47, $P < 0.001$, Wilcoxon rank sum test). We reasoned that task experience was the likely driver of this effect. To test this, we asked the human subjects to perform the experiment two more times. Reassuringly, they eventually caught up with the monkeys (Fig. 1*D*; median σ_m for humans in block 3 = 0.52, median σ_m for monkeys = 0.47, $P = 0.53$). This change in meta-uncertainty with experience was not due to perceptual learning. The slope of high and low contrast psychometric functions did not significantly change from block 1 to 3 (Fig. 1*E*; median high contrast slope for block 1 = 0.57, median high contrast slope for block 3 = 0.63, $P = 0.55$, Wilcoxon rank sum test; median low contrast slope for block 1 = 0.60, median low contrast slope for block 3 = 0.56, $P = 0.73$). We conclude that our animal paradigm invites high-quality metacognitive behavior.

What is the nature of the confidence assignment process that underlies this metacognitive capacity? Previous work has shown that choice-confidence data in tasks like ours are often well captured by a hierarchical process model in which confidence reflects an observer's estimate of the reliability of their decision (26, 31–33). In these models, a stimulus gives rise to a noisy, one-dimensional decision variable (for example, a perceptual orientation estimate). Comparison of this decision variable with a fixed criterion yields a perceptual decision ("clockwise" or "counterclockwise;" Fig. 1 *F*, *Top*). The reliability of this decision is revealed by computing the distance between the decision variable and the decision criterion, and normalizing this distance by an estimate of the uncertainty of the decision variable (Fig. 1 *F*, *Middle*). Comparison of this decision reliability estimate with a fixed confidence criterion yields a confidence report ("confident" or "not confident;" Fig. 1 *F*, *Bottom*). The quality of the confidence reports is limited by a subject's uncertainty about the uncertainty of the decision variable ("meta-uncertainty") (26), or by an analogous noise term, depending on the specific model variant (31, 32). As can be seen for an example dataset, this computational framework captures how the monkey's tendency to choose the "confident" response option jointly depends on stimulus orientation and stimulus contrast (Fig. 1*G* and *SI Appendix*, Fig. S1 *E* and *F*).

Decision-making areas downstream of sensory cortex do not get one-dimensional perceptual estimates as input, but high-dimensional population responses. They implement operations akin to these idealized model computations by mapping this population activity onto the available choice options. To gain an intuition for these mapping rules, consider a pair of hypothetical V1 neurons whose responses selectively depend on stimulus orientation and stimulus contrast (Fig. 1*H*). One of these neurons prefers orientations smaller than 0°, while the other one on average responds more vigorously to orientations larger than 0°. Thus, their joint activity pattern contains information about stimulus orientation, regardless of stimulus contrast. Specifically, when neuron 2 is more active than neuron 1, the stimulus is more likely to be oriented clockwise from vertical and vice versa (Fig. 1 *I*, *Left*). The mapping rule used by a downstream decision-making circuit can thus be understood as projecting the population activity onto a one-dimensional axis perpendicular to a linear hyperplane that separates clockwise from counterclockwise response patterns (Fig. 1 *I*, *Middle*). The resulting decisions will not be flawless—due to neural response variability, there is considerable overlap between both response distributions, making errors inevitable. Crucially, the population response also contains information about the probability of such an error. The closer the population activity is to the hyperplane, the more probable an error. This effect is amplified for activity patterns that reside close to the bottom left corner of this state space. This part of the space is visited when the stimulus-drive is weak, for example because stimulus contrast is low or stimulus size is small. Here, response patterns are dominated by spontaneous activity, resulting in high levels of sensory uncertainty (34–37) and many incorrect decisions (Fig. 1 *I*, *Middle*). These geometrical considerations yield two testable predictions. First, that decision confidence is related to the structure of population activity in the sensory cortex. And second, that this relation differs from the one between sensory activity and decision content.

Predicting Perceptual Decision Confidence from V1 Activity. While the animals performed the perceptual confidence task, we used multilaminar electrode arrays to record population activity from ensembles of V1 units whose receptive fields overlapped with the stimulus location (*Materials and Methods*). Populations ranged in size from 8 to 46 units (median = 15 units). Consider the activity of three simultaneously recorded units. Stimulus onset elicited a strong transient response, followed by a weaker sustained response (Fig. 2*A*). Some units were better driven by counterclockwise orientations (Fig. 2 *A*, *Top*), some by clockwise orientations (Fig. 2 *A*, *Bottom*), and some did not differentiate between these stimulus conditions (Fig. 2 *A*, *Middle*). We first asked whether the observed populations could in principle provide the sensory signals to support the perceptual task. To this end, we trained linear stimulus decoders to discriminate between clockwise and counterclockwise stimuli and tested them on nonambiguous hold-out trials (Fig. 2*B*; *Materials and Methods*). We found that each recorded population could support the perceptual task above chance level (neural performance ranged from 57.4 to 96.9% correct, median = 69.2%). These decoders have only been provided with neural population responses and stimulus labels ("clockwise" or "counterclockwise"). Yet it is

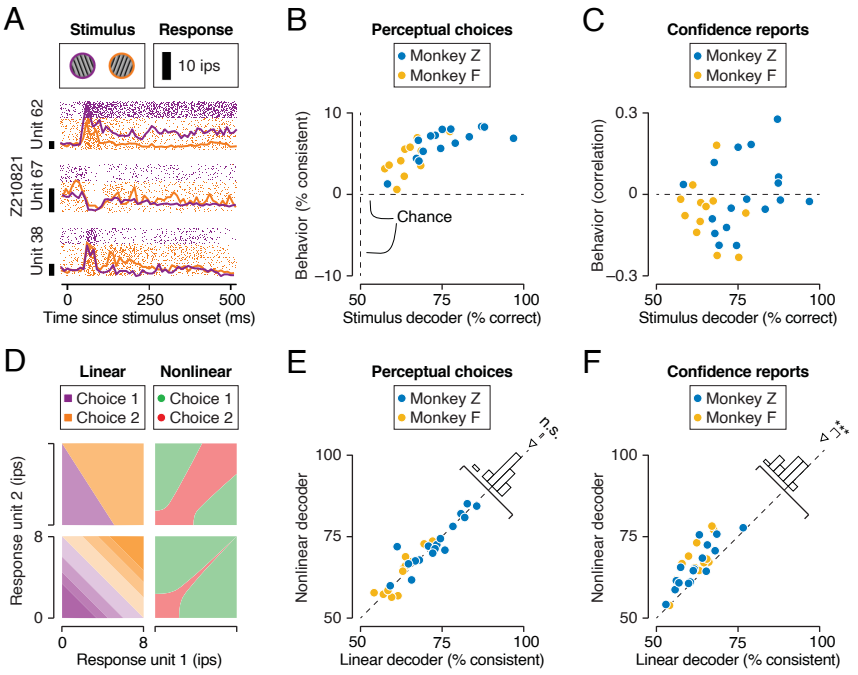

Fig. 2. Predicting perceptual decisions and decision confidence from V1 population activity. (*A*) Spike rasters (dots) and PSTHs (lines) of three example units during presentation of a clockwise (gray) and counterclockwise (black) stimulus. (*B*) Analysis of the linear stimulus decoder. The proportion of correctly predicted perceptual choices minus the proportion expected by chance is plotted against the decoder's task performance. Each symbol represents a recording session. (*C*) The correlation between the monkeys' confidence reports and the linear stimulus decoder's output is plotted against the decoder's task performance. (*D*) Example mapping rules that can be implemented by a linear (*Left*) and nonlinear (*Right*) decoder. (*E*) Comparison of the proportion correctly predicted perceptual choices by a linear (abscissa) and nonlinear (ordinate) choice decoder. (*F*) Comparison of the proportion correctly predicted confidence reports by a linear (abscissa) and nonlinear (ordinate) confidence decoder. n.s. not significant, $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$.

natural to ask whether they can offer some insight into the monkeys' behavior. We compared the stimulus decoders' choices with the animals' reports on a trial-by-trial basis while controlling for the stimulus and found that the fraction of correctly predicted perceptual decisions exceeded the number expected by chance (Fig. 2*B*; median difference = 6.3%, $P < 0.001$, Wilcoxon signed rank test). Both variables exhibited a clear relationship; the better the neural populations could support the perceptual task, the better the stimulus decoder predicted perceptual decisions (Fig. 2*B*; Pearson correlation: $r = 0.79$, $P < 0.001$; *SI Appendix*, Fig. S2). This finding is consistent with the hypothesis that decision-making circuits downstream of visual cortex estimate the most likely interpretation of a sensory response, just like these stimulus decoders do. We also compared the stimulus decoders' output with the animals' confidence reports and found them to be unrelated (Fig. 2*C*; median Pearson correlation = −0.03, $P = 0.16$). This result is not surprising. It simply confirms that in our task, the relationship between sensory activity and decision content cannot account for decision confidence.

To expose the relationship between sensory population activity and decision confidence, we trained confidence decoders to discriminate between trials in which monkeys reported choices with either high or low confidence. For this analysis, we considered both linear and nonlinear decoders (specifically, one-hidden layer neural networks; *Materials and Methods*). Linear decoders can slice a high-dimensional space in various ways (Fig. 2 *D*, *Left*), but none of the possible variations fully captures the hypothesized confidence mapping rule (Fig. 1 *I*, *Right*). Nonlinear decoders can implement both linear mappings or more complex input–output relations (Fig. 2 *D*, *Right*), making them better suited to test the hypothesis. Importantly, this additional complexity is not guaranteed to be beneficial. This will only be the case if the brain's mapping rule requires the complexity (compare Fig. 1 *I*, *Middle* and *Right*). To connect this concept to our data, we first compared linear and nonlinear choice decoders trained to predict perceptual decisions (*Materials and Methods*). We orthogonalized the neural choice and confidence subspaces by curating the decoders' training sets such that the animals' perceptual choice contained no information about their confidence report and vice versa (*Materials and Methods*; *SI Appendix*, Fig. S3*A*). As expected, linear and nonlinear choice decoders performed similarly well, suggesting that a linear mapping rule suffices to relate sensory population activity to perceptual decisions (Fig. 2*E*; median performance linear choice decoder = 66.8% correctly predicted choices; nonlinear choice decoder = 68.8%, median difference = −0.7%, $P = 0.34$, Wilcoxon signed rank test). We then performed the same comparative analysis on the confidence reports and obtained a different result. The nonlinear confidence decoders consistently outperformed their linear counterparts in predicting confidence (Fig. 2*F*; median linear confidence decoder = 61.8% correctly predicted confidence reports; nonlinear confidence decoder = 65.6%, median difference = 4.0%, $P < 0.001$). In general, confidence reports could be predicted about as well as perceptual decisions (median performance difference between nonlinear choice and confidence decoders = 3.5%, $P = 0.19$). Together, these results confirm that perceptual decision confidence is related to the structure of population activity in the sensory cortex, and that this relationship is more complex than the relation between this activity and decision content.

Interrogating the Confidence Decoder. We seek to understand how sensory population activity informs confidence in perceptual decisions. So far, our analysis suggests that nonlinear decoders

trained to predict behavioral choice-confidence reports from neural population activity are a powerful tool in this endeavor. Of course, this is only true to the extent that the mapping relation learned by the decoders resembles the one used by the brain. This need not be the case. Clearly, the confidence decoders are imperfect predictors of the animals' behavior. It is possible that their success is based on exploiting idiosyncratic relationships between neural responses and confidence reports that are distinct from the brain's confidence computation (38). If this were the case, the confidence decoders' output should not exhibit the key signature of sensible confidence assignments, nor should they be able to generalize to new testing conditions. We investigated both issues. We first computed the slope of the psychometric function, conditioned on the confidence decoder's output (*Materials and Methods*; Fig. 3*A*). Higher confidence outputs were associated with a steeper psychometric function (Fig. 3*B*; median slope ratio = 0.88, $P = 0.02$, Wilcoxon signed rank test). This pattern is significant for monkey Z (median slope ratio = 0.85, $P = 0.03$, Wilcoxon signed rank test) and suggestive for monkey F (median slope ratio = 0.89, $P = 0.22$, Wilcoxon signed rank test). This pattern recapitulates a key feature of the animals' behavior and implies that the confidence decoder's outputs are sensible. The confidence decoder recognizes which neural responses are more likely to result in a reliable perceptual decision.

If the mapping rule learned by the confidence decoder resembles the one used by the brain, it should transfer to more challenging testing conditions, such as input patterns it has not been exposed to during training. To test this, we probed the confidence decoders with synthetic patterns of neural activity. We designed these patterns such that they would expose the "pure" effects of stimulus orientation and stimulus contrast on decision confidence. Specifically, for every stimulus orientation, we created a synthetic pattern by computing the trial-averaged population response, thus removing the effects of neural response variability (*Materials and Methods*). We approximated the effects of manipulating stimulus contrast by multiplying these synthetic neural responses by different levels of gain (39–41). Here, we explored gain changes that went far outside the range driven by our experimental contrast manipulation to create out-of-distribution responses (*Materials and Methods*). Consider the output of the confidence decoder for an example recording session. More extreme orientations are always associated with more high confidence outputs, for both the mean population responses (Fig. 3*C*, gray line) as well as across all artificially induced levels of gain (Fig. 3*C*, colored lines). Additionally, higher levels of response gain are always associated with more high confidence outputs, regardless of the stimulus orientation (Fig. 3*C*), similar to the behavioral effect of increasing stimulus contrast (Fig. 1*G*). These effects were evident across datasets (Fig. 3*D*; median difference in predicted proportion high confidence outputs for more vs. less extreme stimulus orientations = 15%; median difference for a response gain of 0.5 and 2 = 38%). Thus, the decoder's confidence output jointly depends on stimulus orientation and stimulus contrast, thereby recapitulating the second key feature of the animals' behavior (*SI Appendix*, Fig. S1 *E* and *F*). Moreover, the mapping rule learned by the decoder generalizes to new testing conditions. We conclude that the confidence decoder evaluates neural activity in a sensible and robust manner.

Relationship Between Choice and Confidence Computations. Our analysis of neural activity was inspired by a computational framework in which confidence reflects an observer's estimate of the reliability of their decision(26, 31, 32). In this framework, the computations that form a decision are distinct from the ones that assign confidence in these decisions. However, there is a direct relationship between the latent variables that underlie the overt perceptual choices and confidence reports. Specifically, more extreme decision variable values will yield higher confidence variable values (Fig. 1 *F*, *Middle*). The decoders we trained on neural data use a latent variable to predict behavioral choice-confidence reports (*Materials and Methods*). We wondered whether these latent variables would be related as predicted by the computational framework. If this were the case, it would provide direct evidence for the notion that the brain's confidence computation evaluates the quality of the sensory evidence that informed the decision.

Consider the neurally decoded decision variable for an example recording session. There are three important effects. The decision variable varies linearly with stimulus orientation (Fig. 4*A*). The slope of this relationship depends on stimulus contrast (Fig. 4 *A*, *Left* vs. *Right* panel). And trials that culminate in a "clockwise

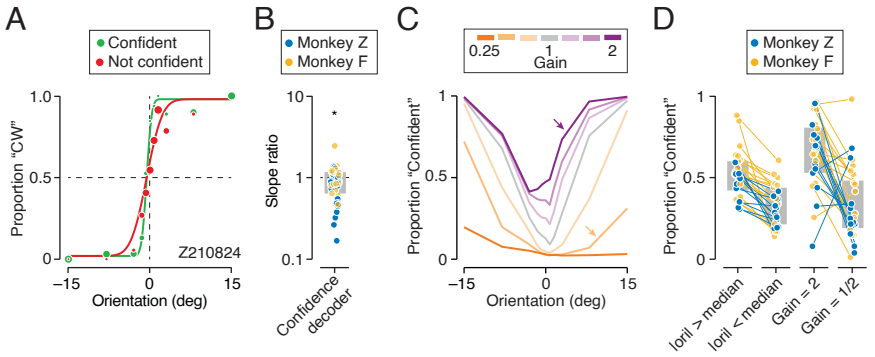

Fig. 3. The nonlinear confidence decoder yields sensible and robust outputs. (*A*) Psychophysical performance for monkey Z in an example recording session. Proportion of clockwise (CW) choices for high-contrast stimuli is shown as a function of stimulus orientation, conditioned on the confidence decoder's output. Symbol size reflects the number of trials (total 520 trials, slope ratio = 0.88). The curves are fits of a behavioral model. (*B*) The ratio of the slope of both psychometric functions. Each symbol represents the slope ratio in one recording session at one contrast ($n = 54$; see *Materials and Methods*). Gray bars indicate interquartile range. (*C*) Illustration of the confidence decoder's output for various synthetic patterns of neural activity for an example recording session. (*D*) Summary of the synthetic confidence experiments for all recording sessions. (*Left*) Proportion of predicted high confidence outputs elicited by stimuli whose orientation is more or less extreme than the median stimulus orientation. (*Right*) Proportion of high confidence outputs elicited by sensory input patterns with a high or low response gain [indicated by the colored arrows in panel (*C*)]. Each symbol represents the proportion of confident responses in one recording session at one contrast. Gray bars indicate interquartile range. n.s. not significant, $^*P < 0.05$.

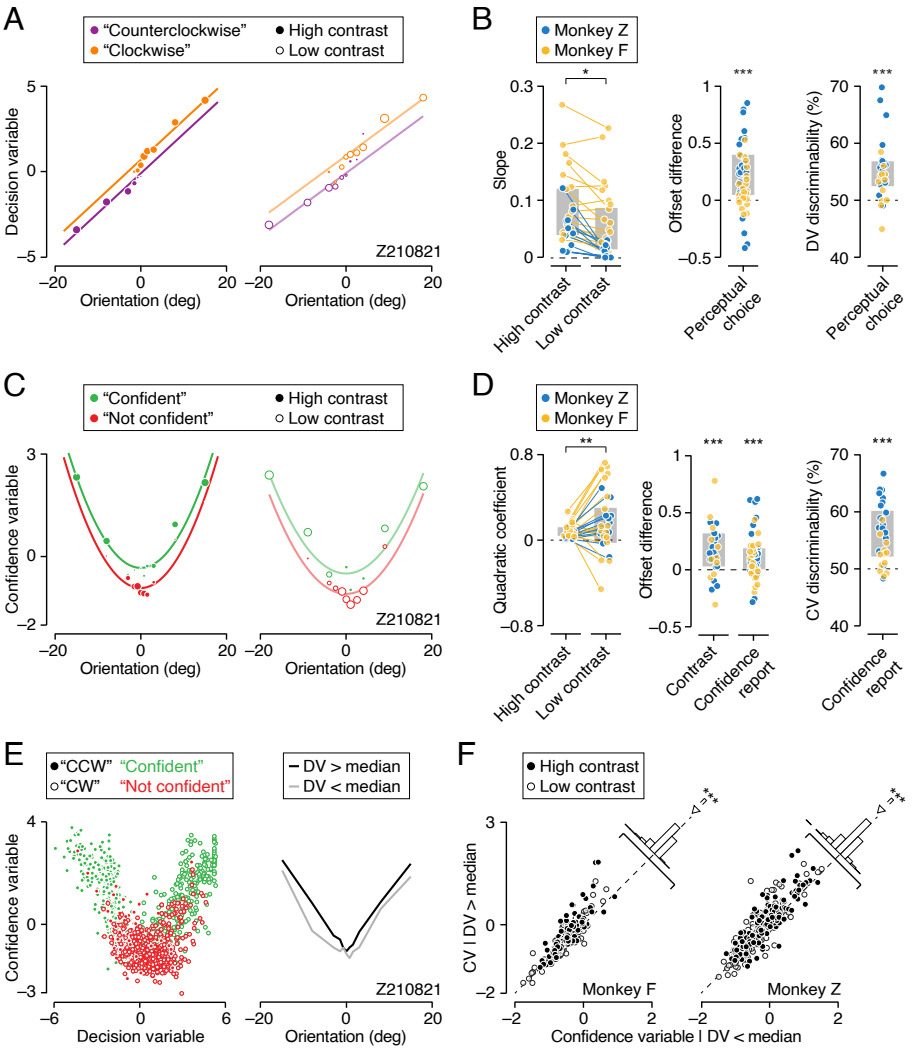

Fig. 4. The decoders' latent variables follow the predictions of a computational framework of perceptual decision confidence. (*A*) The latent variable of the nonlinear choice decoder plotted against stimulus orientation, for high and low contrast trials (*Left* vs. *Right*), conditioned on the animal's perceptual choice (yellow vs. blue) for an example dataset. (*B*) Summary of the decision variable's statistical structure across all recording sessions. (*Left*) Regression slope between orientation and average decision variable at low and high contrast. (*Middle*) Difference in regression offset between counterclockwise and clockwise responses. (*Right*) Accuracy of decision variable at discriminating behavioral decisions. Gray bars indicate interquartile range. (*C*) The latent variable of the nonlinear confidence decoder plotted against stimulus orientation, for high and low contrast trials (*Left* vs. *Right*), conditioned on the animal's confidence report (red vs. green) for an example dataset. (*D*) Summary of the confidence variable's statistical structure across all recording sessions. (*Left*) Quadratic coefficient of parabola fit to average confidence variable as a function of orientation at low and high contrast (*Middle*) Difference in offset of parabolas fit to average contrast or confidence conditioned confidence variables. (*Right*) Accuracy of confidence variable at discriminating confidence reports. Gray bars indicate interquartile range. (*E*) Direct comparison of the confidence variable and the decision variable. (*Left*) Each point represents a single trial in an example recording session. (*Right*) The mean confidence level plotted against stimulus orientation for trials with a decision variable value that is more (black) or less (gray) extreme than the stimulus-specific median. (*F*) Summary of the median-split analysis illustrated in panel (*E*) for all recording sessions. Each data point represents a single stimulus condition. n.s. not significant, $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$.

decision" are associated with a higher decision variable value (Fig. 4*A*, orange vs. purple). These effects were present in most of our datasets (Fig. 4 *B*, *Left* and *Middle*; median slope for high contrast stimuli = 0.064, $P < 0.001$; median reduction in slope for low contrast stimuli = 0.02, $P = 0.04$; median change in offset with perceptual choice = 0.22, $P < 0.001$, Wilcoxon signed rank test). The neurally decoded confidence variable exhibits a different structure. It varies parabolically with stimulus orientation (Fig. 4*C*). The width and offset of the parabola depend on stimulus contrast (Fig. 4*C*). And trials that culminate in a "high confidence" report are associated with a higher confidence variable value (Fig. 4*C*). Again, these effects were present in most of our datasets (Fig. 4 *D*, *Left* and *Middle*; median quadratic coefficient for high contrast stimuli = 0.002, $P < 0.001$; median change in this coefficient for low contrast stimuli = 0.001, $P = 0.005$; median change in offset for low contrast stimuli = 0.15, $P < 0.001$; median change in offset with confidence report = 0.11, $P < 0.001$ Wilcoxon signed rank test). To compare the strength of the association of both decision and confidence latent variables with the overt behavior, we computed their ability to discriminate behavior in the absence of stimulus variation (*Materials and Methods*). A discriminability value of 50% corresponds to chance performance, while 100% means that the behavior can be perfectly predicted from the latent variable. For both perceptual choices and confidence reports, we found a modest association (Fig. 4 *B* and *D*, *Right*; median decision variable discriminability = 54%, $P < 0.001$; median confidence variable discriminability = 56%, $P < 0.001$, Wilcoxon signed rank test). This association tended to be stronger for populations that contained more stimulus

information (*SI Appendix*, Fig. S3*B*). Thus, the decoders' latent variables provide insight into the neural processes underlying the overt choice-confidence reports.

Plotting the confidence variable against the decision variable for all trials of an example recording session reveals a U-shaped relationship, consistent with the proposed confidence computation (Fig. 4 *E*, *Left*). If this relationship arises from this computation, it should leave a signature even within the same stimulus conditions. Specifically, trials that yield a more extreme decision variable value should result in a higher value of the confidence variable. To test this prediction, we computed the median of the absolute value of the decision variable for every stimulus condition and computed the average of the confidence variable separately for trials above and below the median. As can be seen for an example recording session, more extreme decision variable values were systematically associated with higher levels of confidence (Fig. 4 *E*, *Right*). This effect was evident across all datasets, for both monkeys (Fig. 4*F*; median difference in confidence variable: monkey F = 0.06, $P < 0.001$, monkey Z = 0.11, $P < 0.001$). Our analysis ensured that choice and confidence signals occupied orthogonal neural dimensions (*Materials and Methods*). We therefore conclude that the brain's confidence computation evaluates the same sensory population activity that informed the decision.

Discussion

In this study, we investigated neural population activity in V1 during a perceptual confidence task. We sought to understand the brain's confidence assignment process. This process underlies the metacognitive ability to evaluate the quality of perceptual interpretations. We suggest that confidence arises from a nonlinear transformation of the same sensory signals that inform perceptual decisions. When the sensory population response is strong and unambiguous, this transformation results in high decision confidence (Fig. 1*I*, green zone). Conversely, when the sensory population response is weak or ambiguous, it results in low levels of confidence (Fig. 1*I*, red zone). Our proposal is supported by three distinct observations. First, nonlinear decoders of V1 population activity can predict monkeys' confidence in perceptual orientation judgments (Fig. 2*F*), establishing a direct link between the structure of sensory activity and decision confidence. Second, these decoders yield sensible and robust confidence outputs when presented with synthetic patterns of neural activity (Fig. 3), suggesting they capture the essence of the brain's confidence computation. Third, trials that yield stronger and less ambiguous V1 responses as evidenced by a neurally decoded decision variable also result in higher levels of neurally decoded confidence (Fig. 4).

In our task, subjects simultaneously reported decision content and confidence. This is not always the case in confidence experiments. Some tasks instead use sequential reports, meaning that the confidence report is probed at a later moment in time (7, 42). While this may seem like a subtle distinction between experimental designs, it has deep implications. In sequential reports tasks, information that was not available during decision formation can contribute to the final confidence assignment (43, 44). On the flip side, information that was available during decision formation may get lost during the intervening period (44). It thus seems likely that the connection between the sensory information that guides perceptual decisions and confidence in these decisions will be weaker in sequential reports tasks. In extreme cases, both types of confidence could be considered different subtypes (sometimes referred to as propositional and reflective confidence) (42). This study investigated propositional confidence in perceptual decisions. This is the type of confidence we rely on when deciding whether or not we have enough time to cross the street given the speed and distance of approaching traffic. It differs from the confidence we have in a good test score after talking to classmates who took the same test.

Our experimental paradigm enabled us to compare choice and confidence decoders trained on the same neural responses. We found that the relationship between sensory activity and decision confidence is as strong as the relationship between sensory activity and perceptual choice. Decoders predict the observers' confidence report and perceptual choice with similar accuracy across stimuli (Fig. 2 *E* and *F*) and in the absence of stimulus variation (Fig. 4 *B* and *D*). However, we suggest that there is a fundamental difference between both relationships. Perceptual choices arise from a neural computation downstream of sensory cortex that identifies the most likely interpretation of the sensory response; decision confidence instead arises from a computation that evaluates whether this sensory response will produce a reliable decision (Fig. 1*F*). Because these computations are distinct, they can manifest as mapping rules of sensory population activity that occupy orthogonal neural subspaces (Fig. 1*I*). Previous work offers indirect support for these ideas. Specifically, microstimulating sensory neurons in a postdecision wagering task altered monkeys' opt-out choice behavior as if they experienced a change in the sensory signal (10). These results suggest that the same sensory signals that inform decision content inform decision confidence. A different study employing the postdecision wagering paradigm found that pulvinar neurons represent decision confidence but not perceptual choice (9). Inactivating the pulvinar altered monkeys opt-out choice behavior but not their perceptual sensitivity (9). These results suggest that distinct brain circuits may be responsible for decision formation and confidence assignment. Consistent with this, studies that employed a postdecision time investment task found that orbitofrontal cortex neurons in rats play a similar role (12, 13) and represent an abstract decision confidence signal (14). Our work clarifies how neural circuits can extract such pure decision confidence signals from sensory population activity. However, note that in some experimental paradigms, the same neurons may represent decision content and confidence (7, 10).

We have shown that decoders of V1 activity capture behavioral effects of stimulus manipulations as well as behavioral variability under repeated presentations of the same stimulus (Fig. 4 *A* and *C*). Correlations between neural and behavioral responses can illuminate their causal relationship. However, these correlations can also arise for spurious reasons. Previous studies employing binary perceptual decision-making tasks found that choice-related signals in the sensory cortex reflect a combination of factors. These include the perceptual decision-making process (45–48), but also choice-aligned fluctuations in attention (49, 50), expectation (51), motor planning (52), and in other unspecified sources that impact sensory activity (53, 54). Could spurious reasons underlie the association between neural activity and overt behavior in our study? This concern is warranted. There is no statistical guarantee that the associations we reported primarily reflect the confidence assignment process. However, our task-design has a unique strength compared to binary decision-making tasks. Subjects generated a two-dimensional choice-confidence report. Our analysis ensured that both dimensions were orthogonal in neural population space. Nevertheless, we found that trials that yielded a more extreme perceptual decision variable also resulted in a higher level of confidence (Fig. 4 *E* and *F*). For this association to arise for spurious reasons, there would

need to be a factor whose properties are more complex than simple choice-alignment. It would need to jointly align with perceptual choices and confidence reports. While we cannot rule out this possibility, we hope that the richness of our behavioral paradigm has helped to expose the confidence computations implemented by neural circuits downstream of sensory cortex.

The ability to recognize which perceptual interpretations of the environment are at risk of being flawed is a hallmark of metacognition and as such often associated with higher intelligence. Our findings suggest that the confidence computations underlying this ability in the primate brain at least in part arise from simple deterministic transformations of sensory population activity. These transformations can in principle be realized in basic neural circuits. As such, our findings highlight how sophisticated behavior can arise from a cascade of simple operations. This is well appreciated in the domain of artificial intelligence. We suggest that it may also be true for certain components of biological intelligence. In general, metacognitive judgments are imperfect (26, 29, 31). Here, this was evident from the levels of meta-uncertainty displayed by our human and nonhuman subjects (Fig. 1*D*). As of yet, we do not know the neural causes of this. Metacognitive inefficiencies may originate in noise in sensory representations (35, 55–57). Alternatively, these inefficiencies may arise downstream of sensory cortex, for example from suboptimal confidence mapping rules (58). The task-paradigm and computational framework we have developed offer promising vehicles to address these outstanding questions and achieve a more complete understanding of the neural mechanisms that underlie and constrain our sense of confidence.

Materials and Methods

Animal Subjects. Our experiments were performed on two adult male macaque monkeys (*Macaca mulatta*, aged 7 and 10 y old at the time of the experiments). The animals were trained to perform an orientation discrimination task with saccadic eye movements as operant responses. Monkey F had previously participated in another research study (22, 23), Monkey Z had not previously participated in research studies. All training, surgery, and recording procedures were approved by the University of Texas Institutional Animal Care and Use Committee and conformed to the NIH Guide for the Care and Use of Laboratory Animals. Under general anesthesia, both animals were implanted with three custom-designed titanium head posts and a titanium recording chamber which enabled access to V1 (59).

Apparatus. The monkeys were seated in a custom-designed primate chair in front of a gamma-corrected 22-inch CRT monitor (Sony Trinitron, model GDM-FW900), with their heads restrained using three surgical implants. Stimuli were shown on the CRT monitor, which was positioned approximately 60 cm away from the monkeys' heads. The CRT had a resolution of 1,280 by 1,024 pixels with a refresh rate of 75 Hz. Eye position was tracked continuously with an infrared eye tracking system at 1 kHz (EyeLink 1000, SR Research). Stimuli were presented using the Psychophysics Toolbox (60) in MATLAB (MathWorks). Neural activity was recorded using the Plexon OmniPlex System (Plexon). Precise temporal registration of task events and neural activity was obtained through a Datapixx system (Vpixx). All of these systems were integrated using the PLDAPS software package (61) (https://github.com/HukLab/PLDAPS). An analogous setup was used for the human psychophysical experiment, except that head position was stabilized using a chin rest and the monitor was a Hewlett Packard, model A7217.

Visual Stimuli. We constructed oriented visual stimuli by bandpass filtering 3-D luminance noise with a filter organized on a velocity plane. The filter's spatial frequency passband was centered at a spatial frequency of 2.5 cycles per degree and had a bandwidth of 0.5 octaves. Its velocity passband was centered at a speed of 2.5° per second (which corresponds to a central temporal frequency of 5 Hz) and had a bandwidth of 1 octave. The filter's orientation bandwidth was 3°. For each stimulus condition, the stimulus set contained five unique filtered noise movies. Each orientation discrimination experiment included stimuli that varied in orientation and contrast. Stimulus contrast was computed by normalizing the summed orientation amplitude spectrum of each stimulus frame with the summed amplitude spectrum of a reference grating with matching spatial frequency. There was one high and one low contrast level per experiment. The high contrast value was constant across experiments, the low contrast value varied somewhat across experiments in an attempt to elicit a difference in the slope of high and low contrast psychometric curves. The ratio between low and high contrast values ranged between 0.37 and 0.63 with a median value of 0.56. The high contrast stimuli spanned a range of 11 different orientations, the number of low contrast orientations varied across experiments (15 experiments had 11 orientations, 7 had 9, 3 had 7, and 4 had 2). Orientation values were chosen to maintain approximately equivalent performance across high and low contrast stimuli. Stimuli were presented with equal frequency in 14 of 29 recordings. In 15 of 29 recordings, stimuli were presented with equal frequency, except for three high contrast stimulus orientations in which a single filtered noise iteration was overrepresented. For all illustrated example sessions (Figs. 1 *B* and *G*, 3*A*, and 4 *A* and *C* and *SI Appendix* Fig. S1 *A* and *E*) stimuli were presented with equal frequency.

Fixation Task. At the beginning of each recording session, monkeys first performed a passive fixation task. We used a hand-mapping procedure to estimate the location of the spatial receptive fields of visually responsive units. The average receptive field center estimate served as the center location for the visual stimuli presented during the rest of the recording session. We conducted an initial fixation task during which we presented sinusoidal gratings of varying orientation for 1,000 ms each. This was followed by the orientation discrimination task.

Orientation Discrimination Task. The orientation-discrimination task is a variant of classical visual categorization tasks in which the subject uses a saccadic eye movement as operant response (51, 62, 63). We used a richer version of this task in which subjects are invited to additionally report their confidence in each perceptual decision. Each trial began when the subject fixated a small white dot at the center of the screen. Upon fixation, four black choice targets appeared–one in each quadrant of the screen. Targets to the left of the fixation point represented counterclockwise decisions, targets to the right clockwise decisions. Upper choice targets indicated high decision confidence, lower choice targets low confidence. After a variable prestimulus fixation period, the stimulus appeared in the near periphery (average eccentricity: monkey F = 4.32°, monkey Z = 3.00°) for 500 ms. Subjects judged the orientation of the stimulus relative to vertical. The stimulus then disappeared along with the fixation mark and subjects reported their decision and confidence with a saccadic eye movement to one of the four choice targets. Auditory feedback was given to indicate the accuracy of the decision and the chosen level of confidence. Specifically, the tone differed for correct and incorrect trials and the sound was played twice in quick succession for high confidence reports. If the decision was correct, a liquid reward was delivered via a computer operated reward system (New Era). Vertically oriented stimuli received random feedback. Trials in which the monkey did not saccade to one of the choice targets within 3 seconds were aborted. To incentivize meaningful confidence reports, there were four possible reward levels. It required one correct decision to move from level 1 to 2, 3 further correct decisions to move from level 2 to 3, and 3 more to reach level 4. Subjects remained at level 4 until they reported an incorrect decision with high confidence, which reset the score to level 1. Resets could occur at any level. The higher the reward level, the larger the reward for a correct decision. In addition, correct decisions reported with high confidence were rewarded more generously than correct decisions reported with low confidence. High confidence rewards for each level were 0.04, 0.16, 0.32, 0.64 ml for monkey F and 0.116, 0.232, 0.464, 0.928 ml for monkey Z. Low confidence rewards were a scalar function of high confidence reward. This scalar value varied across sessions and was adjusted to titrate the proportion of high and low confidence responses (average 0.68 ± 0.04 for monkey Z and average 0.82 ± 0.04 for monkey F. Lower scalar values encouraged more high confidence responses due to a larger reward difference between high and low

confidence. Each trial, the current reward level was indicated to the monkey by the duration of the prestimulus fixation period (the lower the reward level, the longer this duration). Both monkeys managed to stay at the highest reward level for the majority of trials (fraction of trials at reward level 4: monkey F = 64%, monkey Z = 80%). We conducted 12 successful recordings from monkey F and 17 from monkey Z (average number of reward level 4 trials per session, monkey F = 753; monkey Z = 1,026). In four sessions, only two different low contrast stimulus orientations were used. The low-contrast psychometric functions from these four sessions were excluded from Fig. 3*B*.

Human Psychophysical Experiment. Nineteen human subjects (10 male, 9 female; ages 19 to 32) with normal or corrected-to-normal vision participated in the experiment. The experimental protocol was approved by the local ethics committee (Institutional Review Board of The University of Texas at Austin) and all participants gave written informed consent. The human behavioral task was the same as the animals' orientation discrimination task, with the exception that the stimulus was presented more centrally and subjects earned points instead of liquid reward (points per high confidence correct response as a function of reward level: 2, 4, 8, 16; points per low confidence correct response: 1, 2, 4, 8). Human subjects began by completing 175 training trials. We used these initial trials to estimate each subject's orientation sensitivity. This sensitivity estimate determined the range of stimulus orientations used in the main experiment. We chose the range such that the subjects' overall task performance level would resemble that of the animals. This procedure worked well for all but two subjects for whom we discarded the first block of trials. Subjects performed the main task in subblocks of 50 trials. Subjects were rewarded with monetary points in the same manner as the macaques were rewarded with liquid reward, and received analogous auditory feedback at the end of each trial. Every 50 trials, subjects were given additional visual feedback on their total point count. Subjects completed three blocks of 1,100 trials. Eleven subjects judged the same filtered noise stimuli as the monkeys did, eight subjects were presented with deterministic sinusoidal gratings instead. Because meta-uncertainty did not systematically differ across both groups of subjects, we included all these datasets in our analysis except for the two subjects who had poorly calibrated first blocks yielding a total of 17 human observers in this analysis.

Behavioral Analysis. Each session we measured observers' behavioral capability to discriminate stimulus orientation by fitting the relationship between stimulus orientation and probability of a "clockwise" choice with a psychometric function consisting of a lapse rate and a cumulative Gaussian function. To compare the behavioral capability associated with low and high confidence reports, each psychometric function had its own steepness parameter (the inverse of the SD of the cumulative Gaussian). The parameters controlling lapse rate and the point of subjective equality (the mean of the cumulative Gaussian) were shared across both psychometric functions. Model parameters were optimized by maximizing the likelihood over observed data, assuming responses arise from a Bernoulli process. For the analysis documented in Fig. 1*C*, each dataset was analyzed independently.

CASANDRE (26) is a two-stage process model in which comparison of a noisy internal representation of the stimulus (V_d) with a decision criterion (C_d) yields a perceptual choice (3). The decision variable is on average unbiased, but it is subject to Gaussian noise with SD σ_d. This noise determines the level of perceptual uncertainty. A confidence variable (V_c) is computed by normalizing the strength of the decision variable ($|V_d - C_d|$) with an estimate of perceptual uncertainty ($\widehat{\sigma_d}$), modeled as a sample from a lognormal distribution whose mean is the true perceptual uncertainty, σ_d, and whose SD is σ_m. Because σ_m reflects the uncertainty about perceptual uncertainty, it is referred to as meta-uncertainty. It is the sole parameter in the model that limits the quality of confidence judgments (with higher values indicating lower metacognitive ability). For each dataset, we obtained an estimate of the subject's level of meta-uncertainty by fitting the CASANDRE model to the choice-confidence data using a fitting procedure described previously (26). In brief, the model had eight parameters: the SD of the decision variable (σ_d) (one per contrast level, two in total), the decision criterion (C_d) (one per contrast level, two in total), the level of meta-uncertainty (σ_m), the confidence criterion (C_c) (we allowed for choice-dependent asymmetries, two in total), and lapse rate (λ). For each dataset, we computed the log-likelihood of a given set of model parameters across all choice-confidence reports and used an iterative procedure to identify the most likely set of parameter values (specifically, the interior point algorithm used by the Matlab function "fmincon"). For the analysis documented in Fig. 1 *D* and *E*, the first and last blocks of trials completed by the human subjects were analyzed independently.

Electrophysiological Recordings. During the orientation discrimination task, we recorded extracellular spiking activity from populations of V1 neurons through a chronically implanted recording chamber. Every recording session, we used a microdrive (Thomas recording) to mechanically advance one or two linear electrode arrays (Plexon S- and V-probes; 32 or 24 contacts) into the brain. We positioned the linear arrays so that they roughly spanned the cortical sheet and removed them after each recording session. Continuous neural data were acquired and saved to disk from each channel (sampling rate 30 kHz, Plexon Omniplex System). To extract responses of individual units, we performed offline spike sorting. We first automatically spike-sorted the data with Kilosort (64), followed by manual merging and splitting as needed (with the "phy" user interface, https://github.com/kwikteam/phy). Given that the electrodes' position could not be optimized for all contact sites, most of our units probably consist of multineuron clusters. We used the fixation task to identify visually responsive units whose activity selectively depended on stimulus orientation. We measured each unit's response by expressing spike times relative to stimulus onset and counting spikes within a 1,000-ms window following response onset. For each unit, we chose a response latency by maximizing the stimulus-associated response variance (65). We visually inspected orientation tuning curves and excluded untuned units from further analysis.

Decoders. We ensured that both linear and nonlinear choice decoders could not use decision confidence to predict choices and that confidence decoders could not use perceptual choice to predict confidence. Specifically, we orthogonalized choice and confidence information in the training trials by maintaining a fixed ratio of high and low confidence reports across clockwise and counterclockwise choices and a fixed ratio of clockwise and counterclockwise choices across high and low confidence reports. To do so, we randomly selected trials from underrepresented trial types (e.g "high-confident counterclockwise") and concatenated them to the training set. To minimize potentially confounding influences of cross-trial variation in the animals' motivation, attention, and alertness, we only included "reward level 4" trials in the training set. Training sets on average contained 1,088 trials and were identical for linear and nonlinear decoding analyses.

Linear Decoders. To assess how well the recorded populations could support the perceptual task, we trained linear *stimulus decoders* to discriminate between clockwise and counterclockwise stimuli. We used all stimuli whose orientation differed from 0°. We first z-scored each unit's spike counts. We then used these z-scored responses to estimate the set of linear weights, $\mathbf{w} = (w_1, \ldots w_n)$, where n is the neuronal population size, that best separate clockwise and counterclockwise stimulus response patterns, assuming a multivariate Gaussian response distribution:

$$\mathbf{w} = \Sigma^{-1} s, \qquad [1]$$

Where s is the mean difference of the stimulus-category conditioned z-scored responses and Σ is the covariance matrix of the z-scored responses. The decoder weights are calculated from observed trials. To avoid double-dipping, we excluded the trial under consideration from the calculation and solely used all other trials to estimate the weights. This way, we obtained a "cross-validated" stimulus judgment from the linear stimulus decoder for each trial. We quantified how well these decoders captured the animals' behavior by computing the fraction of consistent perceptual choices between behavior and stimulus decoder's predictions and subtracting the fraction expected by chance based on the decoder's and the animal's overall success rate (Fig. 2*B*). We quantified how well the output of the stimulus decoders could predict confidence behavior by computing the Pearson correlation between the binary vector of behavioral confidence reports and the binary vector of the stimulus decoder's

choice predictions (Fig. 2*C*). In a later analysis, we compared nonlinear *choice* and *confidence decoders* with their linear counterparts (Fig. 2 *E* and *F*). For this analysis, we used exactly the same set of training and hold-out trials for the linear decoders as we used for the nonlinear decoders.

Nonlinear Decoders. We trained feed-forward multilayer perceptron neural networks on z-scored V1 responses to either predict the animals' perceptual choice or their confidence report. We implemented networks within the TensorFlow framework using the AdamW optimizer with an objective to minimize binary cross-entropy. Models consisted of 1 hidden layer with 15 hidden units per layer, had a dropout rate between layers of 0.1, and the learning rate was set to 0.001. We explored various hyperparameter settings and found the results presented here to be robust across settings. We trained networks on 80% of trials (training/validation set) and obtained a cross-validated prediction on the held-out 20% of trials, rotating trials between training and held-out set such that each trial had a cross-validated prediction (Fig. 2 *E* and *F*). To ensure that every trial would be part of the hold out set, we trained 30 different networks per dataset. Just like we did for the linear decoders, we solely used cross-validated choice and confidence predictions in our analysis. For each trial, we selected the decoder's prediction from one random network in which this trial was held out. We found 30 networks to be sufficient for every trial to be held out at least once.

To interrogate whether the confidence decoders extracted meaningful information from V1 responses, we compared the slope of the psychometric function conditioned on the confidence decoder's output (Fig. 3*A*). This comparison is most reliable when both psychometric functions contain a similar number of trials. We achieved this by using the median of the latent confidence variable as confidence criterion. We did this for both high and low contrast trials.

We probed the confidence decoders with synthetic patterns of neural activity (Fig. 3*C*). To create these patterns, we first computed the cross-trial average firing rate per unit for a given stimulus orientation using only high contrast, reward level 4 trials. We manipulated the gain of these responses by multiplying this average population response vector with a single scalar factor. For each recording session, we used this scalar-multiplied activity as input to one randomly picked network out of 30 which has previously been trained on real, nonscalar-multiplied data (Fig. 3*D*).

We studied the decoders' latent variables (Fig. 4). This analysis involved computing a discriminability index. To do so, we first z-scored the latent variables per stimulus condition, thus removing stimulus-driven effects. We then created two groups of trials based on the animals' behavioral reports (either their perceptual choice or their confidence report). We included all stimulus conditions for which both response options had been used at least 5 times. Finally, we computed the area under the curve for both sets of trials (45).

Data, Materials, and Software Availability. The data and analysis code that support the findings of this study are available in the public GitHub repository associated with this manuscript: https://github.com/zoebsinger/V1-Confidence-Physiology.git (66).

ACKNOWLEDGMENTS. This work was supported by the US NSF (Graduate Research Fellowship to Z.M.B.-S., and CAREER award 2146369 to R.L.T.G.), the US NIH (grant nos. T32 EY021462 to Z.M.B.-S and C.M.Z., K99 EY032102 to C.M.Z., EY032999 to R.L.T.G., and Intramural Research Program of the NIH, National Eye Institute to C.M.Z.) and the Whitehall Foundation (grant no. UTA19-000535 to R.L.T.G.).

1. S. Hecht, S. Shlaer, M. H. Pirenne, Energy, quanta, and vision. *J. Gen. Physiol.* **25**, 819–840 (1942).
2. W. P. Tanner Jr., J. A. Swets, A decision-making theory of visual detection. *Psychol. Rev.* **61**, 401–409 (1954).
3. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966), vol. 1.
4. L. Festinger, Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *J. Exp. Psychol.* **32**, 291–306 (1943).
5. P. Mamassian, Visual confidence. *Annu. Rev. Vis. Sci.* **2**, 459–481 (2016).
6. A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
7. R. Kiani, M. N. Shadlen, Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
8. P. Middlebrooks, M. Sommer, Metacognition in monkeys during an oculomotor task. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 325–337 (2011).
9. Y. Komura, A. Nikkuni, N. Hirashima, T. Uetake, A. Miyamoto, Responses of pulvinar neurons reflect asubject's confidence in visual categorization. *Nat. Neurosci.* **16**, 749–755 (2013).
10. C. R. Fetsch, R. Kiani, W. T. Newsome, M. N. Shadlen, Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* **83**, 797–804 (2014).
11. B. Odegaard *et al.*, Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1588–E1597 (2018).
12. A. Kepecs, N. Uchida, H. A. Zariwala, Z. F. Mainen, Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
13. A. Lak *et al.*, Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
14. P. Masset, T. Ott, A. Lak, J. Hirokawa, A. Kepecs, Behavior- and modality-general representation of confidence in orbitofrontal cortex. *Cell* **182**, 112–126.e18 (2020).
15. B. A. Purcell, R. Kiani, Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4531–E4540 (2016).
16. J. Drugowitsch, A. G. Mendonça, Z. F. Mainen, A. Pouget, Learning optimal decisions with confidence. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24872–24880 (2019).
17. M. Sarafyazd, M. Jazayeri, Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, eaav8911 (2019).
18. B. Bahrami *et al.*, Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
19. N. K. Logothetis, J. D. Schall, Neuronal correlates of subjective visual perception. *Science* **245**, 761–763 (1989).
20. M. N. Shadlen, W. T. Newsome, Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* **86**, 1916–1936 (2001).
21. J. I. Gold, M. N. Shadlen, The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
22. J. A. Charlton, R. L. T. Goris, Abstract deliberation by visuomotor neurons in prefrontal cortex. *Nat. Neurosci.* **27**, 1–9 (2024).
23. T. A. Langlois, J. A. Charlton, R. L. T. Goris, Bayesian inference by visuomotor neurons in the prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2420815122 (2025).
24. L. S. Geurts, J. R. H. Cooke, R. S. van Bergen, J. F. M. Jehee, Subjective confidence reflects representation of Bayesian probability in cortex. *Nat. Hum. Behav.* **6**, 1–12 (2022).
25. C. R. Fetsch *et al.*, Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *eLife* **7**, e36523 (2018).
26. Z. M. Boundy-Singer, C. M. Ziemba, R. L. T. Goris, Confidence reflects a noisy decision reliability estimate. *Nat. Hum. Behav.* **7**, 142–154 (2023).
27. M. Vivar-Lazo, C. R. Fetsch, Neural basis of concurrent deliberation toward a choice and degree of confidence. bioRxiv [Preprint] (2024). https://www.biorxiv.org/content/10.1101/2024.08.06.606833.abstract (Accessed 27 September 2024).
28. D. H. Hubel, T. N. Wiesel, Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968).
29. T. O. Nelson, A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.* **95**, 109–133 (1984).
30. S. M. Fleming, H. C. Lau, How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
31. M. Shekhar, D. Rahnev, The nature of metacognitive inefficiency in perceptual decision making. *Psychol. Rev.* **128**, 45–70 (2021).
32. P. Mamassian, V. de Gardelle, Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychol. Rev.* **129**, 976–998 (2022).
33. B. Caziot, P. Mamassian, Perceptual confidence judgments reflect self-consistency. *J. Vis.* **21**, 8 (2021).
34. G. Orban, P. Berkes, J. Fiser, M. Lengyel, Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
35. O. J. Henaff, Z. M. Boundy-Singer, K. Meding, C. M. Ziemba, R. L. T. Goris, Representation of visual uncertainty through neural gain variability. *Nat. Commun.* **11**, 2513 (2020).
36. D. Festa, A. Aschner, A. Davila, A. Kohn, R. Coen-Cagli, Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nat. Commun.* **12**, 3635 (2021).
37. R. L. T. Goris, R. Coen-Cagli, K. D. Miller, N. J. Priebe, M. Lengyel, Response sub-additivity and variability quenching in visual cortex. *Nat. Rev. Neurosci.* **25**, 237–252 (2024).
38. R. Geirhos *et al.*, Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
39. B. C. Skottun, A. Bradley, G. Sclar, I. Ohzawa, R. D. Freeman, The effects of contrast on visual orientation and spatial frequency discrimination: A comparison of single cells and behavior. *J. Neurophysiol.* **57**, 773–786 (1987).
40. D. J. Heeger, Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
41. D. Ferster, K. D. Miller, Neural mechanisms of orientation selectivity in the visual cortex. *Annu. Rev. Neurosci.* **23**, 441–471 (2000).
42. S. M. Fleming, Metacognition and Confidence: A Review and Synthesis. *Annu. Rev. Psychol.* **75**, 241–268 (2024).
43. A. Resulaj, R. Kiani, D. M. Wolpert, M. N. Shadlen, Changes of mind in decision-making. *Nature* **461**, 263–266 (2009).
44. T. Balsdon, V. Wyart, P. Mamassian, Metacognitive evaluation of postdecisional perceptual representations. *J. Vis.* **24**, 2 (2024).
45. K. H. Britten, W. T. Newsome, M. N. Shadlen, S. Celebrini, J. A. Movshon, A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13**, 87–100 (1996).
46. J. V. Dodd, K. Krug, B. G. Cumming, A. J. Parker, Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *J. Neurosci. Off. J. Soc. Neurosci.* **21**, 4809–4821 (2001).
47. A. G. Bondy, R. M. Haefner, B. G. Cumming, Feedback determines the structure of correlated variability in primary visual cortex. *Nat. Neurosci.* **21**, 598–606 (2018).
48. R. M. Haefner, P. Berkes, J. Fiser, Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–690 (2016).

49. H. Nienborg, B. G. Cumming, Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature* **459**, 89–92 (2009).
50. K. R. Quinn, L. Seillier, D. A. Butts, H. Nienborg, Decision-related feedback in visual cortex lacks spatial selectivity. *Nat. Commun.* **12**, 4473 (2021).
51. R. L. T. Goris, C. M. Ziemba, G. M. Stine, E. P. Simoncelli, J. A. Movshon, Dissociation of choice formation and choice-correlated activity in macaque visual cortex. *J. Neurosci.* **37**, 5195–5203 (2017).
52. P. Laamerad, L. D. Liu, C. C. Pack, Decision-related activity and movement selection in primate visual cortex. *Sci. Adv.* **10**, eadk7214 (2024).
53. A. J. Levi, Y. Zhao, I. M. Park, A. C. Huk, Sensory and choice responses in MT distinct from motion encoding. *J. Neurosci.* **43**, 2090–2103 (2023).
54. C. M. Ziemba *et al*., Neuronal and behavioral responses to naturalistic texture images in macaque monkeys. *The J. Neurosci.* **44**, e0349242024 (2024).
55. P. Heggelund, K. Albus, Response variability and orientation discrimination of single cells in striate cortex of cat. *Exp. Brain Res.* **32**, 197–211 (1978).
56. R. L. T. Goris, J. A. Movshon, E. P. Simoncelli, Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865 (2014).
57. Z. M. Boundy-Singer, C. M. Ziemba, O. J. Hénaff, R. L. T. Goris, How does V1 population activity inform perceptual certainty? *J. Vis.* **24**, 12 (2024).
58. J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, A. Pouget, Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
59. D. L. Adams, J. R. Economides, C. M. Jocson, J. M. Parker, J. C. Horton, A watertight acrylic-free titanium recording chamber for electrophysiology in behaving monkeys. *J. Neurophysiol.* **106**, 1581–1590 (2011).
60. D. H. Brainard, The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
61. K. Eastman, A. Huk, PLDAPS: A hardware architecture and software toolbox for neurophysiology requiring complex visual stimuli and online behavioral control. *Front. Neuroinf.* **6**, 00001 (2012).
62. W. T. Newsome, K. H. Britten, J. A. Movshon, Neuronal correlates of a perceptual decision. *Nature* **341**, 52–54 (1989).
63. H. Nienborg, B. G. Cumming, Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *J. Neurosci.* **34**, 3579–3585 (2014).
64. M. Pachitariu, N. A. Steinmetz, S. N. Kadir, M. Carandini, K. D. Harris, "Fast and accurate spike sorting of high-channel count probes with KiloSort" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), vol. 29.
65. M. A. Smith, N. J. Majaj, J. A. Movshon, Dynamics of motion signaling by neurons in macaque area MT. *Nat. Neurosci.* **8**, 220–228 (2005).
66. Z. M. Boundy-Singer, C. M. Ziemba, R. L. T. Goris, V1-Confidence-Physiology. GitHub. https://github.com/zoebsinger/V1-Confidence-Physiology/. Deposited 28 April 2025.