Distinct encoding of decision confidence in human medial prefrontal cortex

Dan Bang[a,1] and Stephen M. Fleming[a,b]

[a]Wellcome Centre for Human Neuroimaging, University College London, WC1N 3BG London, United Kingdom and [b]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, WC1B 5EH London, United Kingdom

Edited by Earl K. Miller, Massachusetts Institute of Technology, Cambridge, MA, and accepted by Editorial Board Member Michael S. Gazzaniga May 3, 2018 (received for review January 22, 2018)

Our confidence in a choice and the evidence pertaining to a choice appear to be inseparable. However, an emerging computational consensus holds that the brain should maintain separate estimates of these quantities for adaptive behavioral control. We have devised a psychophysical task to decouple confidence in a perceptual decision from both the reliability of sensory evidence and the relation of such evidence with respect to a choice boundary. Using human fMRI, we found that an area in the medial prefrontal cortex, the perigenual anterior cingulate cortex (pgACC), tracked expected performance, an aggregate signature of decision confidence, whereas neural areas previously proposed to encode decision confidence instead tracked sensory reliability (posterior parietal cortex and ventral striatum) or boundary distance (presupplementary motor area). Supporting that information encoded by pgACC is central to a subjective sense of decision confidence, we show that pgACC activity does not simply covary with expected performance, but is also linked to within-subject and between-subject variation in explicit confidence estimates. Our study is consistent with the proposal that the brain maintains choice-dependent and choice-independent estimates of certainty, and sheds light on why dysfunctional confidence often emerges following prefrontal lesions and/or degeneration.

confidence | decision making | fMRI | medial prefrontal cortex

Decisions are often made in the face of uncertainty and in the absence of immediate feedback. Accompanying such decisions is a sense of confidence in having made the correct choice, which can be used to guide behavior (1, 2). For example, after having made a difficult choice, an animal might correctly estimate that its decision is unlikely to be correct and thus avoid wasting time waiting for a reward that may not arrive (3). Humans may communicate such estimates of decision confidence to make more accurate decisions together than are made alone (4). Despite widespread agreement that decision confidence is a useful quantity for adaptive control of behavior, neurobiological support for a distinct encoding of decision confidence is lacking.

Several computational models propose that decision confidence reflects an internal estimate of the probability that a choice is correct (1, 2). A ubiquitous paradigm for studying the neural basis of this computation is sensory psychophysics. On a typical trial, subjects first make a categorical choice about an ambiguous stimulus, such as deciding whether a cloud of dots is moving left or right or whether a contrast grating is tilted counterclockwise or clockwise of vertical. Subjects then make a secondary judgment that requires estimating the probability that the initial choice is correct. For example, subjects may have to decide whether to opt out of the choice for a sure but small reward or, in humans, explicitly estimate their confidence in the choice. Variation in these confidence-based behaviors may be induced by manipulation of task features, such as stimulus reliability (e.g., percentage of coherently moving dots) or the distance between the stimulus and a choice boundary (e.g., angular deviation from vertical axis), and/or intrinsic stochasticity in neural processing.

This general approach has been used with various species and techniques to identify neural areas that are involved in, or at least predict, confidence-based behaviors. In monkeys, single-unit recording and functional inactivation have identified the lateral intraparietal sulcus (5), thalamic pulvinar (6), supplementary eye fields (7), and dopaminergic neurons in the substantia nigra (8) as contributing to confidence-based behaviors, such as opt-out responses and postdecision wagers. In rodents, similar approaches have identified the orbitofrontal cortex as involved in confidence-based behaviors, such as willingness to wait for a potential reward (3). In humans, fMRI has identified neural areas that track explicit confidence estimates, including the striatum (9), medial prefrontal cortex (10, 11), dorsal anterior cingulate cortex (12), and rostrolateral prefrontal cortex (11, 12).

In the aforementioned tasks, there are at least two distinct components to a computation of decision confidence: the reliability of sensory evidence and the relation of such evidence with respect to a choice boundary (1). When the choice boundary is known in advance or fixed, one of these components may be a sufficient statistic for estimating decision confidence. For example, it has been proposed that algorithmically, decision confidence is a function of the sensory evidence in favor of a choice and elapsed time (5) or the (absolute) distance between such evidence and a choice boundary (3, 6, 9). However, such close coupling of decision confidence and its component parts can make it difficult to evaluate the contribution of distinct neural signals to confidence-based behaviors. For example, neural activity in the parietal cortex might predict an animal's opt-out responses (5), because the area encodes a probability distribution over sensory states given sensory evidence—a key component of decision confidence—rather than decision confidence per se. Conversely, neural activity in the orbitofrontal cortex might predict an animal's willingness to wait for a potential reward (3), because the area encodes the distance between sensory

Significance

Recent computational models propose that our sense of confidence in a choice reflects an estimate of the probability that the choice is correct. However, it has proven difficult to experimentally separate decision confidence from its component parts, such as our certainty about perceptual evidence or choice requirements. We have devised a task to dissociate these quantities and isolate a distinct encoding of decision confidence in the medial prefrontal cortex of the human brain. We show that activity in this area not only tracks expected performance on a task, but also is related to both within-subject and between-subject variation in a subjective sense of confidence. Our study illuminates why dysfunctional confidence often emerges following damage to the prefrontal cortex.

Author contributions: D.B. and S.M.F. designed research; D.B. performed research; D.B. contributed new reagents/analytic tools; D.B. analyzed data; and D.B. and S.M.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. E.K.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. Email: danbang.db@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1800795115/-/DCSupplemental.

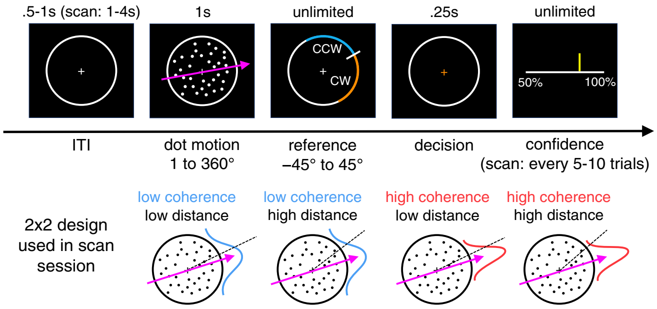

Fig. 1. Continuous-direction random-dot motion task with variable reference. Subjects had to judge whether the net direction of dot motion was counterclockwise or clockwise to a reference that appeared after stimulus offset. We varied the percentage of coherently moving dots and the absolute angular distance between the motion direction and the reference. In the prescan session, confidence estimates (50–100% in steps of 10%) were elicited on every trial. In the scan session, confidence estimates were elicited every 5–10 trials. Information on stimulus calibration is provided in *SI Appendix*, Fig. S1.

evidence and a choice boundary—another component of decision confidence—rather than decision confidence per se. While the distinction between these encoding schemes may sometimes prove moot, there are many situations in which decision confidence cannot be readily estimated from the choice process and/or must be represented separately (1), such as when a choice boundary is not known at the start of a trial (13) or when faced with a series of interdependent choices (14).

Here we tested for a distinct encoding of decision confidence in the human brain by devising a task that isolates decision confidence from its component parts. Subjects performed a continuous-direction, variable-reference random dot-motion task that in aggregate separated the probability that a motion discrimination judgment was correct (expected performance) from the reliability of a percept of motion direction (sensory reliability) and the distance between a motion percept and a choice boundary (boundary distance). Our approach builds on previous behavioral paradigms that have examined components of decision confidence (e.g., stimulus mean and variance) (15, 16). Using fMRI, we show that both aggregate and single-trial signatures of decision confidence are tracked by an area in the medial prefrontal cortex, the perigenual anterior cingulate cortex (pgACC).

Results

Experimental Isolation of Decision Confidence. Subjects (n = 32) viewed a field of moving dots inside a circular aperture (Fig. 1). On each update of the motion display, a fraction of dots moved coherently in a prespecified direction, sampled anew on each trial from the range 1–360°, whereas the remainder moved randomly. After the motion display, a line transecting the aperture appeared. Subjects had to decide whether the net direction of dot motion was counterclockwise (CCW) or clockwise (CW) to this reference. Finally, the subjects indicated their confidence in the choice being correct.

Using a factorial design, we varied the fraction of coherently moving dots (coherence) and the absolute angular distance between the motion direction and the reference (distance). Our aim was to separate a subject's internal estimate of the probability that a motion discrimination judgement is correct from the reliability of his or her percept of motion direction and the distance between that motion percept and a choice boundary—components that bear on a confidence computation but are in and of themselves not sufficient for confidence estimation in our task.

Subjects performed the task in separate prescan and scan sessions. In the prescan session, we calibrated a set of coherences and distances (2 × 4 design) associated with target levels of choice accuracy and evaluated the behavioral effects of our task manipulation. In the scan session, we simplified the design to ensure sufficient trials per condition for fMRI analysis, using a subset of the calibrated task parameters (2 × 2 design). To avoid confounding neural responses related to decision confidence with neural responses related to explicit reports, we elicited confidence estimates every 5–10 trials.

Behavioral Validation of the Experimental Approach. We first validated that subjects' expected performance varied with changes in coherence and distance. Indeed, we found that choice accuracy was affected by both factors; subjects were more likely to be correct when coherence was high and distance was high [Fig. 2*A*, logistic regression; coherence: t (31) = 11.46, $P < 0.001$; distance: t (31) = 19.00, $P < 0.001$; interaction: t (31) = 9.85, $P < 0.001$]. These effects were mirrored in choice reaction time; subjects made faster decisions when coherence was high and distance was high [Fig. 2*B*, linear regression; coherence: t (31) = −4.97, $P < 0.001$; distance: t (31) = −11.58, $P < 0.001$; interaction: t (31) = −9.26, $P < 0.001$].

Critically, consistent with the aim of our design, the effects of coherence and distance on choice accuracy were reflected in explicit confidence estimates; subjects reported higher confidence when coherence was high and distance was high [Fig. 2*C*, ordinal regression; coherence: t (31) = 8.45, $P < 0.001$; distance: t (31) = 11.15, $P < 0.001$; interaction: t (31) = 9.40, $P < 0.001$]. These confidence effects survived controlling for choice reaction time and nuisance factors, such as the initial position of the confidence marker and the cardinality of motion direction (*SI Appendix*, Fig. S2). Finally, the effects of coherence and distance on choice behavior and confidence estimates were replicated in the scan session (*SI Appendix*, Fig. S3).

To further unpack the drivers of the choice process, we modeled subjects' responses using a hierarchical instantiation of the drift-diffusion model (DDM) (17). We remained agnostic as to how our task affected DDM parameters by fitting drift rate (signal-to-noise ratio of evidence accumulation), threshold (amount of evidence needed for a choice), and nondecision time (e.g., stimulus encoding and response preparation) separately for each condition. In both sessions, we found that only drift rate varied between conditions, whereas threshold and nondecision time were stable; see the model predictions in Fig. 2 *A* and *B* and the parameter estimates in *SI Appendix*, Fig. S4. In our task, the momentary evidence entering into the accumulation process can be thought of as the signed difference between a noisy sample from a sensory representation of motion direction held in visual short-term memory and a choice boundary (18). Here the reliability of the sensory representation, controlled by coherence, and the placement of the choice boundary, controlled by distance, jointly determine the signal-to-noise ratio of the accumulation process and thereby the probability that a choice is correct. Taken together, the DDM analysis shows that subjects' choice strategy (threshold) and task engagement (nondecision time) were stable across conditions and sessions and supports that our task separates expected performance from sensory reliability and boundary distance.

Isolating Neural Signatures of Expected Performance. Having validated our experimental approach, we next estimated a general

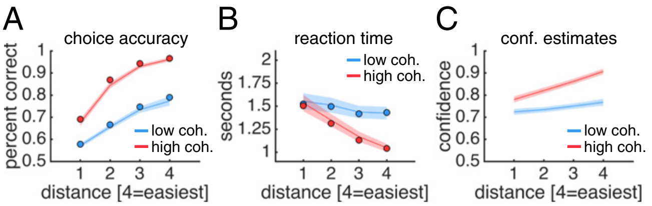

Fig. 2. Behavioral results. (*A*) Choice accuracy. (*B*) Reaction time measured from reference onset. (*C*) Confidence estimates. In *A* and *B*, the solid dots represent posterior predictive values from a hierarchical DDM fit to subjects' responses separately for each condition. In *A*–*C*, data are from the prescan session. *SI Appendix*, Fig. S3 provides equivalent plots from the scan session. Data are presented as group mean ± SEM.

linear model (GLM) of the fMRI data. As noted above, the probability that a subject's motion discrimination judgment is correct is a function of both the reliability of a subject's percept of motion direction (coherence) and the distance between a subject's motion percept and a choice boundary (distance) (Fig. 2*A*). Thus, we would expect a brain region involved in tracking expected performance, an aggregate signature of decision confidence, to carry the main effects of coherence and distance and, importantly, an interaction between these two factors.

To identify such activity patterns, we adopted a masking approach. At a whole-brain level, we first searched for main effects of coherence and distance and then, applying an inclusive mask constructed from the intersection of the two main effects (each map thresholded at $P < 0.05$, uncorrected), searched for an interaction between coherence and distance [$P < 0.05$, family-wise error rate (FWE)-corrected]. This analysis identified a single cluster in the medial prefrontal cortex, the pgACC. In this area, activity tracked changes in both coherence and distance and, importantly, an interaction between these two factors (Fig. 3 *A* and *B*), reflecting the pattern of both choice accuracy (Fig. 2*A*) and explicit confidence estimates (Fig. 2*C*).

We next identified areas that selectively tracked changes in coherence independent of distance. At a whole-brain level, we applied an exclusive mask constructed from the intersection of the main effect of distance and the coherence × distance interaction (each map thresholded at $P < 0.05$, uncorrected) and searched for a main effect of coherence ($P < 0.05$, FWE-corrected). This analysis identified clusters in the extrastriate cortex, posterior cingulate cortex, parietal cortex, and striatum, extending into the thalamus; in these areas, activity was higher when coherence was high but was unaffected by distance (Fig. 3 *A* and *B*). The extrastriate and parietal clusters encompassed area MT+ and the intraparietal sulcus, respectively, areas that are sensitive to motion coherence and motion direction (19, 20).

Finally, we identified areas that selectively tracked distance independent of coherence. At a whole-brain level, we applied an exclusive mask constructed from the intersection of the main effect of coherence and the coherence × distance interaction (each map thresholded at $P < 0.05$, uncorrected) and searched for a main effect of distance ($P < 0.05$, FWE-corrected). This analysis identified clusters in the posterior cingulate cortex, with higher activity when distance was high, and the presupplementary motor area (pre-SMA), with higher activity when distance was low (Fig. 3 *A* and *B*). We obtained comparable whole-brain effects when including correct trials only (*SI Appendix*, Fig. S5).

Controlling for Choice Reaction Time and Value. We next considered alternative explanations of our neural results in terms of choice reaction time and choice value. The effects of coherence, distance and the coherence × distance interaction on pgACC activity survived the inclusion of choice reaction time, both in a regression analysis of activity time courses (Fig. 3*C*) and in a series of control GLMs (*SI Appendix*, Fig. S5). We also took advantage of the fact that we varied the reward magnitude associated with the scoring rule on trials in which an explicit confidence estimate was required, such that a correct decision was three times more valuable in one-half of these confidence trials. While reward magnitude modulated activity time courses in the ventral striatum, in line with its role in encoding reward expectation (21), we did not observe an effect of reward magnitude in the pgACC (Fig. 3*D*). Taken together, these analyses indicate that the neural activations identified by our factorial design were not simply due to variation in choice reaction time and/or choice value.

Neural Basis of Decision Confidence. Having demonstrated that the pgACC tracks expected performance as specified by our factorial design, we sought to establish a role for the pgACC in the construction of a subjective sense of decision confidence. Normatively, decision confidence should reflect an internal estimate of the probability that a choice is correct and thus, if computed accurately, should track expected performance as assayed using our factorial design. Indeed, the above analyses show that the pgACC satisfies such a requirement for a neural signature of decision confidence. However, a direct correspondence between expected performance and decision confidence should not always be expected, especially in the absence of any feedback. At the single-trial level, subjective confidence is an internal state that can vary even when external variables are held constant, and at the aggregate level, subjects might not have assigned appropriate weights to the components of confidence formation. Thus, to establish that the pgACC is central to a subjective sense of decision confidence, it is important to show that the pgACC also tracks such "residual" variation in subjective confidence over and above the expected performance.

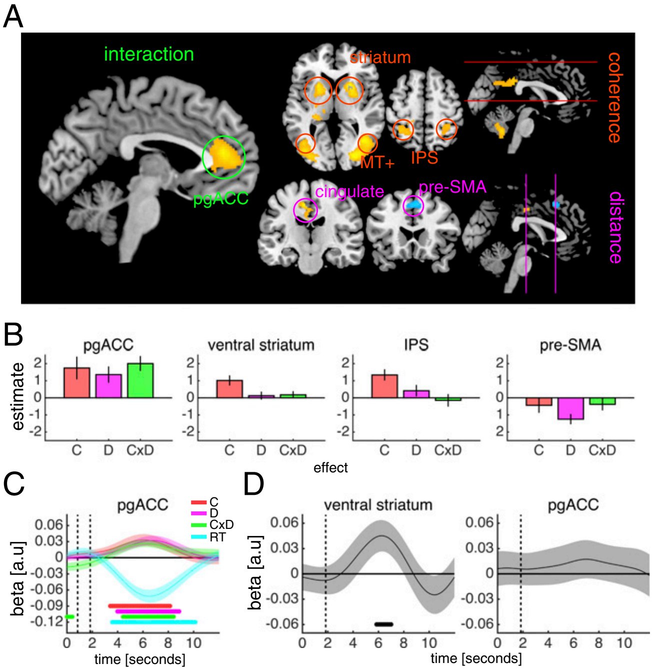

Fig. 3. Neural signatures of expected performance, sensory reliability, and boundary distance. (*A*) Whole-brain factorial analysis of the effects of coherence, distance, and the coherence × distance interaction. Activations are masked as detailed in the text. Cluster colors denote positive (warm) and negative (cold) effects. Clusters are significant at $P < 0.05$, FWE-corrected for multiple comparisons; the cluster-defining threshold is $P < 0.001$, uncorrected. Images are shown at $P < 0.001$, uncorrected. All clusters surviving whole-brain correction postmasking and premasking are detailed in *SI Appendix*, Tables S1 and S2. *SI Appendix*, Fig. S5 presents control GLMs. (*B*) ROI contrast estimates from factorial analysis of the effects of coherence (C), distance (D), and the coherence × distance interaction (C × D). (*C*) GLM analysis of the effects of coherence (C), distance (D), the coherence × distance interaction (C × D), and choice reaction time (RT) on ROI activity time courses. Vertical dashed lines indicate the onset of the motion stimulus and the choice phase. *SI Appendix*, Fig. S6 shows additional ROIs. (*D*) GLM analysis of the effect of reward magnitude on ROI activity time courses on confidence trials. The vertical dashed line indicates the onset of the reward magnitude cue. *SI Appendix*, Fig. S7 shows additional ROIs. In *B*–*D*, to avoid biasing subsequent analyses, ROIs were specified using simple contrasts from our factorial analysis (coherence, distance, and coherence × distance) before masking, except for the ventral striatum, which was specified anatomically. To avoid circularity, a leave-one-out cross-validation procedure was used for ROI specification. Data are represented as group mean ± SEM. In *C* and *D*, dots below the time course indicate significant excursions of *t* statistics assessed using two-tailed permutation tests.

We first sought to establish a trial-by-trial relationship between pgACC activity and subjective confidence. We fitted an ordinal regression model to each subject's explicit confidence estimates in the prescan session and used this model to generate out-of-sample predictions about their subjective confidence in the scan session (Fig. 4*A*). Supporting that the pgACC is central to a subjective sense of decision confidence, pgACC activity estimates [Fig. 4*B*, linear

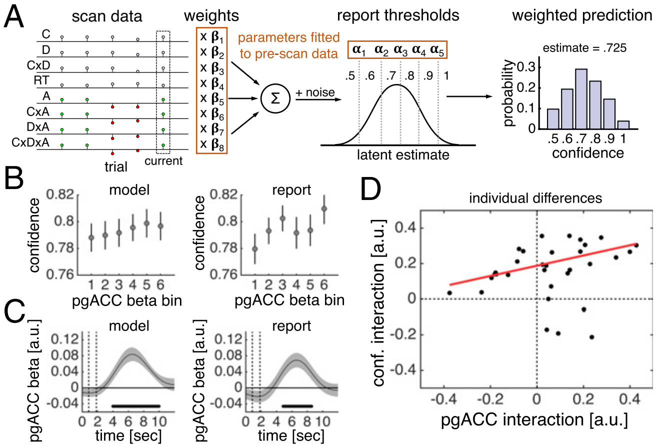

Fig. 4. Activity in the pgACC predicts decision confidence. (*A*) Model of subjective confidence. We fitted an ordinal regression model to each subject's confidence estimates in the prescan session. The model has a set of weights, which parameterize the effects of stimulus and choice features on confidence estimates, and a set of thresholds, which parameterize report biases. By applying the fitted model to each trial of a subject's scan data (stimulus and choice features), we generated a prediction about the subject's subjective confidence in that trial. The prediction is a probability distribution over possible responses (e.g., 0.5 has a 10% probability, 0.6 has a 20% probability, and so forth). We used the expectation over possible responses as our current estimate of subjective confidence. *SI Appendix*, Fig. S8 presents model evaluation. A, accuracy; C, coherence; D, distance; RT, reaction time. (*B*) Visualization of encoding of model-derived subjective confidence (all trials) and reported confidence (confidence trials) in single-trial pgACC activity estimates. (*C*) GLM analysis of encoding of model-derived subjective confidence (all trials) and reported confidence (confidence trials) in pgACC activity time courses. Dots below the time course indicate significant excursion of *t* statistics assessed using two-tailed permutation tests. In *B* and *C*, data are presented as group mean ± SEM. (*D*) Correlation between the interaction of coherence and distance in pgACC activity and confidence estimates. Interaction effects were calculated as a "difference of differences" in our 2 × 2 design. The pgACC interaction was calculated using all trials; the confidence interaction was calculated using confidence trials only. We used robust linear regression because in high-coherence trials, four outlying subjects were more confident when distance was low than when distance was high—a pattern inconsistent with normative predictions for decision confidence and the behavioral results shown in Fig. 2*C*.

regression; *t* (31) = 2.14, *P* = 0.041] and pgACC activity time courses (Fig. 4*C*) predicted trial-by-trial variation in this model-derived variable. Further, pgACC activity estimates [Fig. 4*B*, linear regression; *t* (31) = 1.26, *P* = 0.22] and pgACC activity time courses (Fig. 4*C*) also predicted trial-by-trial variation in the explicit confidence estimates elicited every 5–10 trials.

We next assessed the relationships between individual differences in the neural profile of pgACC activity and the behavioral profile of confidence estimates. There was substantial variation in the extent to which subjects' explicit confidence estimates reflected an interaction between coherence and distance (*SI Appendix*, Fig. S2). Notably, the degree to which coherence and distance interacted in subjects' pgACC activity predicted the degree to which the factors interacted in the subjects' explicit confidence estimates [Fig. 4*D*; robust linear regression; *t* (30) = 2.78, *P* = 0.009]. This relationship survived controlling for the corresponding effect for expected performance [robust linear regression; *t* (29) = 2.44, *P* = 0.021], further supporting that pgACC activity is key to the construction of a subjective sense of decision confidence.

Discussion

In studies of the neural basis of decision confidence, it has proven difficult to dissociate a neural representation of decision confidence from neural representations of its component parts. For example, in the context of the classic random dot-motion task, a neural area may predict opt-out responses because it tracks the reliability of a percept of motion direction—a key component of a confidence computation—rather than decision confidence itself, a quantity that often requires the integration of multiple components (1). Recognizing this distinction is key for advancing our understanding of the neurobiology of decision confidence.

We have developed a psychophysical approach to dissociate a neural representation of decision confidence from neural representations of its component parts. Extending the classic random dot-motion task, we varied both the percentage of coherently moving dots, which could move in any direction along the full circle, and the angular distance between the net direction of dot motion and a variable reference against which motion direction had to be judged. We showed that in this design, subjects' expected performance, an aggregate signature of decision confidence, is a function of both sensory reliability and boundary distance. Using our task to interrogate fMRI data, we observed that activity in an area in the medial prefrontal cortex (the pgACC) uniquely tracked expected performance, whereas neural areas previously proposed to track decision confidence tracked sensory reliability (posterior parietal cortex and ventral striatum) or boundary distance (pre-SMA). Supporting that the information encoded by the pgACC is central to a subjective sense of decision confidence, we found that pgACC activity predicted both within-subject and between-subject variation in explicit confidence estimates.

Intriguingly, the pgACC may be involved in the formation not only of a local estimate of the probability of correct choice on a single trial, the computational definition of decision confidence, but also of a more global estimate of the probability of correct choice across trials, an estimate critical for assessing one's general ability on a task. Evidence for the latter function comes from a recent study showing that pgACC activity tracks a running average of subjects' performance history and predicts their explicit evaluations of expected performance independent of the expected value of a trial, as in our study (22). Unlike our study, in which performance was designed to be independent from one trial to another, in that study performance was rigged such that it was autocorrelated across trials and could be learned only by tracking performance feedback. The diverse connectivity profile of the pgACC is consistent with this area being central to the formation of both local and global estimates of the probability of correct choice across different task domains. The pgACC is connected to the surrounding medial prefrontal cortex, dorsolateral and ventrolateral prefrontal cortex, cingulate cortex, subcortical areas such as the hippocampus and striatum, and posterior areas, including the parietal cortex (23). Taken together, these findings may help explain why metacognition, the ability to monitor and evaluate the success of one's task performance, is often impaired after prefrontal lesions and/or degeneration (24). If the pgACC is critical for confidence formation, then compromising the pgACC or connections to and/or from the pgACC should naturally lead to discrepancies between subjective evaluation and objective performance.

It is noteworthy that our factorial analysis did not identify the rostrolateral prefrontal cortex (rlPFC). Previous studies have consistently shown that rlPFC activity tracks explicit confidence estimates (11, 12), and that the microstructure of the rlPFC predicts the degree to which an individual's subjective evaluation reflects objective performance (25). One hypothesis, which may reconcile those results with ours, is that the rlPFC itself is not involved in computing an internal estimate of the probability that a choice is correct, but instead governs the mapping of this variable onto an explicit confidence estimate for report. In support of this hypothesis, we found that rlPFC activity time courses also predicted trial-by-trial variation in the model-derived predictor of subjective confidence and subjects' explicit confidence estimates (*SI Appendix*, Fig. S9). There is evidence that the rlPFC manages task sets and rules (26), functions presumably involved in maintaining a consistent confidence mapping or in updating confidence mapping in response to a particular communicative context (4). Future studies directly manipulating confidence mapping are needed to test this hypothesis.

Several studies have proposed that decision confidence is a function of the absolute distance between sensory evidence, x, and a choice boundary, b: $|x - b|$ (3, 6, 8, 9). This formulation makes the prediction that decision confidence is higher for larger distances in correct trials but lower for larger distances in error trials—a qualitative pattern that is mirrored in the activity of putative neural substrates of decision confidence, such as rat orbitofrontal cortex (3). Interestingly, we found that the pre-SMA, which tracked boundary distance in our task, also showed such an activity pattern (*SI Appendix*, Fig. S10). This area has been implicated in conflict monitoring (27), confidence formation (12), and changes of mind (28). A unifying explanation of these results may be that the pre-SMA encodes evidence in the coordinate frame of response options—a signal that can be used to guide subsequent behavior and cognition, such as increasing response caution (29) and assigning increased weight to postchoice evidence (28). However, in our task, decision confidence is a function not only of boundary distance, but also of sensory reliability—as expected under normative models of decision confidence and as shown by subjects' explicit confidence estimates (Fig. 2*C*).

Our study may prompt reconsideration of the contribution of the intraparietal sulcus (IPS) and ventral striatum to confidence-based behaviors. For example, in the context of the classic random dot-motion task, neurons in the IPS, including the lateral intraparietal area, which receives input from motion-sensitive area MT+ (30), have been shown to encode the accumulation of evidence toward a choice boundary, with the amplitude and the temporal profile of neuronal activity varying with motion coherence and choice reaction time (31). Because such activity patterns carry information about the probability that a choice is correct, the IPS has been proposed to be central to not only choice, but also confidence formation (5). A similar argument has been made for dopaminergic neurons in the substantia nigra connected with the ventral striatum (8). However, in the context of our continuous-direction, variable-reference random dot-motion task, we observed that fMRI activity in a putative human homolog of the IPS and ventral striatum tracked sensory reliability, but not the integration of sensory reliability with boundary distance. These results are in line with a recent study reporting that the superior colliculus, which together with the IPS forms part of the oculomotor planning circuit, tracks choice but not confidence formation (32).

We remain agnostic as to the source of these reliability-related signals in our task; however, our results are consistent with a hypothesis that such areas encode a probability distribution over sensory states given sensory evidence (33). In this regard, the amplitude of neural activity in these areas may reflect the reliability associated with this distribution. First, activity in the IPS and ventral striatum cannot be explained by a bottom-up response to coherent motion, but instead appears to track changes in motion coherence. In a separate motion-localizer scan, we found that the extrastriate cortex, including MT+, but notably not the IPS or ventral striatum, was activated in a contrast between coherently moving and static dots (*SI Appendix*, Fig. S11). Second, activity in these areas is selective for changes in motion coherence, as evidenced by our masking approach (Fig. 3*A*). Finally, subjects' percepts of motion direction were more reliable in high-coherence trials (Fig. 2*A*), and estimates of sensory reliability were clearly used to inform confidence estimates (Fig. 2*C*). An alternative interpretation is that activity in the IPS and/or ventral striatum reflects expected reward across possible choices (33), a quantity that would be larger in high-coherence trials. It remains to be seen how activity in these areas varies with the space of sensory states (e.g., binary or continuous direction) and the onset of the choice boundary (e.g., concurrent with or after stimulus).

Why should the brain maintain both "choice-dependent" and "choice-independent" estimates of certainty? In many tasks, to perform optimal inference, it is useful to represent the certainty associated with relevant sensory or cognitive variables independent of any future choice (1, 34). For example, when inferring the length of an object from visual and haptic information, the brain needs to know which source is more reliable and thus should dominate the integrated visual-haptic percept, regardless of which response may eventually be required (35). However, after making a choice, it may be efficient to combine these estimates into a statistic summarizing the probability that a choice is correct (1), which, for example, can be used to guide multistage decisions, control learning from feedback (36), and optimize group decisions (4). Our study supports a distinction between choice-dependent and choice-independent estimates of certainty (1) and indicates a dissociation in the neural encoding of these quantities.

Methods

Subject Details. Thirty-five adults (17 females; mean age, 23.60 ± 4.31 y) with normal or corrected-to-normal vision participated in the study. The subjects performed separate prescan and scan sessions, with 2–14 d between sessions. Three subjects were excluded due to excessive motion and/or sleeping during the scan session; their data were not included in any analyses. All subjects provided informed consent, and the study was approved by the Ethics Committee of University College London. Each subject received a flat rate for participation (£40) and could earn an additional performance-based bonus (up to £12).

Experimental Details.

Task. Subjects performed a continuous-direction, variable-reference random dot-motion task as described above (Fig. 1). Choices and confidence estimates were submitted to a variant of the Brier score, with a subject's score (reward or loss) on trial t calculated as follows: $score_t = r \times (0.5 - (confidence_t - accuracy_t)^2)$, where *confidence* is {0.5, 0.6, 0.7, 0.8, 0.9, 1}, *accuracy* is {0, 1}, and r is a scaling factor specifying the maximum reward or cost. In the prescan session, the reward factor was fixed at £4. In the scan session, the reward factor was £2 in one half of the confidence trials and £6 in the other half (indicated by "£" or "££", respectively, displayed above the scale). Subjects received the sum of their average trial-by-trial earnings calculated separately for each reward factor. Details of trial events and timing, response mode, and stimulus presentation are provided in *SI Appendix*, *Methods*.

Prescan procedure. The prescan session consisted of five runs. Subjects first viewed motion stimuli of variable duration and coherence. Subjects then practiced the task (40 trials) with high coherence and high distances. In this run only, subjects received trial-by-trial feedback, with the aim of familiarizing them with making direction judgments in continuous space. Subjects then performed calibration phase 1 (120 trials) and phase 2 (260 trials), involving estimation of a set of coherences and distances (2 × 4 design) associated with target levels of choice accuracy. Finally, subjects performed the main experiment (540 trials). The calibration procedure is described in detail in *SI Appendix*, *Methods*.

Scan procedure. The scan session comprised seven runs. Subjects first performed a calibration phase during the acquisition of structural images (180 trials), then performed the main experiment over five runs (5 × 112 = 560 trials). We used a subset of the calibrated task parameters (2 × 2 design) from the prescan session. In the final scan run, subjects viewed alternating displays (12 s) of static and dynamic dots (2 × 12 = 24 displays). Details are provided in *SI Appendix*, *Methods*.

Behavioral Analysis.

Choice reaction time. We excluded trials in which subjects' choice reaction times were 2.5 SD below or above their grand mean reaction time computed separately for the prescan and scan sessions. This procedure resulted in the exclusion of approximately 2% of the trials per subject per session.

Regression models. We used logistic regression to predict choice accuracy, linear regression to predict choice reaction time, and ordinal regression to predict confidence estimates (Fig. 2). We log-transformed choice reaction time, contrast-coded coherence and distance, and then z-scored all variables. We performed a separate regression for each subject and tested the group-level significance of a predictor by comparing the coefficients pooled across subjects to 0 (one-sample t test). We used ordinal regression to construct the computational model of subjective confidence for fMRI analysis (Fig. 4*A*). The model included eight predictors: log-transformed choice reaction time, accuracy, coherence, distance, and interaction terms for coherence × distance, accuracy × coherence, accuracy × distance and accuracy × coherence × distance. We did not z-score variables to facilitate between-session compatibility. The model provided a good fit to the prescan data and generalized well to the scan data (*SI Appendix*, Fig. S8).

Hierarchical drift-diffusion model. We estimated subjects' DDM parameters using hierarchical Bayesian estimation within the HDDM toolbox (17) (ski.clps.brown.edu/

hddm_docs/). We fitted drift rate (*v*), decision threshold (*a*), and nondecision time (*t*) separately for each condition of our factorial design (prescan: eight conditions; scan: four conditions). We also included intertrial variability in drift rate (*sv*) and nondecision time (*st*) as free parameters that were constant across conditions. We extracted mean group-level posterior estimates for visualization of DDM parameters (*SI Appendix*, Fig. S4) and entered these into the *simuldiff* function from the DMAT toolbox (37) (https://ppw.kuleuven.be/okp/software/dmat/) to generate posterior predictive values (Fig. 2 *A* and *B*).

fMRI Analysis.

Whole brain. fMRI analysis was conducted using SPM 12 (www.fil.ion.ucl.ac.uk/spm/). The whole-brain analysis shown in Fig. 3*A* was based on a single event-related GLM (GLM1). We labeled each trial according to whether coherence (C) and distance (D) were low (L) or high (H): CL&DL, CL&DH, CH&DL, and CH&DH. We modeled the four trial types with separate "condition" regressors. Choice reaction time outlier trials were assigned to an "outlier" regressor. These regressors were specified as boxcars time-locked to the onset of the motion stimulus and spanning until the time of choice (38). We modeled confidence events with a separate "rate" regressor, specified as a stick function time-locked to the onset of the scale and parametrically modulated by the (*z*-scored) number of button presses used to navigate the marker. We included motion and biophysical parameters as additional "nuisance" regressors. Regressors were convolved with a canonical hemodynamic response function. Regressors were modeled separately for each scan run, and constants were included to account for between-run differences in mean activation and scanner drifts. A high-pass filter (128-s cutoff) was applied to remove low-frequency drifts.

We assessed group-level significance by applying one-sample *t* tests against 0 to the first-level contrast images. Inclusive/exclusive masks for our factorial analysis were created using SPM's *imcalc* function; each second-level contrast image was thresholded at $P < 0.05$, uncorrected, the default setting for masking analyses in SPM, before creating a mask. We report clusters significant at $P < 0.05$, FWE-corrected for multiple comparisons, with a cluster-defining threshold of $P < 0.001$, uncorrected.

We obtained similar results in GLMs that introduced modifications to GLM1 (*SI Appendix*, Fig. S5). In GLM1C, only correct trials were included. In GLM2, the four condition regressors had a fixed duration of 1 s and were parametrically modulated by (log-transformed) choice reaction time. In GLM2C, only correct trials were included. In GLM3, there was only one "interest" regressor, which had a fixed duration of 1 s and was parametrically modulated by four dummy variables, one for each trial type, and (log-transformed) choice reaction time. In GLM3C, only correct trials were included. Parametric modulators were not orthogonalized, and regressors competed to explain variance.

Clusters surviving whole-brain correction postmasking and premasking for GLM1 and GLM1C are shown in *SI Appendix*, Tables S1–S4. Details on fMRI acquisition, preprocessing, and physiological monitoring are provided in *SI Appendix*, *Methods*.

ROIs. Detailed information on ROI specification, ROI single-trial activity estimates, ROI activity time courses, and permutation testing is provided in *SI Appendix*, *Methods*.

Resource Sharing. Anonymized behavioral data and code supporting our main analyses are available on GitHub (https://github.com/metacoglab/BangFleming/). Unthresholded group-level statistical maps are available on NeuroVault (https://neurovault.org/collections/3792/).

ACKNOWLEDGMENTS. We thank Chris Summerfield for helpful comments on an earlier version of this manuscript, and Dick Passingham for advice on neuroanatomy. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

1. Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat Neurosci* 19:366–374.
2. Meyniel F, Sigman M, Mainen ZF (2015) Confidence as Bayesian probability: From neural origins to behavior. *Neuron* 88:78–92.
3. Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455:227–231.
4. Bang D, et al. (2017) Confidence matching in group decision-making. *Nat Hum Behav* 1:0117.
5. Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324:759–764.
6. Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A (2013) Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat Neurosci* 16: 749–755.
7. Middlebrooks PG, Sommer MA (2012) Neuronal correlates of metacognition in primate frontal cortex. *Neuron* 75:517–530.
8. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A (2017) Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Curr Biol* 27: 821–832.
9. Hebart MN, Schriever Y, Donner TH, Haynes J-D (2016) The relationship between perceptual decision variables and confidence in the human brain. *Cereb Cortex* 26: 118–130.
10. Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18:1159–1167.
11. De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16:105–110.
12. Fleming SM, Huijgen J, Dolan RJ (2012) Prefrontal contributions to metacognition in perceptual decision making. *J Neurosci* 32:6117–6125.
13. Ferrera VP, Yanike M, Cassanello C (2009) Frontal eye field neurons signal changes in decision criteria. *Nat Neurosci* 12:1458–1462.
14. van den Berg R, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM (2016) Confidence is the bridge between multi-stage decisions. *Curr Biol* 26:3157–3168.
15. Boldt A, de Gardelle V, Yeung N (2017) The impact of evidence reliability on sensitivity and bias in decision confidence. *J Exp Psychol Hum Percept Perform* 43: 1520–1531.
16. Zylberberg A, Roelfsema PR, Sigman M (2014) Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious Cogn* 27:246–253.
17. Wiecki TV, Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Front Neuroinform* 7:14.
18. Shadlen MN, Shohamy D (2016) Decision making and sequential sampling from memory. *Neuron* 90:927–939.
19. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J Neurosci* 12: 4745–4765.
20. Braddick OJ, et al. (2001) Brain areas sensitive to coherent visual motion. *Perception* 30:61–72.
21. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
22. Wittmann MK, et al. (2016) Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* 91:482–493.
23. Neubert F-X, Mars RB, Sallet J, Rushworth MFS (2015) Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proc Natl Acad Sci USA* 112:E2695–E2704.
24. Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* 137:2811–2822.
25. Allen M, et al. (2017) Metacognitive ability correlates with hippocampal and prefrontal microstructure. *Neuroimage* 149:415–423.
26. Gilbert SJ, et al. (2006) Functional specialization within rostral prefrontal cortex (area 10): A meta-analysis. *J Cogn Neurosci* 18:932–948.
27. Kouneiher F, Charron S, Koechlin E (2009) Motivation and cognitive control in the human prefrontal cortex. *Nat Neurosci* 12:939–945.
28. Fleming SM, van der Putten EJ, Daw ND (2018) Neural mediators of changes of mind about perceptual decisions. *Nat Neurosci* 21:617–624.
29. Forstmann BU, et al. (2008) Striatum and pre-SMA facilitate decision-making under time pressure. *Proc Natl Acad Sci USA* 105:17538–17542.
30. Maunsell JH, van Essen DC (1983) The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J Neurosci* 3:2563–2586.
31. Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J Neurosci* 22:9475–9489.
32. Odegaard B, et al. (2018) Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc Natl Acad Sci USA* 115:E1588–E1597.
33. Beck JM, et al. (2008) Probabilistic population codes for Bayesian decision making. *Neuron* 60:1142–1152.
34. Ma WJ, Jazayeri M (2014) Neural coding of uncertainty and probability. *Annu Rev Neurosci* 37:205–220.
35. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–433.
36. Purcell BA, Kiani R (2016) Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc Natl Acad Sci USA* 113: E4531–E4540.
37. Vandekerckhove J, Tuerlinckx F (2008) Diffusion model analysis with MATLAB: A DMAT primer. *Behav Res Methods* 40:61–72.
38. Grinband J, Wager TD, Lindquist M, Ferrera VP, Hirsch J (2008) Detection of time-varying signals in event-related fMRI designs. *Neuroimage* 43:509–520.