TITLE

Dissociating decision confidence from evidence reliability in the human brain

AUTHORS

Dan Bang[1,*] & Stephen M. Fleming[1,2]

AFFILIATIONS

[1]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom.

[2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London WC1N 3BG, United Kingdom.

[*]Correspondence: danbang.db@gmail.com

ABSTRACT

Our confidence in a choice and the reliability of evidence pertaining to that choice appear to be inseparable quantities. However, recent computational frameworks suggest that the brain should encode distinct estimates of these quantities for adaptive behavioural control. Consider a line judge in a tennis match deciding whether a ball is out. When the ball is moving quickly, the sensory signal is unreliable. Nevertheless, if the ball lands far from the line, a decision may be easy and confidence high. Decision confidence should reflect the probability that a choice is correct and must therefore, unlike other estimates of certainty such as evidence reliability, be flexibly conditioned on a choice. Here we devised a psychophysical approach to decouple confidence in a sensory choice from the reliability of the sensory evidence upon which a choice is based. Using human fMRI, we observed that an area in medial prefrontal cortex (perigenual cingulate cortex, pgACC) tracked probability correct, the normative definition of decision confidence, whereas neural areas previously thought to track decision confidence, such as ventral striatum and posterior parietal cortex, instead tracked sensory reliability. We show that pgACC activity does not simply co-vary with objective performance but is also linked to within-subject and between-subject variation in a subjective sense of decision confidence. Our study is consistent with the recent proposal that the brain separately maintains choice-dependent and choice-independent estimates of certainty, and sheds light on why dysfunctional confidence estimates often emerge following prefrontal lesions and/or degeneration.

KEYWORDS

confidence; certainty; perception; decision-making; fMRI; medial prefrontal cortex

INTRODUCTION

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 right choice which can be used to guide behaviour (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 alone (4, 5). Despite widespread agreement that decision confidence is a useful quantity for adaptive control of behaviour, neurobiological support for a distinct encoding of decision confidence is lacking.

Several computational models propose that decision confidence reflects an estimate of the probability that a choice is correct (1, 2). A ubiquitous paradigm for studying the neural basis of decision confidence is sensory psychophysics (6). On a typical trial, subjects first make a 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 clockwise or counterclockwise of vertical. Subjects then make a secondary judgement which, to maximise reward, 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. Trial-by-trial variation in these confidence-based behaviours, induced by experimental manipulation of stimulus ambiguity and/or stochasticity in stimulus encoding, are then used to make inferences about the neural basis of decision confidence.

A problem is, however, that such designs induce a correlation between the reliability of the evidence upon which a choice is based and the probability that a choice is correct. In the examples above, the higher the reliability of the sensory evidence (e.g., motion coherence or contrast), the higher the probability that a choice based on such evidence is correct. It is therefore hard to tell whether a putative neural signature of decision confidence is tracking evidence reliability or probability correct. Yet the distinction between these quantities matters: in Bayesian frameworks, estimates of evidence reliability and probability correct should be separately maintained for adaptive behavioural control (1). Consider a line judge in a tennis match deciding whether a ball is out. When the ball is moving quickly, the sensory signal is unreliable. Even then, a decision may be easy and confidence high (e.g., ball lands far from line). When the ball is moving slowly, the sensory signal is reliable. Still, a decision may be hard and confidence low (e.g., ball touches line). The point is that an estimate of probability correct, unlike an estimate of evidence reliability, must be flexibly conditioned on a choice.

A conflation of probability correct with quantities which are relevant to, but do not depend on, a choice but may explain why putative neural signatures of decision confidence have been reported throughout the brain. In monkeys, single-unit recording and functional inactivation have identified neural areas involved in confidence-based behaviours; these areas include lateral intraparietal sulcus (7), thalamic pulvinar (8), and dopaminergic neurons in substantia nigra (9). In rodents, similar approaches have identified orbitofrontal cortex as central to confidence-based behaviours (3, 10). In humans, fMRI has identified areas which track explicit confidence estimates, including ventromedial prefrontal cortex (11, 12), striatum (13), dorsal anterior cingulate cortex (14, 15) and rostrolateral prefrontal cortex (11, 14–16). However, because of the aforementioned conflation of the probability that a choice is correct with the reliability of the evidence upon which a choice is based, the computational role of these neural areas remains unclear.

Here we set out to overcome these issues and test for a distinct neural encoding of decision confidence. We extend a classic random-dot motion task (17, 18) to separate the probability that a subject's motion discrimination judgement is correct from the reliability of a subject's percept of motion direction. Using human fMRI, we show that the former quantity, probability correct, is uniquely tracked by medial prefrontal cortex (perigenual anterior cingulate cortex, pgACC), whereas sensory reliability co-varied with activity in extrastriate and parietal regions and striatum. Supporting that pgACC is central to a subjective sense of decision confidence, we show that its activity predicts both within-subject and between-subject variation in explicit confidence estimates. Our results show that decision confidence is separable from evidence reliability, and reveal a novel dissociation in the neural encoding of these quantities.

RESULTS

Experimental dissociation of probability correct and sensory reliability

Subjects (N = 32) viewed a field of moving dots inside a circular aperture (**Fig. 1**). A fraction of dots moved coherently in a specified direction, sampled anew on each trial from the range 1-360 degrees, whereas the rest moved randomly. Once the motion display had terminated, subjects were presented with a reference line transecting the aperture. They had to decide whether the net direction of dot motion was clockwise (CW) or counterclockwise (CCW) to the reference. Finally, after having made a choice, subjects were asked to indicate the confidence in their choice being correct. We varied the fraction of coherently moving dots and the absolute angular distance between the motion direction and the direction reference. In this design, unlike the classic random dot-motion task, the probability that a subject's motion discrimination judgement is correct is determined by *both* the reliability of a subject's percept of motion direction (coherence) and the placement of a subject's choice criterion (distance). We leveraged this feature of our design to identify putative neural signatures of probability correct over and above simple main effects of sensory reliability.

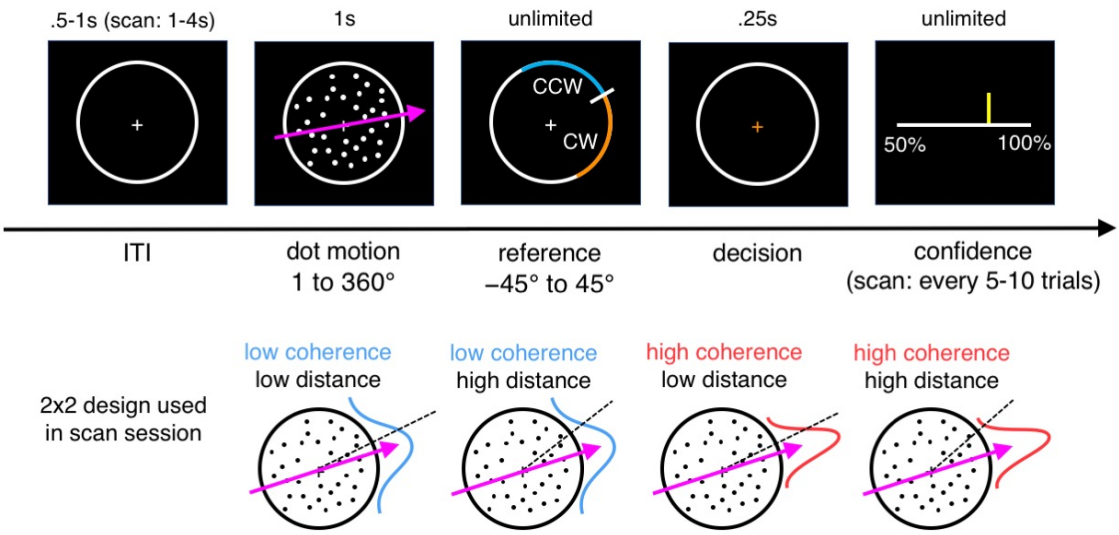

Figure 1. Random-dot motion task with varying reference. Subjects had to judge whether the net direction of dot motion was clockwise or counterclockwise to a reference which appeared after the motion stimulus. We varied the percentage of coherently moving dots and the absolute angular distance between the motion direction and the direction reference. Subjects performed the task in separate pre-

scan and scan sessions. In the pre-scan session, confidence estimates (50% to 100% in steps of 10%) were elicited on every trial. In the scan session, confidence estimates were elicited every 5-10 trials. See **Fig. S1** for information about stimulus calibration.

We used an incentive-compatible scoring rule (19) under which it was in subjects' best interest to make as many correct decisions as possible and to estimate their confidence as accurately as possible. Subjects were told about their earnings at the end of the experiment; we did not provide trial-by-trial feedback to prevent habitual learning of stimulus-response associations (20, 21). In a pre-scan session, we calibrated a set of coherences and distances (2x4 design) associated with target levels of choice accuracy, and evaluated the behavioural 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 (2x2 design in **Fig. 1**). To avoid confounding neural responses related to decision confidence with neural responses related to explicit reports, we elicited confidence estimates every 5-10 trials.

Behavioural validation of experimental approach

In our task, the probability that a subject makes a correct choice should be determined by both the reliability of their percept of motion direction, controlled by changes in coherence, and the placement of their choice criterion, controlled by changes in distance. We found that choice accuracy was indeed affected by both factors: subjects were more likely to be correct when coherence was high and when distance was high (**Fig. 2a**, logistic regression, coherence: $t(31) = 11.757$, $p < .001$, distance: $t(31) = 18.741$, $p < .001$, interaction: $t(31) = 10.440$, $p < .001$). These effects were mirrored in choice reaction time: subjects made faster decisions when coherence was high and when distance was high (**Fig. 2b**, linear regression, coherence: $t(31) = -4.810$, $p < .001$, distance: $t(31) = -10.995$, $p < .001$, interaction: $t(31) = -9.587$, $p < .001$). Factors affecting subjects' choice behaviour also predicted their explicit confidence estimates: subjects reported higher confidence when coherence was high and when distance was high (**Fig. 2c**, ordinal regression, coherence: $t(31) = 8.448$, $p < .001$; distance: $t(31) = 11.148$, $p < .001$; interaction: $t(31) = 9.390$, $p < .001$). These effects on confidence estimates survived controlling for choice reaction time and nuisance factors such as the initial position of the confidence marker and the cardinality of motion direction (**Fig. S2**). Finally, the effects of coherence and distance on choice behaviour and confidence estimates were replicated in the scan session (**Fig. S3**).

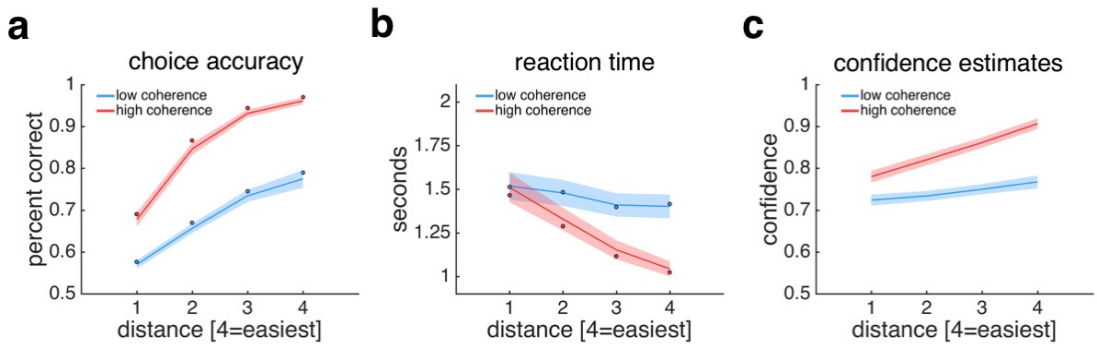

Figure 2. Behavioural results. **a**, Choice accuracy. **b**, Reaction time. **c**, Confidence estimates. See **Fig. S2** for analysis controlling for various nuisance factors. **a-b**, Solid dots are posterior predictives from the hierarchical drift-diffusion model fit to subjects' responses separately for each condition. **a-c**, Data are from the pre-scan session. See **Fig. S3** for equivalent plots from the scan session. Data are represented as group mean ± SEM.

To further unpack the drivers of the choice process, we modelled subjects' responses using a hierarchical instantiation of the drift-diffusion model (27). We remained agnostic as to how our task affected model parameters by fitting drift-rate (i.e. signal-to-noise ratio of evidence accumulation), threshold (i.e. amount of evidence needed for a choice) and non-decision 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 non-decision time were stable (see model predictions in **Fig. 2a-b** and parameter estimates in **Fig. S4**). In our task, the momentary evidence entering into the accumulation process can be thought of as the difference between a noisy sample from a sensory representation of motion direction held in visual short-term memory and a fixed choice criterion (22). Under this account, the reliability of the sensory representation, controlled by coherence, and the placement of the choice criterion, controlled by distance, *jointly* determine the signal-to-noise ratio of evidence accumulation and thereby the probability that a choice is correct. This analysis confirms that our task separates probability correct from lower-level effects of sensory reliability, and show that subjects' choice strategy (threshold) and task engagement (non-decision time) remained stable across conditions and sessions.

Isolating neural signatures of probability correct

Having shown separate behavioural control over probability correct and sensory reliability, we next estimated a general linear model of the fMRI data (2x2 design as shown in **Fig. 1**) to identify their neural signatures. As demonstrated above, the probability that a subject's motion discrimination judgement is correct is a function of both the reliability of a subject's percept of motion direction (coherence) and the placement of a subject's choice criterion (distance) (**Fig. 2a**). A brain region involved in tracking probability correct should therefore show 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 $p < .05$, uncorrected), searched for an interaction between coherence and distance ($p < .05$, FWE-corrected). This analysis identified a single cluster in medial prefrontal cortex (perigenual anterior cingulate cortex, pgACC) which satisfied the requirements for a neural signature of probability correct: in this area, consistent with the effects observed for choice accuracy (**Fig. 2a**) and confidence estimates (**Fig. 2c**), activity tracked changes in coherence and distance and, importantly, an interaction between these two factors (**Fig. 3a-b**).

We next identified neural areas which selectively track changes in coherence, independently 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 x distance interaction (each map thresholded at $p < .05$, uncorrected) and searched for a main effect of coherence ($p < .05$, FWE-corrected). This analysis identified clusters in extrastriate cortex, posterior cingulate cortex, posterior parietal cortex and striatum, extending into the thalamus: in these areas, activity was higher when coherence was high but was unaffected by distance (**Fig. 3a-b**). In keeping with previous studies, the extrastriate and the parietal clusters encompassed area MT+ and the intraparietal sulcus, respectively; both areas have been shown to be sensitive to motion coherence (23, 24) and motion direction (25, 26). In the Discussion, we consider whether these activations may constitute a neural representation of sensory reliability, a key input to a confidence computation in our task.

Finally, we identified neural areas which selectively track distance, independently 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 x distance interaction (each map thresholded at $p < .05$, uncorrected) and searched for a main effect of distance ($p < .05$, FWE-corrected). This analysis identified clusters in posterior cingulate cortex, with higher activity when distance was high, and pre-supplementary motor area, with higher activity when distance was low (**Fig. 3a-b**). The latter area in particular has been implicated in tracking choice difficulty (27) and confidence formation (15). However, we are reluctant to interpret these findings further as our task was not optimised for investigating the effects of criterion difficulty; the primary purpose of the variable reference was to separate probability correct from the effects of sensory reliability in both behaviour and brain activity. We emphasise that all whole-brain effects were comparable when we performed our factorial analysis using correct trials only (**Fig. S5**).

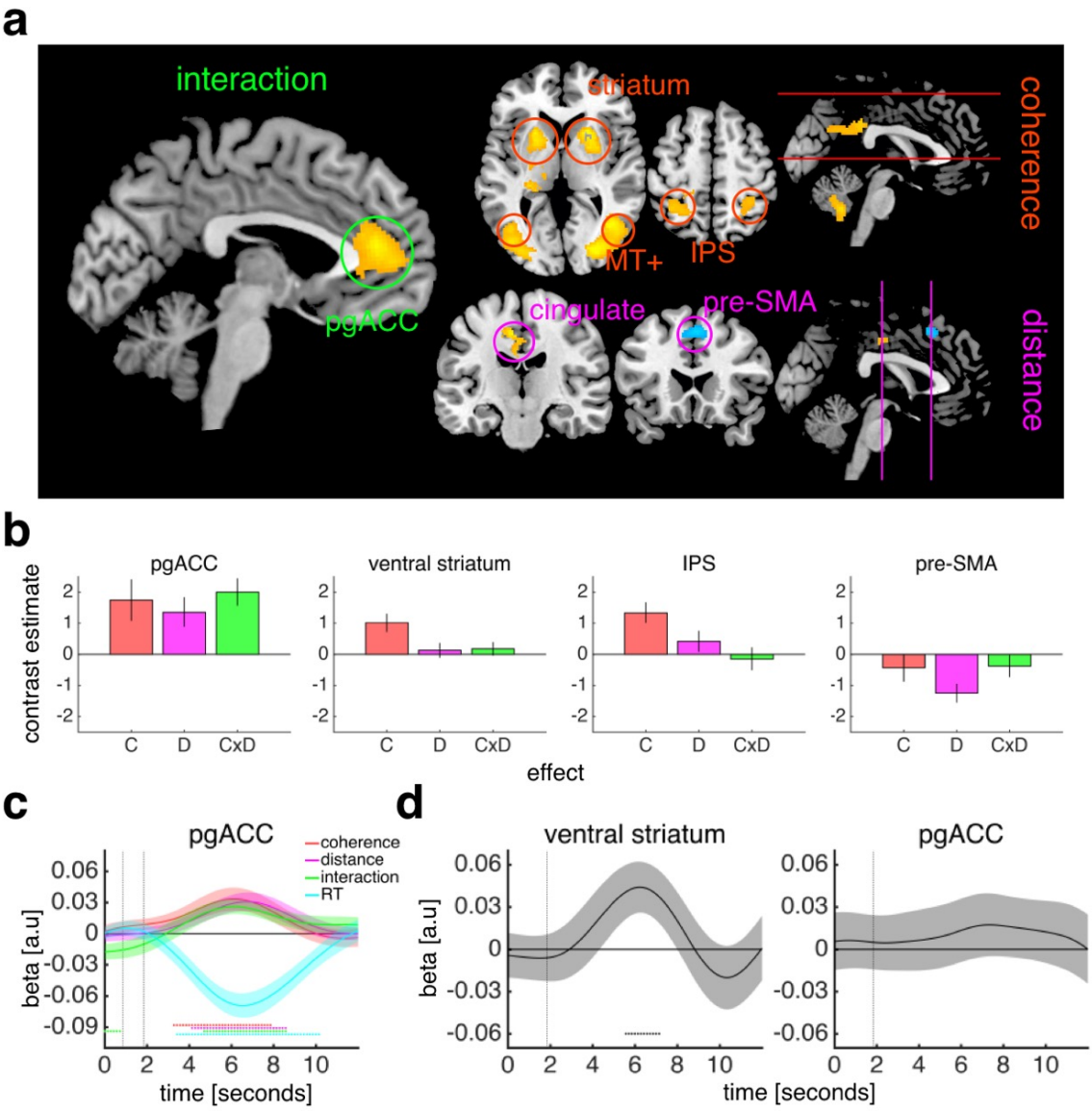

Figure 3. Neural signatures of probability correct and sensory reliability. **a**, Whole-brain factorial analysis of the effects of coherence, distance and the coherence x distance interaction. Activations are masked as detailed in the main text. Cluster colours denote positive (warm) and negative (cold) effects. Clusters are significant at $p < .05$, FWE-corrected for multiple comparisons. Cluster-defining threshold:

$p < .001$, uncorrected. Images are shown at $p < .001$, uncorrected. All clusters surviving whole-brain correction post- and pre-masking are detailed in **Tables S1-2**. See **Fig. S5** for control GLMs. **b,** ROI contrast estimates from factorial analysis of the effects of coherence (C), distance (D) and the coherence x distance interaction (CxD). **c**, General linear model analysis of the effects of coherence, distance, the coherence x distance interaction and choice reaction time on ROI activity timecourses. Vertical dashed lines indicate the onsets of the motion stimulus and the choice phase. See **Fig. S6** for additional ROIs. **d**, General linear model analysis of the effect of reward magnitude on ROI activity timecourses on confidence trials. Vertical dashed line indicates the onset of the reward magnitude cue. See **Fig. S7** for additional ROIs. **b-d**, To avoid biasing subsequent analyses, ROIs were specified using simple contrasts from our factorial analysis (coherence, distance, coherence x distance) prior to masking, except for ventral striatum, which was specified anatomically, as the striatal clusters identified by the factorial analysis extended into non-striatal areas. To avoid circularity, a leave-one-out cross-validation procedure was used for ROI specification. Data are represented as group mean ± SEM. **c-d**, Significance testing (dots) was performed using a permutation procedure.

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, features which often correlate with performance and/or confidence. The effects of coherence, distance and the coherence x distance interaction on pgACC activity survived the inclusion of choice reaction time, both in a regression analysis of activity timecourses (**Fig. 3c**) and a series of control GLMs (**Fig. S5**). We also took advantage of the fact that we varied the reward magnitude associated with the scoring rule on trials where an explicit confidence estimate was required, such that making a correct decision was three times more valuable on half of these confidence trials. While reward magnitude modulated activity timecourses in ventral striatum, consistent with an established role in encoding of upcoming reward (28, 29), we did not observe an effect of reward magnitude in pgACC (**Fig. 3d**). Taken together, these analyses indicate that the neural signatures of probability correct and sensory reliability were independent of choice response time and choice value.

Neural basis of individual differences in subjective confidence

Having demonstrated that pgACC tracks probability correct, we sought to establish a role for pgACC in the construction of a subjective sense of decision confidence. We first analysed the trial-by-trial relationship between pgACC activity and subjective confidence. We fitted an ordinal regression model to each subject's confidence estimates in the pre-scan session and then used the model to generate out-of-sample predictions about their subjective confidence in the scan session (**Fig. 4a**). Supporting a role of pgACC in the construction of a subjective sense of decision confidence, pgACC activity estimates (**Fig. 4b**, linear regression, $t(31) = 2.358$, $p = .025$) and pgACC activity timecourses (**Fig. 4c**) predicted trial-by-trial variation in the model-derived predictor. Further, pgACC activity estimates (**Fig. 4b**, linear regression, $t(31) = 1.612$, $p = .117$) and pgACC activity timecourses (**Fig. 4c**) were also predictive of trial-by-trial variation in the explicit confidence estimates given every 5-10 trials (15% of trials).

We next analysed the relationship between individual differences in the neural profile of pgACC activity and the behavioural profile of confidence estimates across subjects. On a normative view, to compute the probability that a choice is correct, subjects must estimate the reliability of a motion percept, controlled by changes in coherence, and use this information to weight the motion percept with respect to a choice criterion, controlled by changes in distance. There was, however, substantial individual differences in the degree to which the elicited confidence estimates reflected such an interaction (**Fig. S2**). Notably, the degree to which coherence and distance were integrated in pgACC activity predicted the degree to which the same factors were integrated in the explicit confidence estimates given every 5-10 trials (**Fig.**

4d; robust linear regression, $t(30) = 2.894$, $p = .007$), which was not the case for other ROIs. Critically, this relationship survived controlling for the effects of coherence and distance on objective performance (robust linear regression, $t(29) = 2.395$, $p = .023$), showing that pgACC activity is key to the construction of a subjective sense of decision confidence.

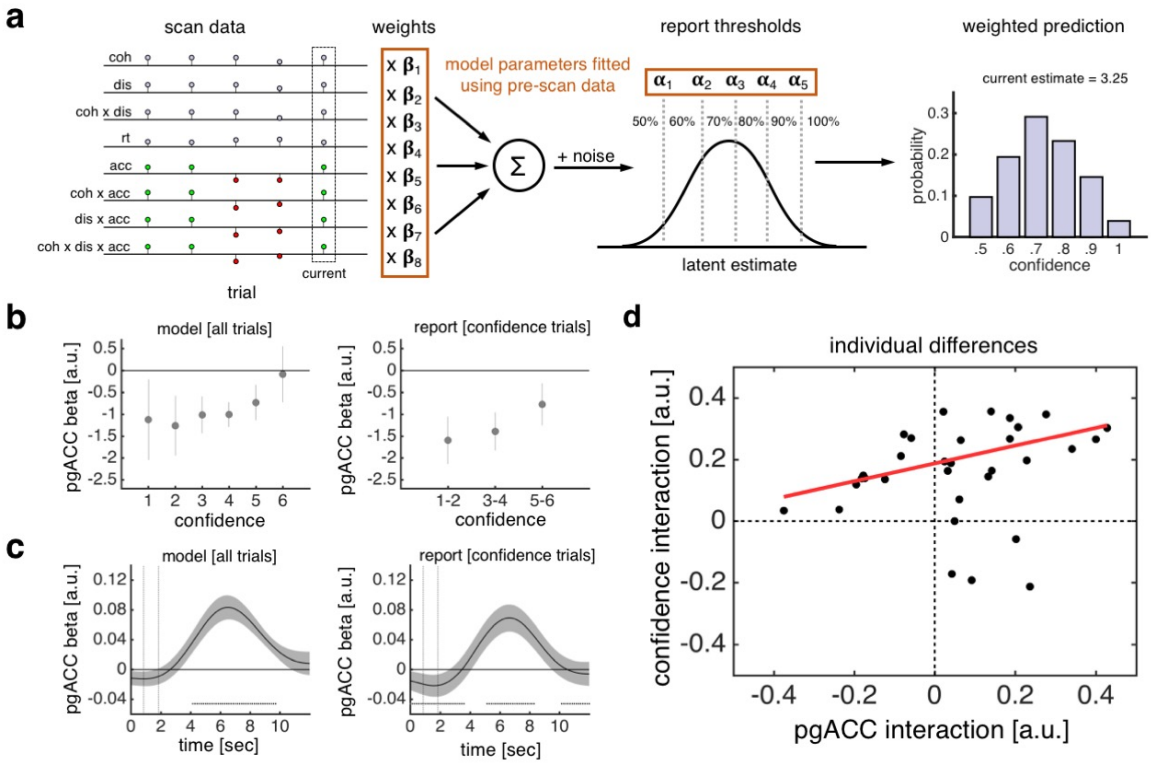

Figure 4. Activity in pgACC predicts decision confidence. **a**, Model of subjective confidence. We fitted an ordinal regression model to a subject's confidence estimates in the pre-scan session. The model has a set of weights, which parameterise the effects of stimulus and choice features on confidence estimates, and a set of thresholds, which parameterise 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 their subjective confidence on that trial. Each prediction takes the form of a probability distribution over possible responses (e.g., '.5' has a 10% probability, '.6' has a 20% probability, and so forth). We used the expectation of this probability distribution as our estimate of subjective confidence. See **Fig. S8** for model evaluation. **b**, Visualisation of encoding of model-derived subjective confidence (all trials) and reported confidence (confidence trials) in single-trial pgACC activity estimates. **c**, General linear model analysis of encoding of model-derived subjective confidence (all trials) and reported confidence (confidence trials) in pgACC activity timecourses. Significance testing (dots) was performed using a permutation procedure. **d**, Correlation between interaction of coherence and distance in pgACC activity and confidence estimates. Interaction effects were calculated as a 'difference of differences' in our 2x2 design. The pgACC interaction was calculated using all trials; the confidence interaction was calculated using confidence trials only. **b-c**, Data are represented as group mean ± SEM.

DISCUSSION

Several computational models propose that decision confidence reflects an estimate of the probability that a choice is correct (1, 2). Surprisingly, despite often being guided by these models, studies on the neural basis of decision confidence have, until now, not separated probability correct from the individual factors which bear on its computation. For example, in the classic random dot-motion task (17), the probability that a subject's motion discrimination judgement is correct is confounded with the reliability of a subject's percept of motion direction. Because of such issues with experimental design, a neural area which appears to track the

probability that a choice is correct may in fact track, or simply be sensitive to the effects of, the reliability of the information upon which a choice is based. Recognising this issue is key for advancing our understanding of the neurobiology of decision confidence.

Here we developed a novel psychophysical approach to dissociate a neural representation of decision confidence from the effects of sensory reliability. Extending the classic random dot-motion task, we systematically manipulated both the percentage of coherently moving dots and the angular distance between the net direction of dot motion and a variable reference against which motion direction had to be judged. In this design, the probability that a subject's motion discrimination judgement is correct is a function of both the reliability of a subject's percept of motion direction (coherence) and the placement of a subject's choice criterion (distance) and is thus not confounded with sensory reliability. Using this design to interrogate human fMRI data, we observed that activity in areas previously reported to track probability correct, such as ventral striatum (9, 13) and posterior parietal cortex (7), tracked sensory reliability. In contrast, activity in pgACC, located in medial prefrontal cortex, tracked probability correct and predicted both within-subject and between-subject variation in explicit confidence estimates.

Intriguingly, pgACC may be functionally involved in the formation of not only a local estimate of probability correct on a single trial, the computational definition of decision confidence, but also a global estimate of probability correct across trials, an estimate which is critical for assessing one's general ability on a task. Evidence for the latter function comes from a recent study which showed that pgACC activity tracks a running average of subjects' performance history and predicts subjects' explicit evaluations of recent performance (30). Unlike our study in which performance was designed to be independent from one trial to another, performance was rigged such that it was auto-correlated across trials and thus could only be learnt from trial-by-trial feedback. Importantly, the diverse connectivity profile of pgACC is consistent with this area as central to the formation of local and global estimates of probability correct: pgACC is connected to surrounding medial prefrontal cortex, dorsolateral and ventrolateral prefrontal cortex, cingulate cortex, subcortical areas, such as hippocampus and striatum, and posterior areas, including parietal cortex (31). These findings may help explain why metacognition, the ability to monitor and evaluate the success of one's performance, is often impaired after prefrontal lesions and/or degeneration (32–35). If pgACC is indeed critical for confidence formation, compromising pgACC, or connections to and/or from pgACC, should naturally lead to discrepancies between subjective evaluations and objective performance.

It is noteworthy that our factorial analysis did not identify rostrolateral prefrontal cortex (rlPFC). Previous studies have consistently shown that fMRI activity in rlPFC tracks explicit confidence estimates (11, 14–16), and that the microstructure of rlPFC can predict the degree to which an individual's subjective evaluations reflect objective performance (36–38). One hypothesis, which may reconcile these results with ours, is that rlPFC is not itself involved in computing a probability that a choice is correct, but instead governs the mapping of internal onto explicit confidence estimates. In support of this hypothesis, we found that activity timecourses in an rlPFC ROI also predicted trial-by-trial variation in our model-derived predictor of subjective confidence as well as subjects' explicit confidence estimates (**Fig. S9**). More broadly, there is evidence that rlPFC manages task sets and rules (39–43), functions which presumably are involved in maintaining a consistent mapping from internal to explicit confidence estimates, or updating a mapping in response to a particular communicative context (5). Future studies directly manipulating confidence mappings are required to test this hypothesis.

Our study may prompt a reconsideration of the putative neural correlates of decision confidence identified in previous studies. In the classic random dot-motion task, neurons in the intraparietal sulcus (IPS), including ventral parietal area (VIP) and lateral parietal area (LIP), both of which receive inputs from motion-sensitive area MT+ (44), have been shown to encode the accumulation of evidence towards a choice boundary, with the amplitude and the temporal profile of neuronal activity varying with motion coherence and choice reaction time (17, 45). Because such activity patterns carry information about the probability that a choice is correct, IPS has been proposed to be central to not only choice formation but also confidence formation (7). However, as noted above, the classic random dot-motion task confounds the probability that a subject's motion discrimination judgement is correct and the reliability of a subject's percept of motion direction. Here, after separating the two quantities, we observed that the BOLD signal in a putative human homologue of IPS tracked sensory reliability and not probability correct.

Another putative neural correlate of decision confidence is ventral striatum. Central to this claim is the signal detection assumption that decisions are made by comparing a continuous decision variable, d, randomly sampled from a Gaussian distribution centred on the stimulus, to a choice boundary, b, and that decision confidence reflects the absolute distance between the decision variable and the criterion, $|d–b|$. Human fMRI has shown that, while activity in frontal and posterior areas, including IPS, reflects the decision variable, activity in ventral striatum reflects the confidence variable (13). Further, monkey single-cell recordings have shown that the activity of dopaminergic neurons in substantia nigra pars compacta, which is preferentially connected with ventral striatum, reflects a similar signal (9). However, the tasks used in these studies also confound the probability that a choice is correct and the reliability (or the strength) of the evidence upon which a choice is based. Here, after separating the two quantities, we observed that a striatal cluster, encompassing ventral striatum, tracked sensory reliability and not probability correct.

We remain agnostic as to the contributions of IPS and striatum in our task. However, our results are consistent with a hypothesis that these areas encode the reliability of a subject's percept of motion direction, a role which has also been suggested for IPS in theoretical work (1, 46), and which is a key input to a confidence computation in our task. First, activity in these areas cannot be explained by a bottom-up response to coherent motion but instead appears to track *changes* in motion coherence: in a separate motion-localiser scan we observed that extrastriate cortex, including MT+, but notably not IPS or striatum, was activated in a contrast between coherently moving and static dots (**Fig. S10**). Second, activity in these areas is *selective* for changes in motion coherence in our task as evidenced by the masking approach (**Fig. 3a**). Finally, subjects' percepts of motion direction were more reliable on high-coherence trials (**Fig. 2a**), and estimates of such reliability were used to inform confidence estimates (**Fig. 2c**). An alternative interpretation of the striatal activity, in line with the role of ventral striatum in reward prediction-error learning (28, 29), is that the responses to motion coherence were in fact secondary to sensory reliability, with motion coherence being a salient predictor of reward (correct choices) in our task.

Why should the brain maintain both 'choice-independent' and 'choice-dependent' estimates of certainty? In many tasks, to perform optimal inference, it is useful to represent the certainty associated with relevant sensory or cognitive variables independently of any future choice (1, 20). For example, when required to infer the length of an object from visual and haptic cues, the brain needs to know which cue is more reliable and thus should dominate the integrated

visual-haptic percept, regardless of which response may eventually be required (47). Choice-independent estimates of certainty can also be useful when the reliability of different cues varies with time and must be updated independently (48, 49). However, after committing to a choice, it may be more efficient to combine these representations into a statistic that summarises the probability that a choice is correct (1). This summary statistic, decision confidence, can be used to guide multi-stage decisions (50), help optimise learning from feedback (51, 52) and be shared with others to make group decisions (4, 5). Our study supports a distinction between choice-independent and choice-dependent estimates of certainty (1), and indicate a dissociation in the neural encoding of these quantities in perceptual decision-making.

AUTHOR CONTRIBUTIONS

D.B. designed the experiment, programmed the task, collected the data, analysed the data, interpreted the results, and wrote the manuscript. S.M.F. designed the experiment, interpreted the results, and wrote the manuscript.

ACKNOWLEDGEMENTS

The Wellcome Centre for Human Neuroimaging is supported by core funding from Wellcome (203147/Z/16/Z).

COMPETING INTERESTS

The authors declare no competing interests.

METHODS

Experimental model and subject details

Subjects performed a random-dot motion task in separate pre-scan and scan sessions (2-14 days between sessions). Thirty-five adults (17 female, mean ± SD age = 23.60 ± 4.31 years) with no history of psychiatric or neurological disorders participated in the study. Three subjects were excluded due to excessive motion and/or sleeping in the scan session; their data was not included in any analyses. Subjects provided informed consent, and the study was approved by the ethics committee of University College London. Subjects received a flat rate for participation (£40) and earnt an additional performance-based bonus (up to £12).

Method details

Task

Random-dot motion task

Each trial began with the presentation of a fixation cross at the centre of a circular aperture. After a uniformly sampled delay (pre-scan: .5-1s; scan: 1-4s), subjects viewed a field of moving dots (1s). Subjects were instructed to fixate during stimulus presentation. Once the motion stimulus terminated, subjects were presented with a direction reference which transected the aperture. Subjects were required to press one of two buttons to indicate whether the net direction of dot motion (angle as measured from aperture centre) was clockwise (CW) or counter-clockwise (CCW) to the reference (angle as measured from aperture centre). As a visual aid, the response buttons and the associated arcs of the aperture were coloured orange and blue, with colour assignment counterbalanced across subjects. Once a choice had been made, the colour of the central cross (orange or blue) confirmed the decision (.25s). We varied the percentage of coherently moving dots (coherence) and the absolute angular distance between the net direction of dot motion and the direction reference (distance). We decoupled the motion stimulus and the direction reference in time as previous studies have shown that concurrent presentation of a reference may alter sensory encoding/decoding (53).

In the pre-scan session, subjects estimated the confidence in their choice (i.e., how likely they thought their decision was to be correct) on every trial. In the scan session, subjects estimated the confidence in their choice every 5-10 trials (15% of trials). Subjects responded by moving a marker along a scale from 50% to 100% in steps of 10%. The marker started randomly in one of the six locations along the scale and was controlled by button press. Once a response had been registered, the marker turned grey (.5s). Choices and confidence estimates were submitted to a variant of the Brier score (19), with a subject's score (reward or loss) on trial t given by: $score_{\mathrm{t}} = r * (.5 - (confidence_{\mathrm{t}} - accuracy_{\mathrm{t}})^2)$ where $confidence \in \{.5, .6, .7, .8, .9, 1\}$, $accuracy \in \{0, 1\}$, and r is a scaling factor which determines the maximum reward or cost. In the pre-scan session, the reward factor was fixed at £4. In the scan session, the reward factor was £2 on half of confidence trials and £6 on the other half; the size of the reward factor was indicated by a symbol above the confidence scale ('£' or '££'). Subjects received the sum of their average trial-by-trial earnings calculated separately for each reward factor (maximum bonus was £12). Subjects were informed about their earnings at the end of the experiment.

Stimulus construction

The circular aperture (8 degrees in diameter) was centred on a fixation cross (0.2 degrees in diameter). The motion stimulus was made up of three sets of dots (each dot was 0.12 degrees in diameter) shown in consecutive frames. Each set of dots was shown for one frame (about 16ms) and then replotted again three frames later (about 50ms) – some dots were displaced in the specified motion direction (see below) at a speed of 2 degrees s^{-1} while the rest of the dots were displaced at random locations within the aperture. We refer to the percentage of dots displayed in the specified motion direction as coherence, k. The dot density was fixed at 16 dots $degrees^{-2}$ s^{-1}. To help subjects maintain fixation, a circular region (0.7 degrees in diameter) at the centre of the circle was kept free of dots. The motion direction was sampled uniformly from the range 1-360 degrees. The direction of the reference (0.8 degrees in length and 0.08 degrees in width) was within 45 degrees of the motion direction. We here refer to the raw difference between the motion direction and the direction reference as signed distance,

δ, with positive values for CW and negative values for CCW. A set of coherences, $\mathbf{K}$, and a set of signed distances, Δ, were calibrated for each subject and crossed in a factorial design in the main experiment.

Stimulus calibration: pre-scan session

The aim of the calibration procedure was to identify a pair of coherences associated with different levels of sensory reliability and a set of signed distances associated with different levels of criterion difficulty. This aim was achieved in two phases. We did not elicit confidence estimates in either phase. In phase 1, subjects performed the task with a fixed absolute distance, $\mathbf{\Delta} \in$ {-10,10}. Coherence was .20 for the first two trials and then updated using a 2-down-1-up procedure: after two correct decisions, coherence was decreased by .01; after one incorrect decision, coherence was increased by .01. We used the median coherence of the last 20 trials of phase 1 to specify a medium coherence, k_{med}, constrained to a lower limit of .12 and an upper limit of .50. In phase 2, subjects performed the task with coherence fixed at k_{med}. We employed a range of signed distances, $\mathbf{\Delta} \in \{-45, -24, -12, -6, -3,3,6,12,24,45\}$, and fitted a psychometric function for each subject using logit regression. This procedure provided us with a measure of a subject's choice bias and sensitivity. We corrected the psychometric function for choice bias and then used the corrected psychometric function to infer a set of positive distances associated with the target accuracies 60%, 72.5% and 85%: $\{\bar{\delta}_{60}, \bar{\delta}_{72.5}, \bar{\delta}_{85}\}$. For the main experiment, we defined the set of coherences as, $\mathbf{K} \in \{k_{\text{low}} = .5 * k_{\text{med}},\ k_{\text{high}} = 2 * k_{\text{med}}\}$, and the set of signed distances as, $\mathbf{\Delta} \in \{-45, -\bar{\delta}_{85}, -\bar{\delta}_{72.5}, -\bar{\delta}_{60}, \bar{\delta}_{60}, \bar{\delta}_{72.5}, \bar{\delta}_{85}, 45\}$. The extreme signed distances were probed three times less often; they served as anchor points for construction of psychometric functions. Pre-scan calibration data are shown in **Fig. S1**.

Stimulus calibration: scan session

Subjects performed one further calibration phase at the start of the scan session during the acquisition of structural images. Coherence was fixed at k_{med} and we calibrated the signed distances associated with 60% and 85% accuracy, $\pm\bar{\delta}_{60}$ and $\pm\bar{\delta}_{85}$, using a QUEST procedure (54). For the main experiment, we defined the set of coherences as $\mathbf{K} \in \{k_{\text{low}},\ k_{\text{high}}\}$ and the set of signed distances as $\mathbf{\Delta} \in \{-\bar{\delta}_{85}, -\bar{\delta}_{60}, \bar{\delta}_{60}, \bar{\delta}_{85}\}$. Scan calibration data are shown in **Fig. S1**.

Procedure: pre-scan session

The pre-scan session consisted of five runs. Subjects first completed a tutorial, with motion stimuli shown for variable durations (1-5s) and at variable coherences (.30-.60). Subjects then practiced the task (40 trials) at relatively high coherence and extreme distance ($\mathbf{K} = \{.30, .60\}$, $\mathbf{\Delta} = \{-30,30\}$). In this run only, subjects received trial-by-trial feedback (central crossed turned either green, 'correct', or red, 'incorrect', after a choice), with the aim to familiarise subjects with angular judgements in circular space. Next, subjects performed calibration phases 1 (120 trials) and 2 (260 trials). Finally, subjects performed the main experiment (540 trials). The pre-scan session lasted 2 hours.

Procedure: scan session

The scan session was made up of seven runs. Subjects first performed the calibration phase during the acquisition of structural images (180 trials) and then the main experiment over five runs (5 x 112 = 560 trials). In the scan session only, sampling of motion direction was yoked such that there was a 50% probability that a given trial had the same motion direction as on the previous trial. This manipulation is orthogonal to coherence and distance, and direction does not tell subjects whether the upcoming choice should be CW or CCW. Effects of repetition on direction-specific neural responses will be analysed in a separate paper. In the final scan run, subjects viewed alternating displays (12s) of static and dynamic dots ($\mathbf{K} = \{.50\}$) (2 x 12 = 24 displays).

FMRI

Acquisition

MRI data were acquired on a 3T Siemens Trio scanner. T1-weighted structural images were acquired using a 3D MDEFT sequence: 1x1x1 mm resolution voxels; 176 sagittal slices, 256x224 matrix; TR =

10.55ms; TE = 3.14ms; TI = 680ms. BOLD $T2^*$-weighted functional images were acquired using a gradient-echo EPI pulse sequence: 3x3x3 mm resolution voxels; 48 transverse slices, 64x74 matrix; TR = 3.36; TE = 30ms; slice tilt = 0 degrees, slice thickness = 2 mm; inter-slice gap = 1mm; ascending slice order. Field maps were acquired using a double-echo FLASH (gradient echo) sequence: TE1 = 10ms; TE2 = 12.46ms; 64 slices were acquired with 2 mm slice thickness and a 1 mm gap; in-plane field of view is 192x192 mm2 with 3x3 mm2 resolution.

Pre-processing

MRI data were pre-processed using SPM12 (Wellcome Trust, London). The first 5 volumes of each functional run were discarded to allow for T1 equilibration. Functional images were slice-time corrected, realigned and unwarped using the field maps (55). Structural T1-weighted images were co-registered to the mean functional image of each subject using the iterative mutual-information algorithm. Each subject's structural image was segmented into grey matter, white matter and cerebral spinal fluid using a nonlinear deformation field to map it onto a template tissue probability map (56). These deformations were applied to structural and functional images to create new images spatially normalised to the Montreal Neurological Institute (MNI) space and interpolated to 2x2x2 mm voxels. Normalized images were spatially smoothed using a Gaussian kernel with full-width half-maximum of 8mm. Motion correction parameters estimated from the realignment procedure and their first temporal derivatives – yielding a total of 12 'motion' regressors – were included as confounds in the first-level analysis for each subject.

Physiological monitoring

Peripheral measurements of subject pulse and breathing were made together with scanner slice synchronisation pulses using a Spike2 data acquisition system (Cambridge Electronic Design Limited, Cambridge UK). The cardiac pulse signal was measured using an MRI compatible pulse oximeter (Model 8600 F0, Nonin Medical, Inc. Plymouth, MN) attached to the subject's finger. The respiratory signal, thoracic movement, was monitored using a pneumatic belt positioned around the abdomen close to the diaphragm. A physiological noise model was constructed to account for artifacts related to cardiac and respiratory phase and changes in respiratory volume using an in-house MATLAB toolbox (57). Models for cardiac and respiratory phase and their aliased harmonics were based on RETROICOR (58) and a similar, earlier method (59). Basis sets of sine and cosine Fourier series components extending to the 3rd harmonic were used to model physiological fluctuations. Additional terms were included to model changes in respiratory volume (60, 61) and heart rate (62). This procedure resulted in a total of 14 'biophysical regressors' which were sampled at a reference slice in each image volume to give a set of values for each time point. The regressors were included as confounds in the first-level analysis for each subject.

Quantification and statistical analysis

Choice reaction time

We excluded a trial when a subject's choice reaction time was 2.5 SD below or above their grand mean reaction time computed separately for the pre-scan and the scan sessions. This procedure resulted in the exclusion of approximately 2% of trials per subject per session.

Behaviour

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 z-transformed all predictors. 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 zero (one-sample *t*-test). We used ordinal regression to construct the computational model of subjective confidence for fMRI analysis (**Fig. 4a**). The model included eight predictors: log-transformed choice reaction time, accuracy, coherence, distance, and interaction terms for coherence x distance, accuracy x coherence, accuracy x distance and accuracy x

coherence x distance. We fitted the model to each subject using the pre-scan data and created trial-by-trial predictions of confidence level for fMRI analysis using the scan data. Since we sub-selected and re-calibrated the stimulus for the scan session, we did not z-transform predictors to facilitate between-session generalisation of model predictions. The model provided a good fit to the pre-scan data and generalised well to the scan data (**Fig. S8**).

Hierarchical drift-diffusion model

Decision formation was modelled using the drift-diffusion model (DDM). The DDM models two-choice decision-making as a process of accumulating noisy evidence over time with a certain speed, or drift-rate (*v*), until one of two decision thresholds is crossed and the associated response is executed. Larger threshold separation (*a*) leads to slower responses but more accurate responding. The DDM includes a non-decisional component (*t*) which captures time needed for stimulus encoding and motor execution. We employed hierarchical Bayesian estimation of subjects' DDM parameters using the HDDM toolbox (63) (http://ski.clps.brown.edu/hddm_docs/). We fitted non-decision time (*t*), decision threshold (*t*) and drift-rate (*v*) separately for each condition of our factorial design (pre-scan: 8 conditions; scan: 4 conditions). We also included inter-trial variability in drift-rate (*sv*) and non-decision time (*st*) as free parameters that were constant across conditions. The HDDM toolbox applies Markov Chain Monte-Carlo (MCMC) sampling methods to approximate posterior distributions over the DDM parameters; we ran 1 chain with 5000 samples, with the first 1000 samples discarded as burn-in. We extracted mean group-level posterior estimates for visualisation of DDM parameters in each condition (**Fig. S4**) and generating posterior predictives (**Fig. 2a-b**). We used the *simuldiff* function from the DMAT toolbox (64) (http://ppw.kuleuven.be/okp/software/dmat/) to generate posterior predictives: for each condition, we simulated 100,000 trials under the hierarchically estimated DDM parameters and calculated mean choice accuracy and mean choice reaction time in that condition.

FMRI

Whole-brain general linear models

The whole-brain analysis shown in **Fig. 3a** was based on a single event-related General Linear Model (GLM1). We categorised each trial according to whether Coherence (C) and Distance (D) was Low (L) or High (H), yielding four trial types: CL&DL, CL&DH, CH&DL, CH&DH. We modelled the four trial types with separate 'condition' regressors. Choice reaction-time outlier trials were assigned to a separate 'outlier' regressor. These regressors were specified as boxcars time-locked to the onset of the motion stimulus and spanning until choice (65). We modelled confidence events with a separate regressor, using a stick function time-locked to the onset of the scale. This 'rate' regressor was parametrically modulated by the (*z*-transformed) number of button presses used to navigate the marker. We included motion and biophysical parameters (see above) as additional 'nuisance' regressors. Regressors were convolved with a canonical hemodynamic response function. Regressors were modelled 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 (128s cut-off) was applied to remove low-frequency drifts.

Group-level significance was assessed by applying one-sample *t*-tests against zero to the first-level contrast images. Contrast images for our factorial analysis were specified as: coherence = [CL&DL: -1, CL&DH: -1, CH&DL: 1, CH&DH: 1]; distance = [CL&DL: -1, CL&DH: 1, CH&DL: -1; CH&DH: 1]; coherence x distance = [CL&DL: 1, CL&DH: -1, CH&DL = -1, CH&DH = 1]. Both negatively and positively signed *t*-statistics arising from these contrasts were investigated. Inclusive/exclusive masks for our factorial analysis were created using SPM's *imcalc* function; each second-level contrast image was liberally thresholded at $p < .05$, uncorrected, the default setting for masking analysis in SPM, before creating a mask. We report clusters significant at $p < .05$, FWE-corrected for multiple comparisons, with a cluster-defining threshold of $p < .001$, uncorrected. Numerical simulations and tests of empirical data collected under the null hypothesis show that this combination of cluster-defining threshold and random field theory produces appropriate control over false positives (66).

We obtained similar results in GLMs that introduced modifications to GLM1 (**Fig. S5**). In GLM1C, only correct trials were included. In GLM2, the four condition regressors had a fixed duration of 1s 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 1s and which 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 were allowed to compete to explain variance. The rationale behind the three families of GLMs is as follows: GLM1s assumes that choice reaction time is not a confound but reflects relevant neural processing; GLM2s controls for within-condition variation in choice reaction time; and GLM3s controls for between-condition variation in choice reaction time.

Clusters surviving whole-brain correction post- and pre-masking for GLM1 and GLM1C are shown in **Tables S1-S4**. Activations were visualised using the xjview toolbox for SPM and MRIcron.

Regions of interest

Subject-specific ROI masks were created by applying a leave-one-out procedure to activations from GLM1. For each subject, we performed second-level analyses excluding that subject's data, and specified individual ROI masks using the second-level *t*-maps (cluster-defining threshold: $p < .001$, uncorrected). We obtained clusters associated with a main effect of coherence (bilateral IPS), a main effect of distance (pre-SMA), and the coherence x distance interaction (pgACC). This ROI procedure ensured unbiased ROI selection as each subject's neural activity did not contribute to the selection of ROIs for that subject. ROI masks for MT+ were based on the localiser scan; we created a bilateral group mask using the second-level contrast between dynamic and static motion, and then, for each subject, created bilateral MT+ masks (8-mm sphere) around the subject-specific peaks inside the bilateral group mask. ROI masks for bilateral ventral striatum were specified using the Oxford-GSK-Imanova Striatal Structural atlas included with FSL. ROI masks for bilateral rostrolateral prefrontal cortex were specified using the maps provided by Neubert and colleagues (67) (union of '46' and 'fpl').

Up-sampled activity estimates

We transformed each ROI mask from MNI to native space and extracted pre-processed BOLD timecourses as the average of voxels within the mask. For each scan run, we regressed out variation due to head motion and biophysical responses, applied a high-pass filter (128s cut-off) to remove low-frequency drifts, and oversampled the BOLD timecourse by a factor of ~23 (time resolution of .144s). For each trial, we extracted activity estimates in a 12s window (84 time points), time-locked to 1s before the onset of the motion stimulus or 2s before the onset of the confidence scale (only for reward-magnitude analysis). We used linear regression to predict ROI activity time courses. More specifically, we applied a linear regression to each time point and then, by concatenating beta-weights across time points, created a beta-weight timecourse for each predictor of a regression model. We performed this step separately for each subject and pooled beta-weight timecourses across subjects for visualisation. We tested group-level significance using a permutation procedure. We repeated the above steps 1,001 times, for each repetition shuffling a subject's trial timecourses, and then asked, at each time point, whether the *t*-statics associated with the empirically observed group-level effect (one-sample *t*-test of beta-weights pooled across subjects against zero) was smaller or larger than 97.5% of the *t*-statistics obtained from the permutation procedure.

Single-trial activity estimates

We estimated single-trial BOLD responses as a beta timeseries for the individual-difference analyses in **Fig. 4c**. This was achieved by means of an event-related model with a separate regressor for each trial. Regressors were boxcars time-locked to the onset of the motion display and spanning until choice (65). Each regressor was convolved with a canonical hemodynamic response function. We included motion and biophysical parameters (see above) as 'nuisance' regressors. Regressors were modelled 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 (128s cut-off) was applied to remove low-frequency drifts. One important consideration in using single-trial activity estimates is that a beta for a given trial can be strongly affected by acquisition artefacts that co-occur with that trial (e.g., scanner

pulse artefacts). Therefore, for each subject, we computed the grand-mean beta estimate across both voxels and trials and excluded any trial whose mean beta estimate across voxels was 3 SDs below or above this grand mean (68). About 2% of trials were excluded per subject. Finally, we used the ROI masks to extract ROI single-trial activity estimates.

Resource sharing

All requests for further information, data and associated code should be directed to and will be fulfilled by the corresponding author: Dan Bang (danbang.db@gmail.com).

REFERENCES

1. Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic quantities for different goals. *Nat Neurosci* 19(3):366–374.
2. Meyniel F, Sigman M, Mainen ZF (2015) Confidence as Bayesian probability: from neural origins to behavior. *Neuron* 88(1):78–92.
3. Lak A, et al. (2014) Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* 84(1):190–201.
4. Bahrami B, et al. (2010) Optimally interacting minds. *Science* 329(5995):1081–1085.
5. Bang D, et al. (2017) Confidence matching in group decision-making. *Nat Hum Behav* 1(6):117.
6. Fleming SM, Dolan RJ, Frith CD (2012) Metacognition: computation, biology and function. *Philos Trans R Soc B Biol Sci* 367(1594):1280–1286.
7. Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324(5928):759–764.
8. 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(6):749–55.
9. 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(6):821–832.
10. Kepecs A, Uchida N, Zariwala H a, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455(7210):227–231.
11. De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16(1):105–110.
12. Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18(8):1159–1167.
13. 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(1):118–130.
14. Fleck MS, Daselaar SM, Dobbins IG, Cabeza R (2006) Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cereb Cortex* 16(11):1623–1630.
15. Fleming SM, Huijgen J, Dolan RJ (2012) Prefrontal contributions to metacognition in perceptual decision making. *J Neurosci* 32(18):6117–6125.
16. Hilgenstock R, Weiss T, Witte OW (2014) You'd better think twice: post-decision perceptual confidence. *Neuroimage* 99:323–331.
17. Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J Neurosci* 22(21):9475–9489.
18. Jazayeri M, Movshon JA (2007) A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* 446(7138):912–915.
19. Brier GW (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78(1):1–3.
20. Ma WJ, Jazayeri M (2014) Neural coding of uncertainty and probability. *Annu Rev Neurosci* 37(1):205–220.
21. Adler WT, Ma WJ (2016) The computations underlying human confidence reports are probabilistic, but not Bayesian. *bioRxiv*:1–14.
22. Shadlen MN, Shohamy D (2016) Decision making and sequential sampling from memory. *Neuron* 90(5):927–939.
23. Braddick OJ, et al. (2001) Brain areas sensitive to coherent visual motion. *Perception* 30(1):61–72.
24. Rees G, Friston K, Koch C (2000) A direct quantitative relationship between the functional properties of human and macaque V5. *NatNeurosci* 3(7):716–723.
25. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci* 12(12):4745–4765.
26. Fanini A, Assad JA (2009) Direction selectivity of neurons in the Macaque lateral intraparietal area. 289–305.
27. Kouneiher F, Charron S, Koechlin E (2009) Motivation and cognitive control in the human prefrontal cortex. *Nat Neurosci* 12(7):939–945.
28. O'Doherty J, et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304(5669):452–454.
29. Pagnoni G, Zink CF, Montague PR, Berns GS (2002) Activity in human ventral striatum locked to errors of reward prediction. *Nat Neurosci* 5(2):97–98.
30. Wittmann MK, et al. (2016) Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* 91(2):482–493.

31. 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 U S A*:1–10.
32. Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* 137(10):2811–2822.
33. David AS, Bedford N, Wiffen B, Gilleen J (2014) Failures of metacognition and lack of insight in neuropsychiatric disorders. *Cogn Neurosci Metacognition* 9783642451:345–365.
34. DeLozier SJ, Davalos D (2016) A systematic review of metacognitive differences between Alzheimer's disease and frontotemporal dementia. *Am J Alzheimer's Dis Other Dementias* 31(5):381–388.
35. Goldstein RZ, et al. (2009) The neurocircuitry of impaired insight in drug addiction. *Trends Cogn Sci* 13(9):372–380.
36. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. *Science* 329(5998):1541–1543.
37. McCurdy LY, et al. (2013) Anatomical coupling between distinct metacognitive systems for memory and visual perception. *J Neurosci* 33(5):1897–906.
38. Allen M, et al. (2017) Metacognitive ability correlates with hippocampal and prefrontal microstructure. *Neuroimage* 149(May 2016):415–423.
39. Badre D, Wagner AD (2004) Selection, integration, and conflict monitoring. *Neuron* 41(3):473–487.
40. Dosenbach NUF, et al. (2006) A Core System for the Implementation of Task Sets. *Neuron* 50(5):799–812.
41. Gilbert SJ, et al. (2006) Functional Specialization within Rostral Prefrontal Cortex (Area 10): A Meta-analysis. *J Cogn Neurosci* 18(6):932–948.
42. Koechlin E, Ody CC, Kouneiher FF (2003) The architecture of cognitive control in the human prefrontal cortex. *Science* 302(November):1181–5.
43. Sakai K (2008) Task Set and Prefrontal Cortex. *Annu Rev Neurosci* 31(1):219–245.
44. 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(12):2563–2586.
45. Cook EP, Maunsell JHR (2002) Dynamics of neuronal responses in macaque MT and VIP during motion detection. *Nat Neurosci* 5(10):985–994.
46. Beck JM, et al. (2008) Probabilistic population codes for Bayesian decision making. *Neuron* 60(6):1142–1152.
47. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415(6870):429–433.
48. Drugowitsch J, DeAngelis GC, Klier EM, Angelaki DE, Pouget A (2014) Optimal multisensory decision-making in a reaction-time task. *Elife* 3:1–19.
49. Drugowitsch J, DeAngelis GC, Angelaki DE, Pouget A (2015) Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *Elife* 4(JUNE2015):1–11.
50. van den Berg R, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM (2016) Confidence is the bridge between multi-stage decisions. *Curr Biol* 26(23):3157–3168.
51. Purcell BA, Kiani R (2016) Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc Natl Acad Sci*:201524685.
52. Meyniel F, Schlunegger D, Dehaene S (2015) The sense of confidence during probabilistic learning: a normative account. *PLOS Comput Biol* 11(6):e1004305.
53. Jazayeri M, Movshon JA (2006) Optimal representation of sensory information by neural populations. *Nat Neurosci* 9(5):690–696.
54. Watson AB, Pelli DG (1983) QUEST: A Bayesian adaptive psychometric method. *Percept Psychophys* 33(2):113–120.
55. Andersson JL, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling geometric deformations in EPI time series. *Neuroimage* 13(5):903–919.
56. Ashburner J, Friston KJ (2005) Unified segmentation. *Neuroimage* 26(3):839–851.
57. Hutton C, et al. (2011) The impact of physiological noise correction on fMRI at 7T. *Neuroimage* 57(1):101–112.
58. Glover GH, Li T-Q, Ress D (2000) Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn Reson Med* 44(1):162–167.
59. Josephs O, Howseman AM, Friston KJ, Turner R (1997) Physiological noise modelling for multi-slice EPI fMRI using SPM. *Proceedings of the 5th Annual Meeting of ISMRM* (Vancouver, Canada), p 1682.

60. Birn RM, Smith MA, Jones TB, Bandettini PA (2008) The Respiration Response Function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *Neuroimage* 40(2):644–654.
61. Birn RM, Diamond JB, Smith MA, Bandettini PA (2006) Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *Neuroimage* 31(4):1536–1548.
62. Chang C, Glover GH (2009) Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. *Neuroimage* 47(4):1448–1459.
63. Wiecki T V., Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform* 7(August):14.
64. Vandekerckhove J, Tuerlinckx F (2008) Diffusion model analysis with MATLAB: A DMAT primer. *Behav Res Methods* 40(1):61–72.
65. Grinband J, Wager TD, Lindquist M, Ferrera VP, Hirsch J (2008) Detection of time-varying signals in event-related fMRI designs. *Neuroimage* 43(3):509–520.
66. Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci* 113(28):7900–7905.
67. Neubert FX, Mars RB, Thomas AG, Sallet J, Rushworth MFS (2014) Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron* 81(3):700–713.
68. Atlas LY, Bolger N, Lindquist MA, Wager TD (2010) Brain Mediators of Predictive Cue Effects on Perceived Pain. *J Neurosci* 30(39):12964–12977.

SUPPLEMENTAL FIGURES

Dissociating decision confidence from evidence reliability in the human brain

Dan Bang* and Stephen M. Fleming

*Correspondence: danbang.db@gmail.com

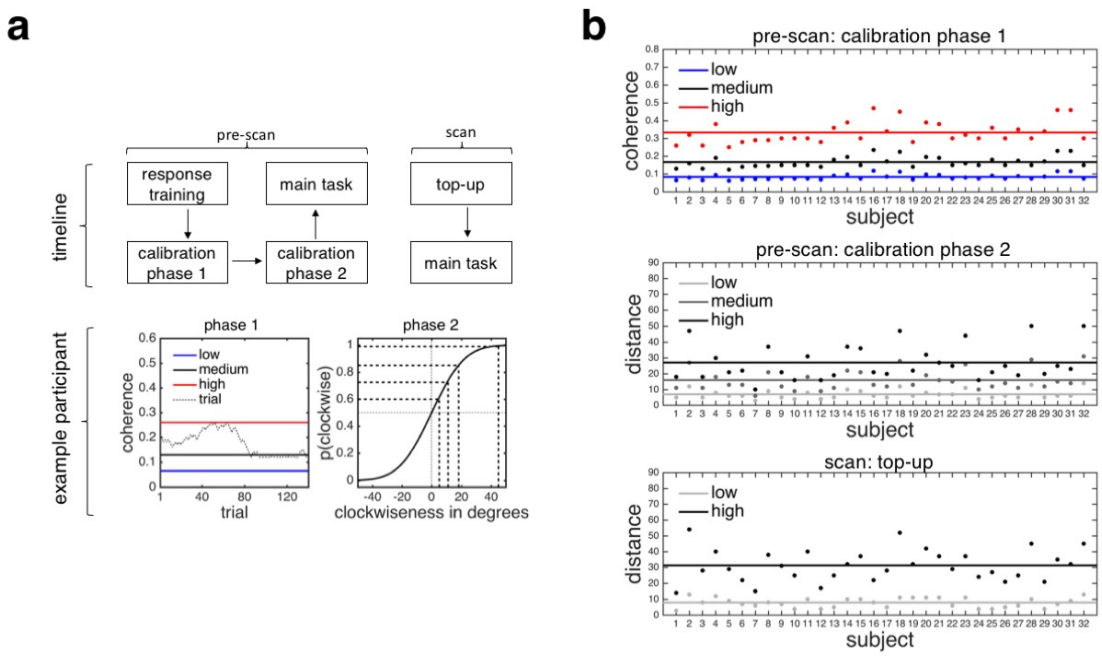

Figure S1. Stimulus calibration. **a**, Experimental procedure. (*top*) Subjects took part in separate pre-scan and scan sessions (2-14 days between sessions). In pre-scan calibration phase 1, we calibrated coherence (medium) at a fixed distance for each subject using a 2-down-1-up procedure. We used 50% (low) and 200% (high) of the medium coherence value in the main task, both in the pre-scan and in the scan session. In pre-scan calibration phase 2, we presented each subject with a range of distances at medium coherence so as to construct their psychometric function. We extracted a set of distances associated with target levels of accuracy for the pre-scan main task. In the scan top-up session, using an adaptive QUEST procedure, we re-calibrated a subset of distances from the pre-scan session at medium coherence for the scan main task. *(bottom*) Example data from pre-scan calibration phase 1 and 2. **b**, Calibration results. Calibrated stimulus values are shown for (*top)* pre-scan calibration phase 1, (*middle*) pre-scan calibration phase 2 and (*bottom*) scan top-up.

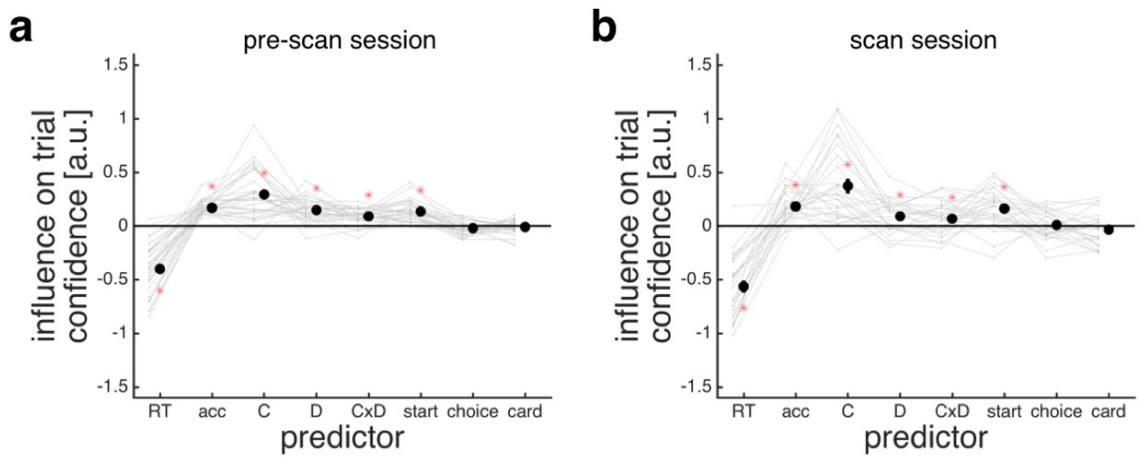

Figure S2. Extended regression analysis of predictors of confidence estimates. **a**, Pre-scan session. Coefficients (y-axis) from a trial-by-trial ordinal regression encoding the influence of stimulus and choice features (x-axis) on current confidence. Predictors included: choice reaction time (RT), choice accuracy (acc), coherence (C), distance (D), the interaction between coherence and distance (CxD), the initial position of the confidence marker (start), whether the choice was clockwise (choice), and the cardinality of the motion direction (card). **b,** Scan session. Same regression analysis as in panel A, except that we also included the reward factor and whether the motion direction was the same as on the previous trial; these predictors were not significant and are not shown here. Confidence estimates were elicited on every 5-10 trials (84 trials in total) in the scan session. **a-b**, We performed a regression for each subject and tested significance (red asterisk) by comparing the coefficients pooled across subjects to zero (one-sample *t*-test). We log-transformed choice reaction time and z-transformed predictors. Faint lines connect the coefficients for each subject. Data are represented as group mean ± SEM.

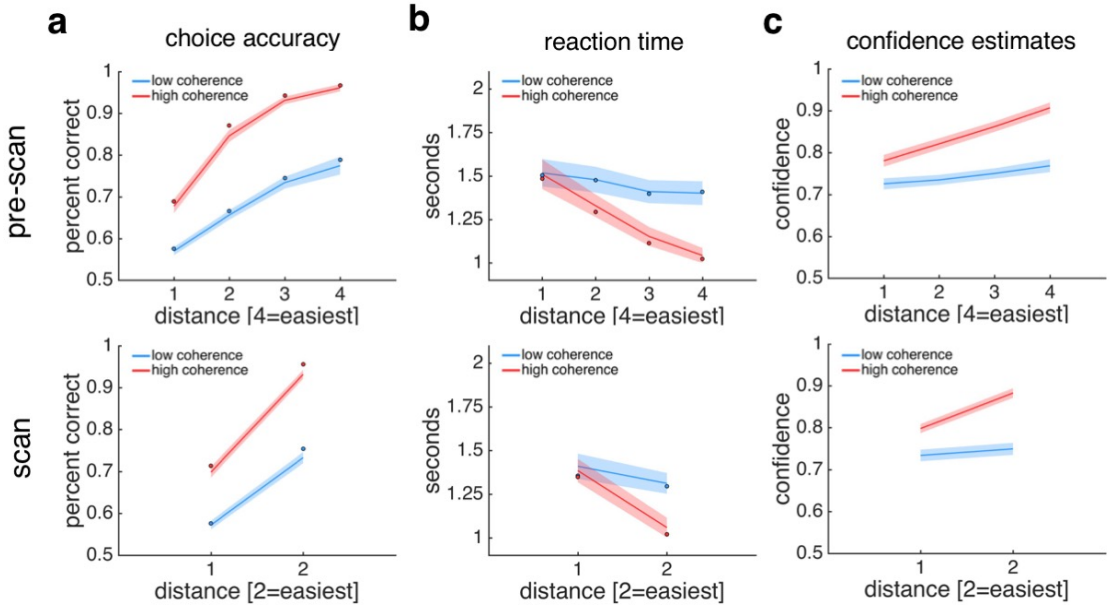

Figure S3. Behavioural results and HDDM fits from pre-scan and scan sessions. **a**, Choice accuracy. In the scan session, subjects were also more likely to be correct when coherence was high and when distance was high (logistic regression, coherence: $t(31) = 13.009$, $p < .001$, distance: $t(31) = 15.172$, $p < .001$, interaction: $t(31) = 10.212$, $p < .001$). **b**, Reaction time. In the scan session, subjects were also faster when coherence was high and when distance was high (linear regression, coherence: $t(31) = -5.031$, $p < .001$, distance: $t(31) = -8.892$, $p < .001$, interaction: $t(31) = -6.390$, $p < .001$). **c**, Confidence estimates. In the scan session, subjects were also more confident when coherence was high and when distance was high (ordinal probit regression, coherence: $t(31) = 5.959$, $p < .001$, distance: $t(31) = 5.786$, $p < .001$, interaction: $t(31) = 4.665$, $p < .001$). **a-b**, Solid dots are posterior predictives from the hierarchical drift-diffusion model fit to subjects' responses separately for each condition. **a-c**, See main text for test statistics for pre-scan session. Data are represented as mean ± SEM.

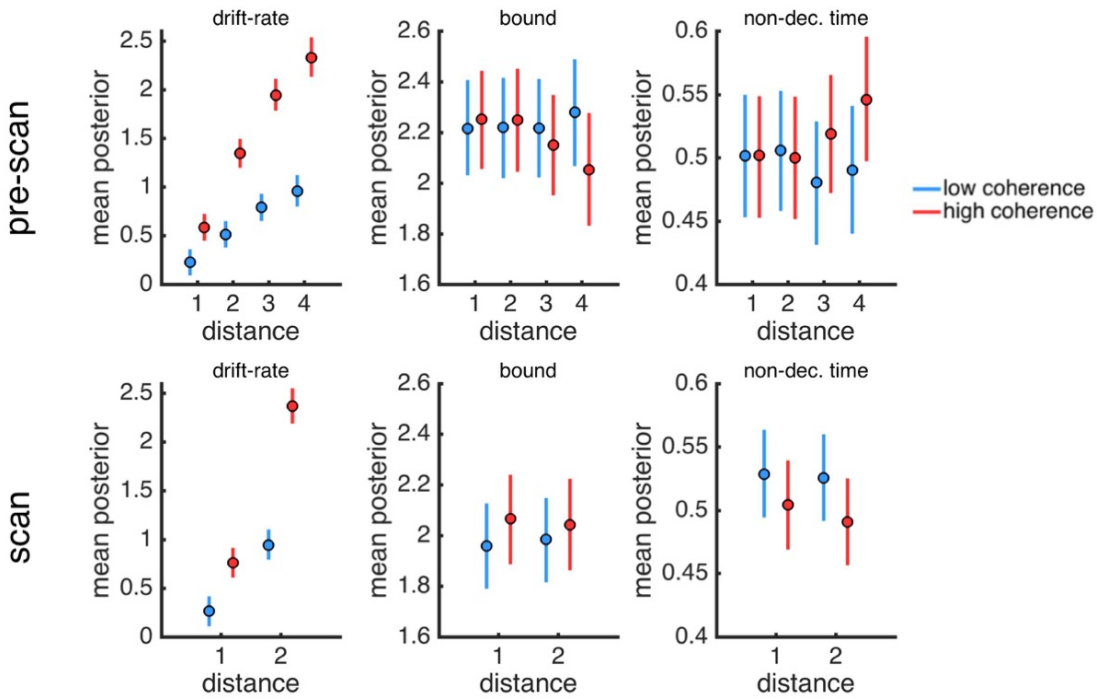

Figure S4. Drift-diffusion model. Mean posterior estimates obtained from hierarchical Bayesian estimation of subjects' drift-diffusion model parameters using the HDDM toolbox for (*top*) the pre-scan session and (*bottom*) the scan session. We fitted non-decision time (*t*), decision threshold (*t*) and drift-rate (*v*) separately for each condition of our factorial design (pre-scan: 8 conditions; scan: 4 conditions). We also included inter-trial variability in drift-rate (*sv*) and non-decision time (*st*) but these parameters were fitted across conditions (pre-scan, *sv* = 0.904, *st* = 0.232; scan, *sv* = 1.063, *st* = 0.270). Error bars indicate 95%-confidence intervals as estimated from the posterior distributions over parameters.

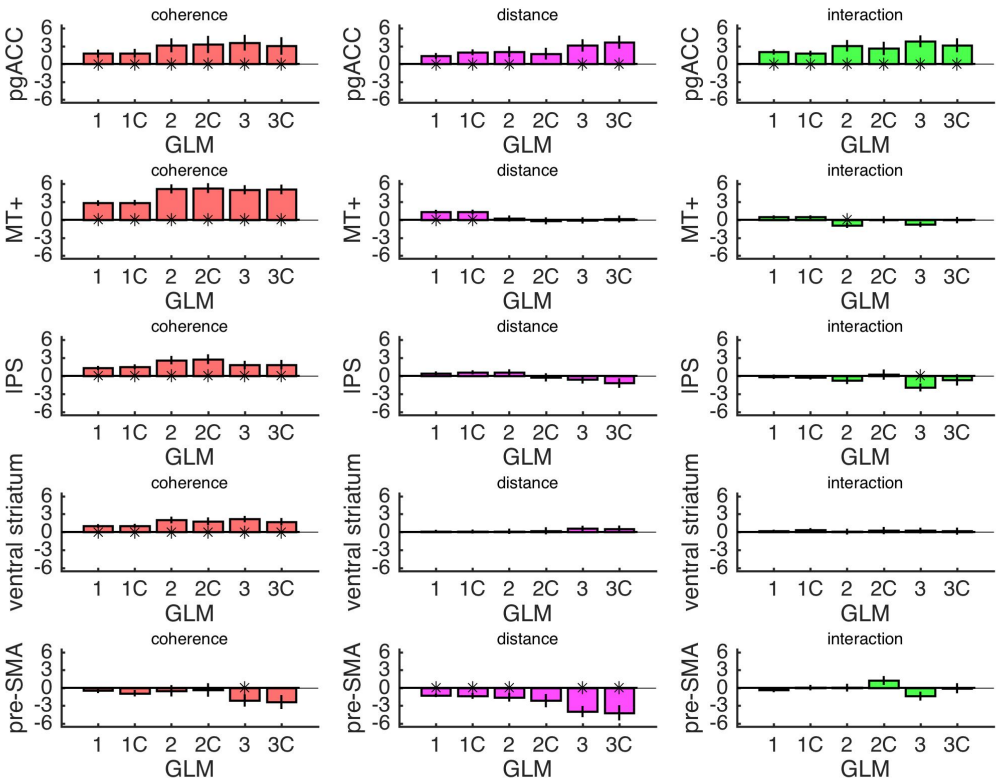

Figure S5. ROI contrast estimates from whole-brain control models. The plots show contrast estimates (y-axis) for coherence, distance and the coherence x distance interaction extracted from a set of control models (x-axis). We tested significance by comparing the contrast estimates pooled across subjects against zero (one-sample *t*-test). The rationale behind the three families of GLMs is as follows: GLM1s assumes that choice reaction time is not a confound but reflects relevant neural processing; GLM2s controls for within-condition variation in choice reaction time; and GLM3s controls for between-condition variation in choice reaction time. C: correct trials only. Data are represented as group mean ± SEM.

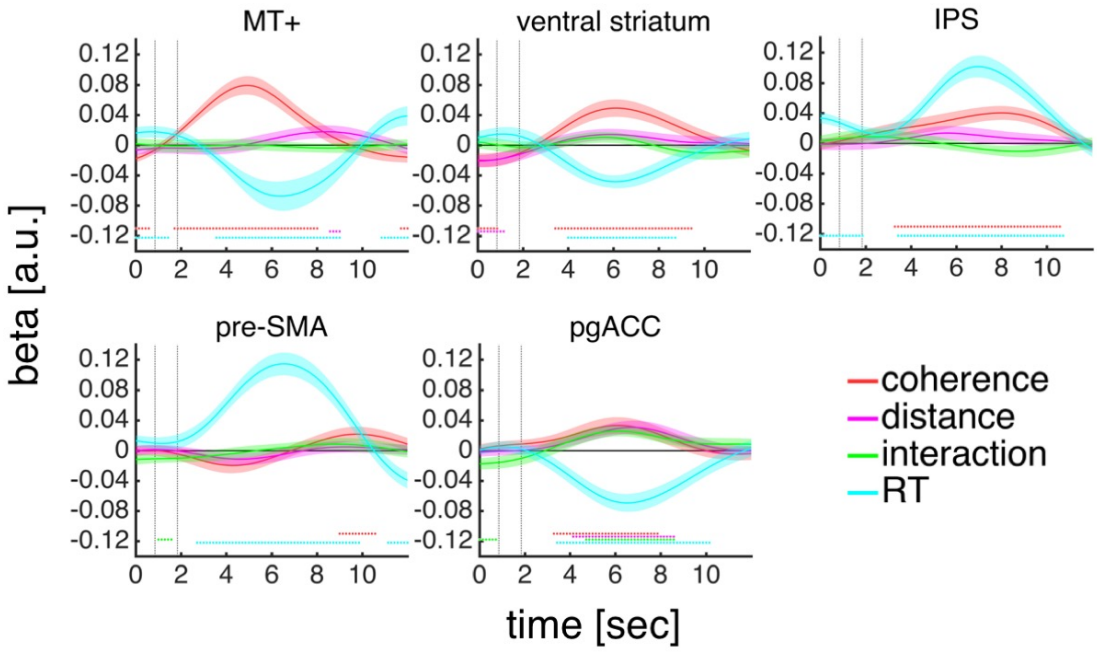

Figure S6. Factorial analysis of ROI activity timecourses. General linear model analysis of the effects of coherence, distance, the coherence x distance interaction and choice reaction time on ROI activity timecourses. We tested group-level significance (dots) using a permutation procedure. We repeated our analysis 1,001 times, each time shuffling the trial timecourses, and then asked, at each time point, whether the *t*-statics associated with the empirically observed group-level effect (one-sample *t*-test of beta-weights pooled across subjects against zero) was smaller or larger than 97.5% of the *t*-statistics obtained through the permutation procedure. Vertical dashed lines indicate the onsets of the motion stimulus and the choice phase. Data are represented as group mean ± SEM.

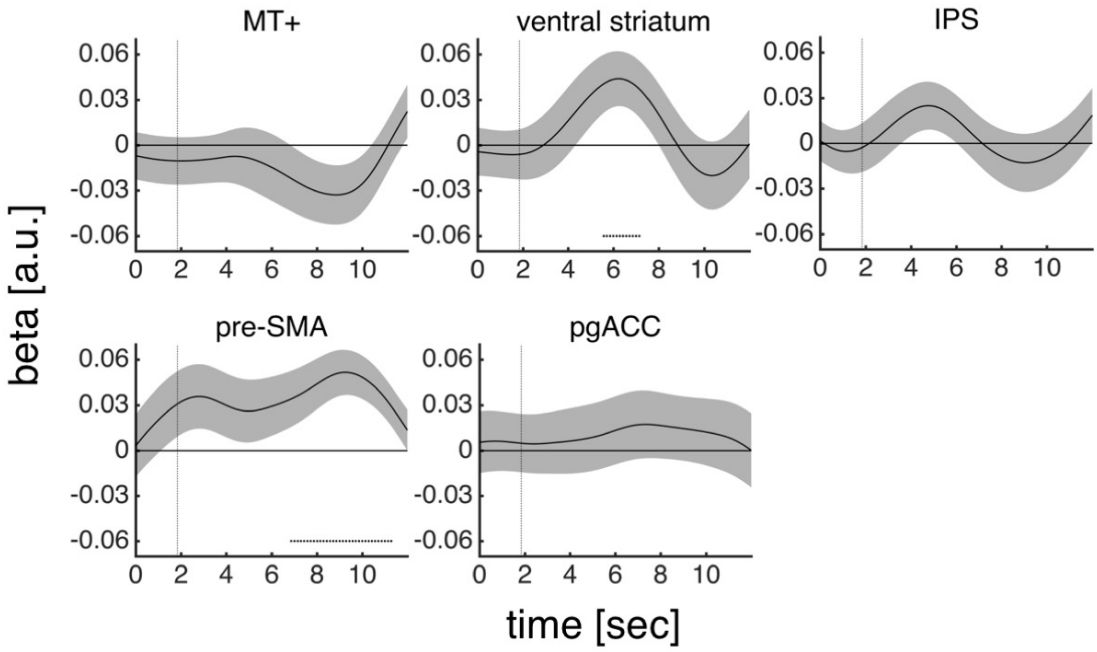

Figure S7. Encoding of reward in ROI activity timecourses. General linear model analysis of the effects of reward magnitude (high versus low) on ROI activity timecourses. We tested group-level significance (dots) using a permutation procedure. We repeated our analysis 1,001 times, each time shuffling the trial timecourses, and then asked, at each time point, whether the *t*-statics associated with the empirically observed group-level effect (one-sample *t*-test of beta-weights pooled across subjects against zero) was smaller or larger than 97.5% of the *t*-statistics obtained through the permutation procedure. Vertical dashed line indicates the onset of the reward magnitude cue. Data, which only include confidence trials (15% of trials), are represented as group mean ± SEM.

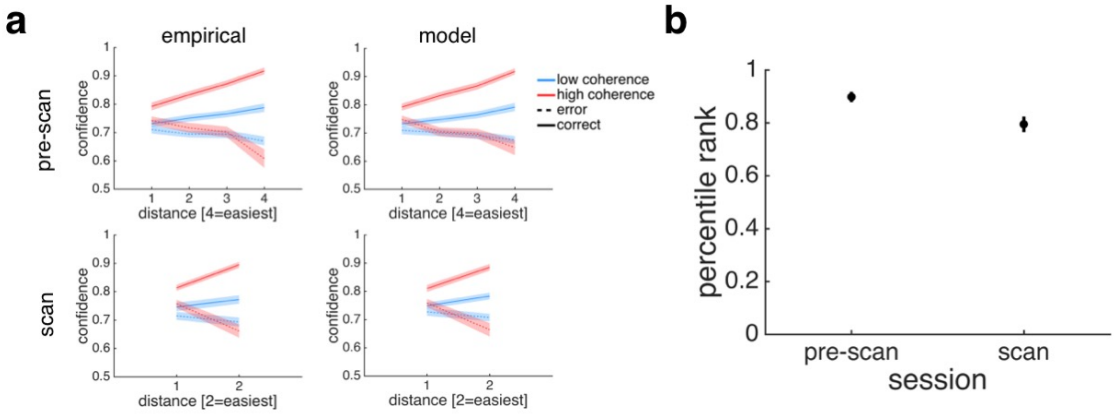

Figure S8. Evaluation of model of subjective confidence. **a**, Model qualitatively reproduces behaviour. The plots show (*left*) empirically observed and (*right*) model-derived confidence in the conditions of our factorial design split by choice accuracy (distance x coherence x choice accuracy) in (*top*) the pre-scan and (*bottom*) the scan session. The model was fitted to data from the pre-scan session and then used to create out-of-sample predictions about the scan session. We show empirical and predicted data from confidence trials only (15% of trials) for the scan session. **b**, Model provided good fits and generalised between sessions. For each subject, $s = i$, we computed the likelihood of the data under a model, m, fitted to the subject's own pre-scan data, $m_{s=i}$, and under the models fitted to the pre-scan data of every other subject, $m_{s\neq i}$. The plot shows (y-axis) the percentile rank of the likelihood under a subject's own model, $m_{s=i}$, compared to all other models, $m_{s\neq i}$, for (x-axis) the pre-scan and the scan sessions. Here higher rank indicates better model fits. As described in **Fig. 4a**, the model predicts, for each trial, the probability that a given response is made. Model predictions were evaluated by summing the likelihood of the observed trial-by-trial responses under this probability distribution. **a-b**, Data are represented as group mean ± SEM.

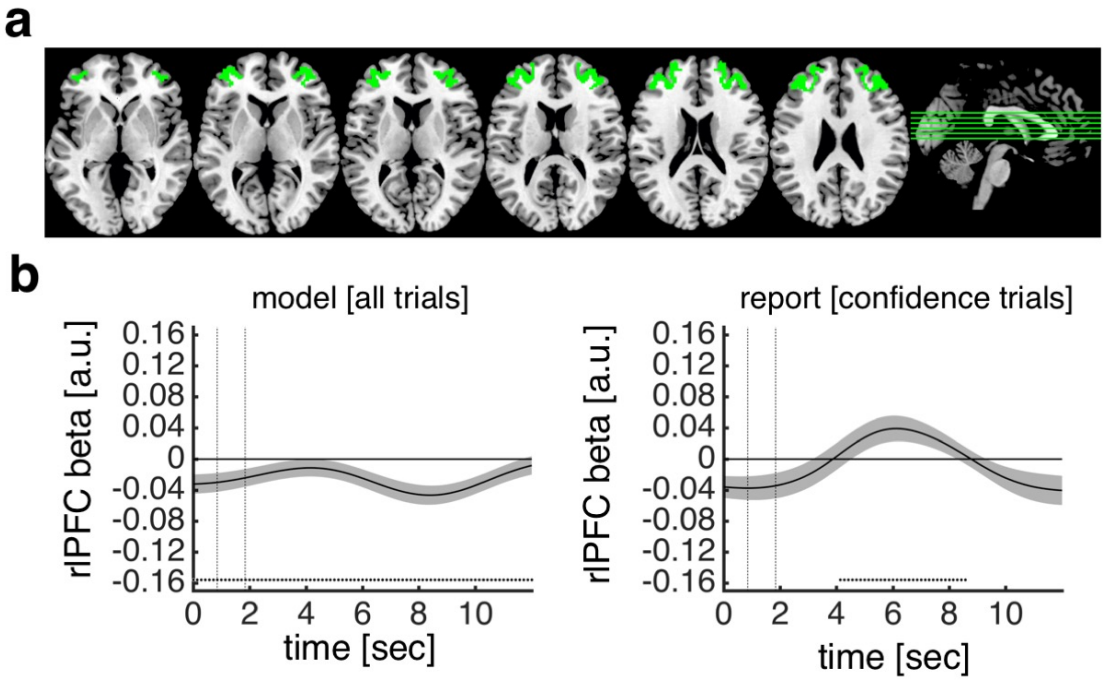

Figure S9. Activity in rlPFC predicts confidence estimates. **a**, ROI mask for rlPFC. **b,** General linear model analysis of encoding of model-derived subjective confidence (all trials) and reported confidence (confidence trials) in pgACC activity timecourses. We tested group-level significance (dots) using a permutation procedure. We repeated our analysis 1,001 times, each time shuffling the trial timecourses, and then asked, at each time point, whether the *t*-statics associated with the empirically observed group-level effect (one-sample *t*-test of beta-weights pooled across subjects against zero) was smaller or larger than 97.5% of the *t*-statistics obtained through the permutation procedure. Vertical dashed line indicates the onset of the reward magnitude cue. Data are represented as group mean ± SEM.

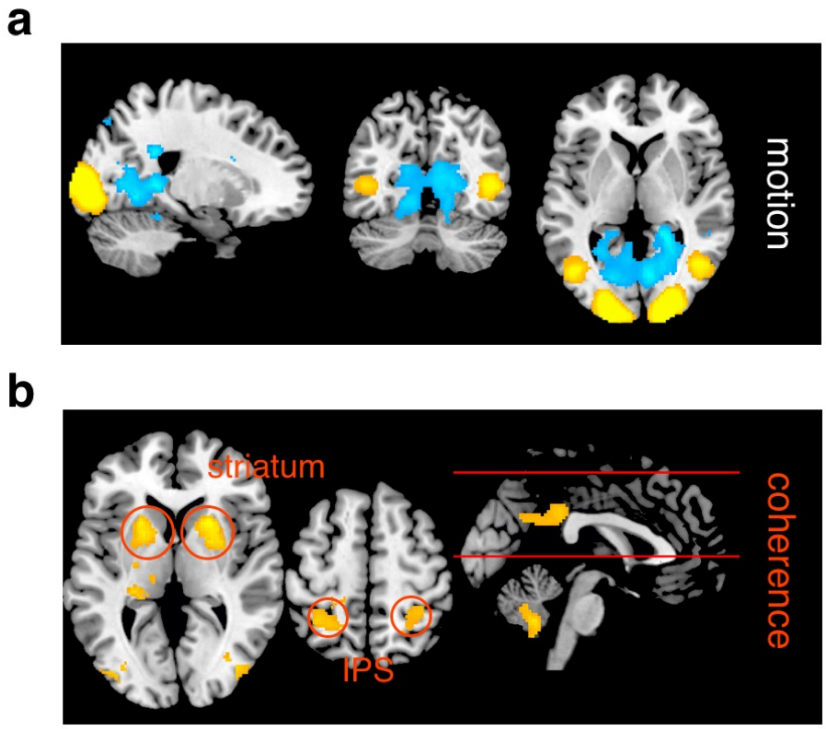

Figure S10. Neural responses to coherently moving dots. **a**, Whole-brain analysis of contrast between coherently moving and static dots in motion-localiser scan. **b**, Whole-brain analysis of main effect of coherence masked by coherent-motion contrast. We applied an exclusive mask constructed from the contrast shown in panel a (thresholded at p < .05, uncorrected) to GLM1. **a-b**, Cluster colours denote positive (warm) and negative (cold) effects. Clusters are significant at $p < .05$, FWE-corrected for multiple comparisons. Cluster-defining threshold: $p < .001$, uncorrected. Images are shown at $p < .001$, uncorrected.

SUPPLEMENTAL TABLES

Dissociating decision confidence from evidence reliability in the human brain

Dan Bang* and Stephen M. Fleming

*Correspondence: danbang.db@gmail.com

GLM1: post-masking

contrast	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	peak voxel MNI coordinates			
high > low coh *exclusive mask: distance and interaction*	striatum	1423	5.41	<.001	26	10	-6	R
	extrastriate (incl. MT+)	2675	5.36	<.001	28	-84	8	R
	striatum	1505	5.14	<.001	-24	10	-10	L
	extrastriate (incl. MT+)	1635	4.72	<.001	-40	-78	2	L
	posterior parietal	373	4.54	.001	38	-26	40	R
	cerebellum	428	4.42	<.001	-4	-56	-42	LR
	posterior cingulate	336	4.39	.001	12	-38	30	LR
	posterior parietal	394	4.25	.001	-32	-50	62	L
high < low coh *exclusive mask: distance and interaction*	----------none----------							
high > low dis *exclusive mask: coherence and interaction*	posterior cingulate	214	4.91	.014	-12	-16	48	LR
	early visual	178	3.77	.016	-16	-98	2	L
high < low dis	pre-SMA	169	4.34	.020	4	20	48	LR
interaction + *inclusive mask: coherence and distance*	pgACC	932	4.82	<.001	-2	44	10	LR
interaction – *inclusive mask: coherence and distance*	----------none----------							

Table S1. Summary of significant activations for GLM1 post-masking. All reported activations are significant at $p < .05$, FWE-corrected for multiple comparisons. Cluster-defining threshold: $p < .001$, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. pre-SMA: pre-supplementary motor area. pgACC: perigenual anterior cingulate cortex.

GLM1: pre-masking

contrast	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	peak voxel MNI coordinates			
high > low coh	striatum	1423	5.41	<.001	26	10	-6	R
	extrastriate (incl. MT+)	2675	5.36	<.001	28	-84	8	R
	striatum	1505	5.14	<.001	-24	10	-10	L
	extrastriate (incl. MT+)	1635	4.72	<.001	-40	-78	2	L
	posterior parietal	373	4.54	.001	38	-26	40	R
	cerebellum	428	4.42	<.001	-4	-56	-42	LR
	posterior cingulate	453	4.39	<.001	12	-38	30	LR
	posterior parietal	394	4.25	.001	-32	-50	62	L
high < low coh	----------none----------							
high > low dis	posterior cingulate	214	4.91	.006	-12	-16	48	LR
	pgACC	246	4.07	.003	12	34	12	LR
	early visual	330	4.02	.006	18	-94	-2	R
	early visual	343	3.86	<.001	-18	-90	2	L
high < low dis	pre-SMA	169	4.34	.020	4	20	48	LR
interaction +	pgACC	1093	4.82	<.001	-2	44	10	LR
	temporal	158	4.43	.027	-60	-18	-18	L
interaction –	precuneus	367	4.44	<.001	0	-64	60	LR

Table S2. Summary of significant activations for GLM1 pre-masking. All reported activations are significant at $p < .05$, FWE-corrected for multiple comparisons. Cluster-defining threshold: $p < .001$, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. pgACC: perigenual anterior cingulate cortex. pre-SMA: pre-supplementary motor area.

GLM1C: post-masking

contrast	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	peak voxel MNI coordinates			
high > low coh *exclusive mask: distance and interaction*	extrastriate (incl. MT+)	791	4.89	<.001	44	-62	4	R
	striatum	467	4.81	<.001	-22	12	-10	L
	striatum	610	4.77	<.001	20	16	-4	R
high < low coh *exclusive mask: distance and interaction*	----------none----------							
high > low dis *exclusive mask: coherence and interaction*	pgACC/vmPFC	570	5.00	.001	10	38	12	LR
	insula	159	4.57	.023	-30	20	-18	L
	insula	140	4.39	.040	32	14	-10	R
	posterior cingulate	150	4.38	.030	0	-12	40	LR
high < low dis *exclusive mask: coherence and interaction*	----------none----------							
interaction + *inclusive mask: coherence and distance*	pgACC	252	4.28	.002	4	44	12	LR
interaction – *inclusive mask: coherence and distance*	----------none----------							

Table S3. Summary of significant activations for GLM1C (correct trials only) post-masking. All reported activations are significant at $p < .05$, FWE-corrected for multiple comparisons. Cluster-defining threshold: $p < .001$, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. pgACC: perigenual anterior cingulate cortex. vmPFC: ventromedial prefrontal cortex.

GLM1C: pre-masking

contrast	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	peak voxel MNI coordinates			
high > low coh	extrastriate (incl. MT+)	2050	4.89	<.001	44	-62	4	R
	striatum	618	4.81	<.001	-22	12	-10	L
	striatum	768	4.77	<.001	20	16	-4	R
	extrastriate (incl. MT+)	1307	4.23	<.001	-24	-88	20	L
	posterior parietal	278	4.01	.002	-24	-52	58	L
	posterior parietal	172	3.97	.027	30	-48	62	R
high < low coh	----------none----------							
high > low dis	pgACC	610	5.00	<.001	10	38	12	LR
	insula	229	4.57	.004	-30	20	-18	L
	insula	182	4.41	.012	32	14	-10	R
	posterior cingulate	293	4.38	<.001	0	-12	40	LR
	premotor	220	4.36	.004	-46	-18	52	L
high < low dis	----------none----------							
interaction +	pgACC	309	4.28	<.001	4	44	12	LR
interaction –	precuneus	520	4.41	<.001	10	-66	64	L

Table S4. Summary of significant activations for GLM1C (correct trials only) pre-masking. All reported activations are significant at $p < .05$, FWE-corrected for multiple comparisons. Cluster-defining threshold: $p < .001$, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. pgACC: perigenual anterior cingulate cortex.