TITLE

Private-public mappings in human prefrontal cortex

AUTHORS

Dan Bang[1,2,*], Sara Ershadmanesh[3], Hamed Nili[4] & Stephen M. Fleming[1,5,6]

AFFILIATIONS

[1] Wellcome Centre for Human Neuroimaging, UCL, London, UK
[2] Department of Experimental Psychology, University of Oxford, Oxford, UK
[3] School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran
[4] Wellcome Centre for Integrative Neuroimaging, Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford, Oxford, UK
[5] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, UCL, London, UK
[6] Department of Experimental Psychology, UCL, London, UK
* Correspondence: danbang.db@gmail.com

ABSTRACT

A core feature of human cognition is an ability to separate private states of mind – what we think or believe – from public actions – what we say or do. This ability is central to navigating social interactions in which different contexts often require different mappings between private states and public actions in order to minimise conflict and facilitate communication. Here we investigated how the human brain supports private-public mappings, using an interactive task which required subjects to adapt how they communicated their confidence about a perceptual decision to the social context. Univariate and multivariate analysis of fMRI data revealed that a private-public distinction is reflected in a medial-lateral division of prefrontal cortex: medial areas (dACC, pgACC) tracked a private sense of confidence, whereas lateral frontal pole (FPl) contextualised this variable for public report. The concept of private-public mappings provides a promising framework for understanding social dysfunction in neuropsychiatric conditions.

INTRODUCTION

A striking aspect of human social behaviour is how often we say things we do not really mean simply because the situation requires it. For example, a politician may have doubts about a policy but, in order to boost voters' faith in her, defend it with great confidence. A dinner guest may want a second piece of cake but, in order to avoid social disapproval, politely decline the offer. Central to these scenarios is a distinction between private states of mind – what we think or believe – and public actions – what we say or do. A general ability to separate private and public aspects of mental states is thought to have evolved under the constraints of social communication (Dennett, 2017). In having a 'control buffer' between private states and public actions, an agent may avoid revealing its current state to competitors (e.g., by inhibiting signs of fragility). It also allows an agent to deceive competitors (e.g., by displaying signs of strength) without interfering with processes that are needed for resolving its current state (e.g., taking steps to replenish energy).

In modern life, where people navigate complex social worlds, this ability to adapt private-public mappings according to the context takes on new functions and helps minimise social friction. For example, social norms pertaining to politeness vary dramatically between contexts, such that a response which is appropriate in one context may be very inappropriate in another (e.g., commenting on another person's appearance at home versus at work). Contextual modulation of private-public mappings is, however, particularly challenging: social norms may be arbitrary, social contexts are typically diffuse over time and space, and there is often a tension between the nature of our private states and the public actions required by the current context. Indeed, context-inappropriate social behaviour is a common symptom in a variety of neuropsychiatric conditions – such as frontotemporal dementia (Ibañez & Manes, 2012), autism (Happé, 1994), schizophrenia (Penn, Ritchie, Francis, Combs, & Martin, 2002) and borderline personality disorder (King-Casas, Sharp, Lomax-Bream, Lohrenz, Fonagy, & Montague, 2008) – at times with profound effects on quality of life.

Understanding why and how social function is impaired in these conditions is likely to be aided by a more thorough characterisation of the cognitive and neural mechanisms that underpin contextual modulation of private-public mappings. Here we studied how the brain solves this problem using decision confidence as a model system. Confidence is well-suited for studying private-public mappings because there is often a dissociation between the confidence that we privately feel and that which we publicly communicate (Aitchison, Bang, Bahrami, & Latham, 2015; Bang et al., 2017; Hertz, Palminteri, Brunetti, Olesen, Frith, & Bahrami, 2017). Further, recent experimental work has provided us with the tools to separately manipulate private and public dimensions of confidence (Bang et al., 2017; Bang & Fleming, 2018). In our experiment, subjects were required to communicate their confidence about simple perceptual decisions in different social contexts, with each context requiring a specific mapping from private to public confidence. For example, one context required subjects to overstate their confidence in order to maximise reward, whereas another required them to understate their confidence.

Preliminary evidence suggests that a private-public distinction for confidence is reflected in a medial-lateral division of prefrontal cortex (PFC). First, recent work has indicated that dorsal anterior cingulate cortex (dACC) (Fleming, Van Der Putten, & Daw, 2018) and perigenual anterior cingulate cortex (pgACC) (Bang & Fleming, 2018; De Martino, Fleming, Garrett, & Dolan, 2013; Lebreton, Abitbol, Daunizeau, & Pessiglione, 2015; Wittmann, Kolling, Faber, Scholl, Nelissen, & Rushworth, 2016) – both in the medial wall of PFC – support the formation of an internal (private) sense of confidence. Second, a large body of literature has observed activation of lateral PFC, in particular the lateral frontal pole (FPl), in relation to explicit (public) reports of confidence (De Martino et al., 2013; Fleming, Huijgen, & Dolan, 2012; Gherman & Philiastides, 2018; Hilgenstock, Weiss, & Witte, 2014; Shekhar & Rahnev, 2018; Yokoyama et al., 2010) – raising the possibility that lateral PFC supports a mapping between private and public confidence. This hypothesis fits with a broader role of lateral PFC in cognitive control

functions such as task switching and set shifting (Badre, 2008; Badre & Nee, 2018). It also fits with observations that injury to lateral PFC is associated with context-inappropriate social behaviour (Ibañez & Manes, 2012). Here, by combining univariate and multivariate analyses of fMRI data, we provide evidence for this division of labour, showing that while dACC and pgACC track the formation of an internal sense of confidence, FPI tracks the transformation of this variable into a context-dependent public report.

RESULTS

Experimental manipulation of private-public mappings

Subjects (n = 28) performed a social perceptual decision task, first in a behavioural session and subsequently during fMRI (**Figure 1**). On each trial, subjects made a group decision about a visual stimulus with one of four partners. Subjects were told that the partners were created by replaying the responses of four people performing the perceptual task on a separate day but, in reality, the partners were simulated. First, subjects judged whether a field of dots was moving left or right. Next, after being informed about the identity of their partner on the current trial, subjects were asked to report their confidence in the perceptual judgement – an estimate which would enter into the group decision. Subjects were then shown the current partner's response for that trial. Finally, implementing a common group decision rule (Bang et al., 2017), the individual decision made with higher confidence was automatically selected as the group decision, after which feedback about its accuracy was delivered. Subjects were incentivised to help each group achieve as many correct group decisions as possible but, by design, could only affect group decisions through their confidence reports.

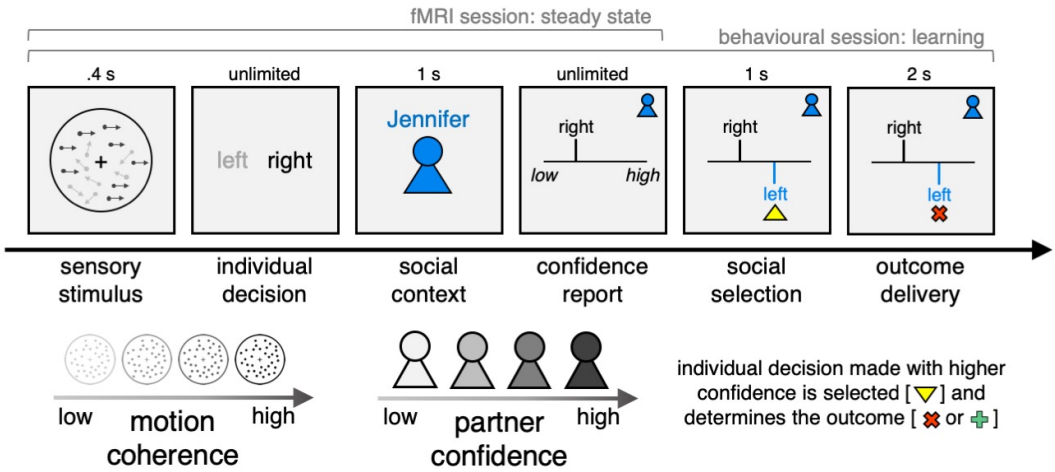


Figure 1. Experimental framework for dissociating private and public confidence. On each trial, subjects made a perceptual group decision with one of four partners. They first decided whether a random dot motion stimulus was moving left or right. We varied the fraction of coherently moving dots in order to manipulate subjects' internal sense of confidence in each decision. Subjects were then informed about their partner on the current trial and were asked to submit a report of confidence in their initial decision (discrete scale from 1 to 6). Subjects were then shown the partner's response, after which the individual decision made with higher confidence was selected as the group decision. Finally, subjects received feedback about choice accuracy, before continuing to the next trial. We engineered the partners to have the same level of performance as subjects but to differ in mean confidence. The structure of the task differed between the behavioural and the fMRI session as explained in the main text.

We varied two features of the task in a factorial (4 × 4) manner. First, we varied the fraction of coherently moving dots (coherence) to manipulate subjects' internal sense of confidence in a perceptual judgement. In general, the higher the coherence, the higher subjects' confidence.

Second, we engineered the partners (context) to have the same choice accuracy as subjects but to differ in mean confidence. In this case, the strategy for maximising group accuracy is to match each partner's mean confidence (Bang et al., 2017). If a subject always reports higher confidence than a partner, then they would not benefit from a partner's correct decisions. In contrast, if a subject always reports lower confidence than a partner, then they would not benefit from their own correct decisions. The design thus allowed us to separate public from private confidence and thereby probe private-public mappings in behaviour and brain activity.

The aim of the behavioural session was to calibrate levels of coherence so as to achieve target levels of choice accuracy and to train subjects on the social task (**Supplementary Figure 1**). The training had three phases. First, subjects were paired with different partners in a block-wise manner. Second, the partners were interleaved, with the identity of each partner revealed after the perceptual choice. Finally, using the same design as in the fMRI session, the 'showdown' stage was played out in the background, with subjects not seeing the partner's response or the group outcome. This change, which was introduced to minimise inter-trial dependencies in behavioural and neural responses, meant that subjects had to rely on their knowledge (expectations) about the partners learned in the preceding behavioural session. In addition, to establish a baseline for behavioural and neural responses, we added a condition where the partner's identity was hidden. The fMRI session consisted of four runs, with the distribution of conditions matched across runs in order to facilitate multivariate analysis of the fMRI data (i.e. 4 trials for each coherence x context condition in each run).

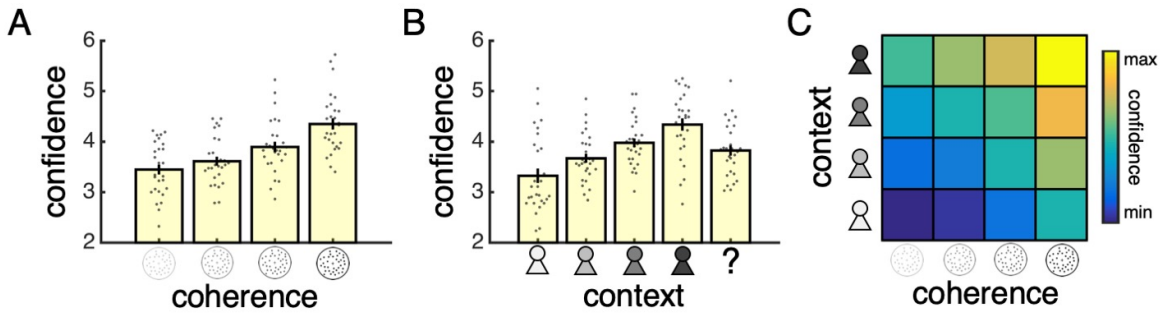


Figure 2. Separate influences of motion coherence and social context on confidence reports. **(A)** Mean confidence reported for each level of coherence. **(B)** Mean confidence reported when playing with each partner. The question mark indicates a condition where the partner's identity was not revealed. **(C)** Heat map visualising mean confidence in each condition of our factorial design aggregated across subjects (confidence was z-scored for each subject). Warmer colours indicate higher confidence. Data are from the fMRI session. In **(A)** and **(B)**, data are represented as group mean ± SEM. Each dot is a subject.

Confidence reports reflect private state and social context

We first tested whether subjects' confidence reports in the fMRI session varied as a function of our factorial design (see **Supplementary Figure 2** for analysis of the behavioural session). As intended, subjects' confidence reports were influenced by both coherence and context (**Figure 2**, ordinal regression; coherence: $t(27)=6.95$, $P<0.001$, context: $t(27)=4.82$, $P<0.001$, interaction: $t(27)=-0.03$, $P=0.975$). In particular, subjects' reported confidence increased with the level of coherence (**Figure 2A**) and the mean confidence of the current partner (**Figure 2B**). In other words, the confidence level reported for a specific level of coherence depended on the current partner's mean confidence (**Figure 2C**) – demonstrating that subjects flexibly adjusted the mapping from an internal sense of confidence in their decision to an explicit report of confidence according to the social context. Critically, subjects were not shown the partners' responses in the fMRI session and therefore had to rely on their knowledge about the partners acquired during the behavioural session.

Encoding of private state and social context in prefrontal cortex

We next turned to the fMRI data to identify neural areas which support a context-dependent mapping from private to public confidence. We focused on three prefrontal ROIs identified in the Introduction: dACC and pgACC, both putative neural substrates for a computation of an internal sense of confidence, and FPl, shown to track reports of confidence but whose role in coordinating private-public mappings has not been tested directly (**Figure 3A**). As an initial assessment of the contribution of these ROIs, we estimated a general linear model (GLM) of neural responses locked to the screen which revealed the partner. Using our factorial design, we included linear and quadratic terms for coherence and context as parametric regressors – with the quadratic context term capturing the intuition that larger behavioural deviations from a default policy (medium confidence) are required when playing with both low-confidence and high-confidence partners (**Figure 3B**).

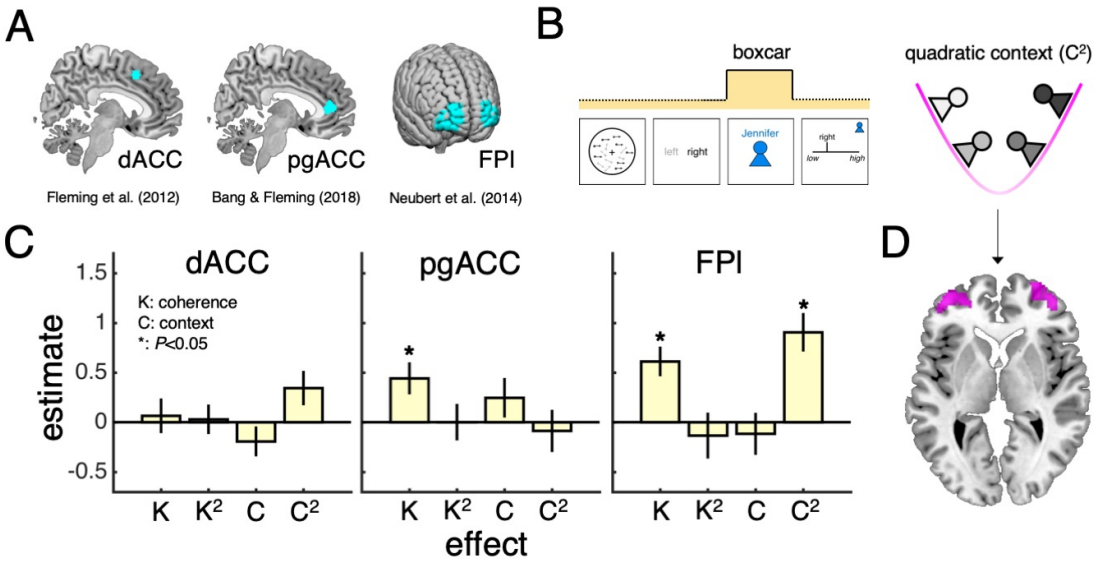


Figure 3. Encoding of private state and social context in lateral frontal pole. **(A)** Region of interest (ROI) masks. **(B)** We modelled neural responses to the context screen, including both linear and quadratic terms for coherence and context as parametric regressors – with a quadratic term for context capturing differential encoding of the low-confidence and high-confidence partners. **(C)** ROI contrast estimates for the effects of coherence (K), quadratic coherence (K^2), context (C) and quadratic context (C^2). We tested significance (asterisk) by comparing contrast estimates across subjects to zero (one-sample *t*-test). Data are represented as group mean ± SEM. **(D)** Visualisation of whole-brain activation for quadratic context in lateral prefrontal areas (clusters significant at $P<0.05$, FWE-corrected for multiple comparisons, cluster-defining threshold at $P<0.001$). Our analyses were ROI-based – motivated by the literature – but see **Supplementary Table 1** for whole-brain activations obtained under GLM. dACC: dorsal anterior cingulate cortex. pgACC: perigenual anterior cingulate cortex. FPl: lateral frontal pole.

Under this model, we would expect activity in neural areas which track private confidence to vary as a function of motion coherence, whereas neural areas which support a mapping from private to public confidence should also track information about the social context. Consistent with a role in private-public mappings, FPl tracked both the level of coherence and the social context (**Figure 3C**). In particular, FPl activity was higher for low-confidence and high-confidence partners – the contexts which required a larger deviation from a default policy for confidence reports – as indicated by an effect of quadratic context (**Figure 3D**). As for the ROIs hypothesised to support the formation of an internal sense of confidence, only pgACC tracked the level of coherence during presentation of the social context (**Figure 3C**). This result is in line with a previous fMRI study which found that pgACC tracked a combination of variables that underpinned changes in private confidence (Bang & Fleming, 2018).

Encoding of trial-by-trial behaviour in dorsal anterior cingulate cortex

We next turned to subjects' confidence reports, which, as shown in **Figure 2C**, reflect factors relating to both the perceptual decision and the social context. To disentangle the contribution of these factors to neural activity, we first used a model to estimate what subjects would have responded on a particular trial had there been no contextual modulation (**Figure 4A**; "what I would have said"). We then related both these model-derived estimates of private confidence ("what I would have said") and the empirically observed reports of public confidence ("what I said") to ROI activity. Of the three ROIs, dACC showed a response profile consistent with a sequential encoding of private and public confidence (**Figure 4B**) – a result which indicates that dACC is involved in the behavioural implementation of a particular private-public mapping.

We reasoned that, if FPl supports the contextualisation of a private state for public action, then connectivity between FPl and dACC/pgACC should vary with the contextual requirements of the task. In particular, connectivity should be higher for larger (absolute) differences between "what I said" and "what I would have said" had there been no contextual requirements. In line with this prediction, a psychophysiological interaction (PPI) analysis showed that connectivity between FPl and dACC increased with behavioural deviations from a default policy (**Figure 4C**). Taken together, these results indicate that dACC integrates a private confidence signal with a context signal from FPl during the formation of explicit confidence reports.

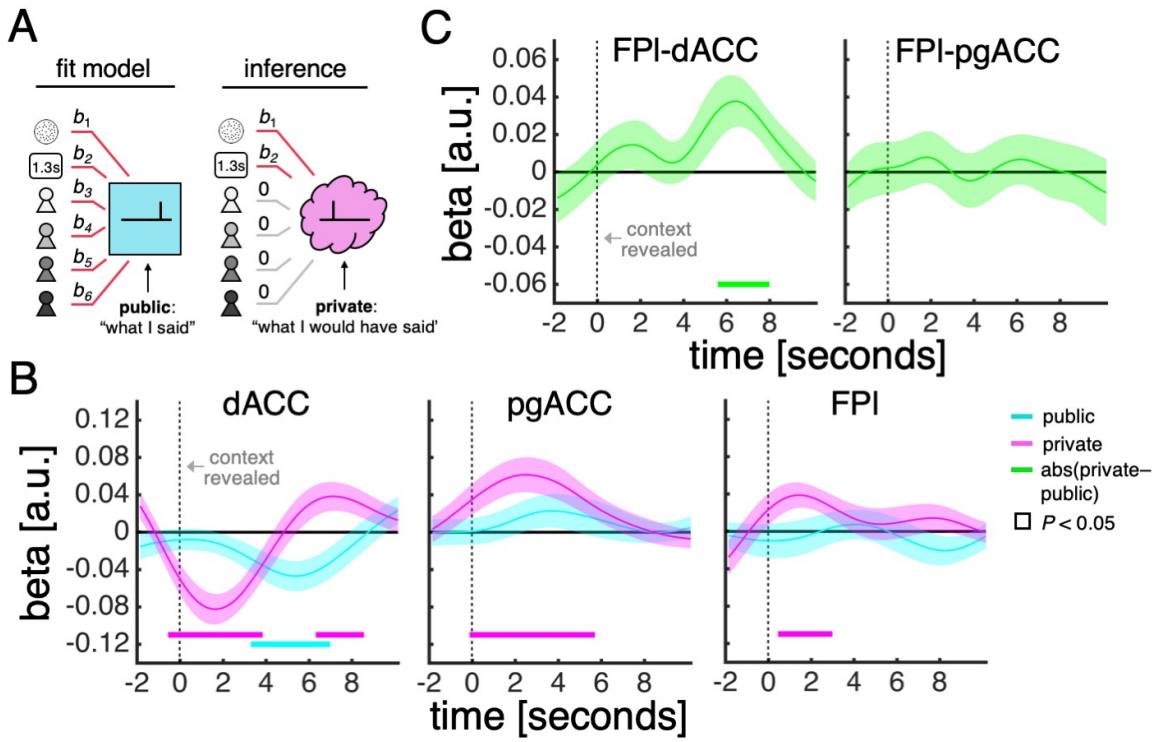


Figure 4. Trial-by-trial neural encoding of private and public confidence. **(A)** We fitted an ordinal probit regression model to a subject's confidence reports, including terms relating to the perceptual judgement (i.e. coherence and reaction time) and the social context (one term for each partner). We then used the fitted model to infer the level of confidence subject would have reported had there been no contextual requirements (a proxy for private confidence), by applying the fitted model to the subject's data, while setting the four social terms to zero. **(B)** GLM analysis of the effects of private confidence (i.e. model-derived estimate) and public confidence (i.e. empirically observed report) on ROI activity time courses. We used activity time courses as there is no task marker of when subjects internally compute a sense of confidence or select a response. **(C)** Psychophysiological interaction (PPI) analysis of ROI activity time courses. Traces are coefficients from a GLM in which we predicted dACC/pgACC activity from the interaction of FPl activity and behavioural deviation (i.e. absolute difference between private and public confidence), controlling for both main effects of FPl activity and behavioural deviation. In **(B)** and **(C)**,

vertical dashed lines indicate the onset of the context screen. We tested significance (coloured square) for each time point by comparing coefficients across subjects to zero (one-sample *t*-test).

Representation of task space in lateral frontal pole

Finally, we reasoned that, if FPl supports private-public mappings, then this area should also carry information about the various social situations engendered by our task. In computational terms, our task is comprised of 16 "states", corresponding to the different combinations of coherence and context (e.g., the state on the current trial may be "low coherence + high-confidence partner", whereas on the next trial a new combination of coherence and context will be encountered). By design, each of these 16 states requires subtly different behavioural responses in order to maximise reward – a pattern reflected by subjects' confidence reports (**Figure 2C**). If FPl indeed supports this contextual modulation of behavioural responses, then it should represent the different states of the task as distinct.

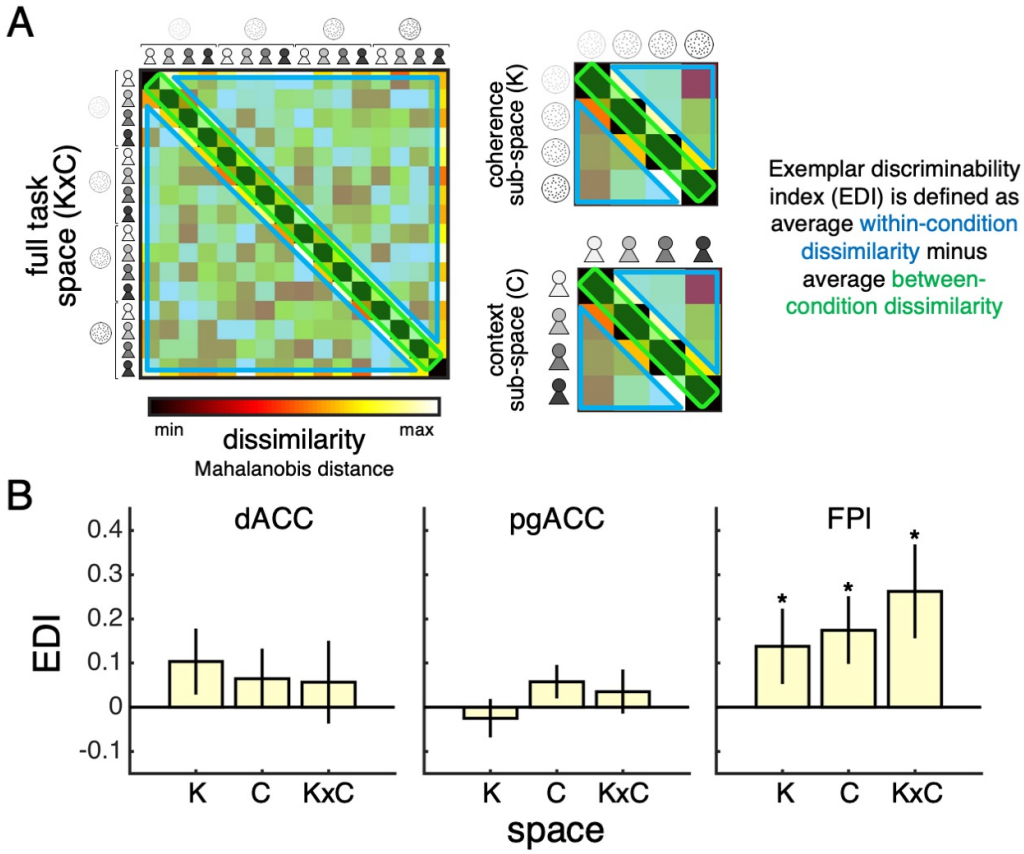


Figure 5. Task representations in lateral frontal pole. **(A)** A split-data representational dissimilarity matrix (sdRDM) is constructed by (1) computing the Mahalanobis distance between the voxel activity pattern in scan run *i* and the voxel activity pattern averaged across scan runs $j \neq i$ for every pair of conditions and (2) averaging the sdRDMs across scan runs. The exemplar discriminability index (EDI) is computed as the average dissimilarity across the off-diagonal elements (blue) minus the average dissimilarity across the diagonal elements (green). A positive EDI indicates that the voxel activity pattern within an ROI is more stable within conditions than between conditions and therefore that the ROI discriminates between the conditions. Data is simulated for visualisation. **(B)** ROI EDIs for the full task space (KxC) and the sub-spaces of coherence (K) and context (C). We tested significance (asterisk) by comparing EDIs across subjects to zero (one-tailed sign-rank test). Data are represented as group mean ± SEM. dACC: dorsal anterior cingulate cortex. pgACC: perigenual anterior cingulate cortex. FPl: lateral frontal pole.

We tested this prediction using representational similarity analysis (RSA) and a metric known as the exemplar discriminability index (EDI) (Nili, Walther, Alink, & Kriegeskorte, 2016). Like other multivariate methods, RSA considers the pattern of activity across voxels within an ROI rather than the mean activity across voxels. The EDI metric asks whether a multivariate pattern is more stable across scan runs *within* conditions than *between* conditions – the intuition being that higher within-condition stability shows that an ROI represents the conditions as distinct (**Figure 5A**). In support of a central role in coordinating private-public mappings, FPl carried a representation of the full task space as well as the sub-spaces of coherence and context – a pattern not found in dACC or pgACC (**Figure 5B**).

DISCUSSION

Contextual modulation of social behaviour – a core part of healthy social function – is facilitated by the ability to flexibly map private states onto public actions. Here we studied how the brain supports private-public mappings, using a social perceptual decision task in which subjects had to learn and apply context-dependent mappings from private to public confidence. Using univariate and multivariate analyses of fMRI data, we found that dACC and pgACC, in line with recent studies (Bang & Fleming, 2018; Fleming et al., 2018), tracked an internal sense of confidence, whereas FPl, through carrying a representation of the social situations induced by our task, supported the transformation of this variable into a context-dependent explicit report.

Our results are consistent with a role of lateral PFC in cognitive control, broadly defined as the ability to use task representations to guide thought and behaviour (Badre, 2008; Badre & Nee, 2018). In a typical study on cognitive control, the appropriate stimulus-response mapping on a particular trial depends on a contextual cue (e.g., respond '1' when stimulus is blue in context A and '2' in context B). The complexity of the control problem is then increased by having additional hierarchical rules (e.g., another cue indicating whether to respond to the colour or size of the stimulus). Both functional imaging and patient studies indicate that lateral PFC is required when control problems increase in complexity – for example, by providing a model of the current task abstracted across individual episodes or acquired through instruction (Badre & D'Esposito, 2007; Badre, Hoffman, Cooney, & D'Esposito, 2009). Our study suggests that similar cognitive and neural mechanisms are recruited to solve the private-public mapping problem inherent to social interaction. An open question is whether private-public mappings involve additional computations beyond those invoked by typical cognitive control paradigms, given that the 'stimulus' requiring a different response involves an intervening subjective state.

Our results also prompt a revaluation of the contribution of lateral PFC to metacognition – the ability to monitor and evaluate ongoing thought and behaviour. Several studies have shown that lateral PFC tracks explicit reports of confidence (De Martino et al., 2013; Fleming et al., 2012), and that the microstructure of lateral PFC predicts the degree to which an individual's confidence reports reflect their objective performance (Allen et al., 2017; Fleming, Weil, Nagy, Dolan, & Rees, 2010). These results are often taken to show that lateral PFC underpins the computation of an internal sense of confidence, or the 'read-out' of such a variable from circuits involved in decision-making. Our study supports and refines this latter hypothesis, by showing that FPl may contextualise an internal sense of confidence for explicit report in accordance with task requirements. On this account, a relationship between prefrontal structure or function and individual differences in metacognition may not reflect a contribution to insight per se but, rather, the ability to maintain a stable private-public mapping.

Recent years has seen an interest in how the brain encodes structured representations of the world for efficient learning and decision-making (Behrens et al., 2018; Niv, 2019; Radulescu, Niv, & Ballard, 2019). In these scenarios, context may be particularly important for adaptive behaviour. In psychology, context is defined as the set of circumstances around an event (e.g., we may hear a ringing phone in different environments such as our home or a friend's house)

(Gershman, 2017; Maren, Phan, & Liberzon, 2013). Representing context is useful because it may carry information about what to expect or what to do (e.g., we should answer a ringing phone at home but not at a friend's house) (Gershman, 2017; Maren et al., 2013). Context may also hold structural information that can be generalised across events (e.g., many behaviours which are acceptable at home are not acceptable when visiting others) (Gershman, 2017). There is evidence that such structural knowledge is encoded in the medial temporal lobe and ventral PFC (Behrens et al., 2018; Niv, 2019). Our results complement this work by identifying representations in FPl that may support a mapping between private states and public actions that are appropriate for the current context.

Our study identifies several questions for future research. First, how are context-dependent private-public mappings learned? Model-free reinforcement learning (e.g., Q-learning) (Sutton & Barto, 1998) may be able to learn the value of taking a public action given a private state and a social context but is likely to be too memory intensive (e.g., storing Q-values for every state-action pair in every context) and too slow (e.g., having to experience every state-action pair in every context) without model-based components (e.g., using a model of the world to generalise across state-action pairs or across contexts based on their similarity) (Gershman, 2017). Second, what happens when there is uncertainty about the current context? In our task, the context was explicitly signalled. However, in many situations, the context must itself be inferred. At a neural level, context uncertainty is likely to require the involvement of additional areas, such as the hippocampus which is hypothesised support inference and generalisation based on structural knowledge (Duncan, Doll, Daw, & Shohamy, 2018; Eichenbaum, 2017a, 2017b). Finally, what is the nature of the PFC task representation identified by our study? In our task, each social context is associated with a distinct private-public mapping. It is therefore hard to tell whether PFC discriminates between social contexts or a more abstract construct such as behavioural policies. This question could be addressed by developing a task in which different social contexts require subjects to adopt the same private-public mapping.

The current conceptualisation of private-public mappings may offer insight into the multiple routes to social dysfunction across a range of neuropsychiatric conditions. Broadly, contextual modulation of social behaviour requires at least three distinct sets of computation – (1) context inference, (2) action selection and (3) learning from outcomes – each presumably supported by distinct neural substrates. In future it may be possible to distinguish different aspects of social dysfunction along each of these dimensions. For example, context-inappropriate social behaviour may arise because an individual cannot identify the current context (e.g., due to hippocampal damage), cannot inhibit prepotent but context-inappropriate actions (e.g., due to prefrontal damage), or does not make correct inferences from past experience (e.g., due to disturbance in neuromodulatory systems supporting learning). Our study provides a starting point for developing such computational-level characterisations of social dysfunction.

METHODS

Subject details

Twenty-eight adults (14 females; mean age ± SD, 26.36 ± 7.00) participated in the study. 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 (£45, £10/hour) and could earn an additional performance-based bonus (see below).

Experimental details

Task and procedure

Subjects performed a social perceptual decision task in separate prescan and scan sessions (**Supplementary Figure 1**). We made modifications to the task between the prescan session and the scan session. Here we first describe the full task, before detailing the changes made.

Each trial began with the presentation of a fixation cross at the centre of a circular aperture. After a uniformly sampled delay (prescan: .5-1s; scan: 1-4s), subjects viewed a field of moving dots inside the aperture (.4s). Once the stimulus terminated, subjects had to press one of two buttons to indicate whether the average direction of dot motion was left or right. Once a choice had been registered, the fixation cross turned grey (.5s). Subjects were then presented with a screen informing them about their partner on the current trial (1s). There were four partners in total. Each partner was indicated by a unique colour and name. Subjects were told that the partners were created by replaying the responses of people performing the perceptual task on another day but, in reality, all partners were simulated.

Next, subjects had to indicate their confidence in the perceptual decision, 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). Subjects were then shown the partner's response on the corresponding trial (1s), before the individual decision made with higher confidence was selected by the computer as the group decision, highlighted by a yellow triangle (2s). Finally, subjects received feedback about the accuracy of the group decision (2s), before continuing to the next trial. The feedback was indicated by replacing the yellow triangle with a green plus sign if the group decision was correct and a red cross if incorrect. Subjects were incentivised to help each group achieve as many correct group decisions as possible: we randomly selected two trials for each group in each session (4 x 2 x 2 = 16 trials) and paid £1 in bonus for every correct group decision (in reality, all subjects were paid £10 in bonus).

We varied two features of the task in a factorial (4 × 4) manner. First, we varied the fraction of coherently moving dots (coherence) to manipulate subjects' internal sense of confidence in a perceptual judgement (see stimulus calibration). Second, we specified the partners (context) such that they had the same choice accuracy as subjects but differed in mean confidence (see simulation of partners).

The behavioural session involved 4 phases. In phase 1, we calibrated the coherence levels so as to achieve target levels of choice accuracy (see stimulus calibration). In phases 2-4, we trained subjects on the social task. In phase 2, subjects were paired with the four partners in a block-wise manner. In particular, there were 4 cycles of blocks of 10 trials per partner (e.g., A-B-C-D-A-B-C-D-…; 4 x 4 x 10 = 160 trials). Information about the partner was shown before each block of 10 trials but not after the perceptual decision. In phase 3, subjects were paired with the four partners in an interleaved manner, with the identity of the partner revealed after a perceptual decision had been made (4 x 40 = 160 trials). In phase 4, the group decision was played out in the background, with the next trial starting after subjects had indicated their confidence. In addition, we introduced a new condition where the social context was hidden

(5 x 40 = 200 trials). In each phase, the coherence levels were counterbalanced across trials within a social context, so that each coherence level was experienced the same number of times for each partner. To help subjects keep track of the behaviour of each group, they were informed every 40 trials how often their individual decision had been selected as the group decision for a partner (15s). The statistics were reset after each phase.

The scan session involved four runs, using the same task design as in phase 4 of the prescan session. We matched the distribution of conditions (coherence × context) across runs in order to facilitate multivariate analysis of the fMRI data, with 4 repetitions per condition (4 x 4 x 5 = 80 trials per run). The screen informing subjects about how often their individual decision had been selected as the group decision for each partner was shown every 20 trials. The statistics were reset after each run.

Random dot kinematograms

Subjects viewed random dot kinematograms (RDKs) contained in a circular white aperture (7 degrees in diameter). Each RDK was made up of three independent sets of dots (each dot was 0.12 degrees in diameter) shown in consecutive frames. Each set of dots were shown for one frame (about 16ms) and then replotted again three frames later (about 50ms). Each time a set of dots was replotted, a subset of the dots, determined by the motion coherence, k, was displaced in the direction of motion at a speed of 5 degrees s^{-1}, whereas the rest of the dots were displaced at random locations within the aperture. The motion direction was to the left or the right along the horizontal meridian. The dot density was fixed at 30 dots $degrees^{-2}$ s^{-1}. To help subjects maintain fixation, a circular region (0.7 degrees in diameter) at the centre of the aperture was kept free of dots. A set of coherence levels, $\mathbf{K}$, was identified for each subject in a separate stimulus calibration session.

Stimulus calibration

Subjects performed the perceptual part of the task in two blocks. In block 1, we deployed a set of prespecified coherence levels, $\mathbf{K}$: $\{.03, .06, .12, .24, .48\}$. Each coherence level was used 20 times for each direction (i.e. 5 x 20 x 2 = 200 trials). We then fitted a simple signal detection theory model with a single noise parameter, σ, governing the statistical relationship between coherence and choice (see simulation of partners for details on model). We selected the noise parameter which minimised the sum of squared errors between predicted and observed choice accuracy across coherence levels. In block 2, we used the fitted noise parameter to select a set of coherence levels associated with the target choice accuracies 60%, 70%, 80% and 90%, $\mathbf{K}$: $\{k_{.6}, k_{.7}, k_{.8}, k_{.9}\}$. Each coherence level was then used 25 times for each direction (i.e. 4 x 25 x 2 = 200 trials). We repeated the fitting procedure and selected a final set of coherence levels for the main task (the final fitted noise parameter was in turn used to simulate the partners).

Simulation of partners

We used a signal detection model to simulate the partners' choices and confidence reports in phases 2 and 3 of the prescan session (Bang et al., 2017). In this model, an agent receives noisy sensory evidence, x, sampled from a Gaussian distribution, $x \in N(dk, \sigma)$, and makes a choice by comparing the sensory evidence to zero, choosing left if $x < 0$ and right if $x > 0$. The mean of the Gaussian distribution is given by coherence, k, and direction, d, with $d = -1$ indicating left and $d = -1$ indicating right. The standard deviation, σ, is the level of sensory noise – specified by fitting the model to a subject's choices in the stimulus calibration session (see stimulus calibration). The agent computes an internal estimate of decision confidence, z, using the absolute value of the evidence strength, $z = |x|$ – a quantity which is monotonically related to the probability that the perceptual choice is correct given the sensory evidence and the level of sensory noise. Finally, the agent maps this internal estimate onto a confidence

report, r, by applying a set of thresholds, $r = f(z)$. Precise control over the number of times that the agent selects a particular confidence report is achieved by first simulating a vector of x's – using the known sequence of stimuli in the task – and then setting the thresholds in z-space so as to achieve a desired confidence distribution. In this way, we created four partners with low, medium-low, medium-high and high mean confidence (see **Supplementary Figure 3** for distributions). We simulated phases 2 and 3 separately, so that a particular partner had the same confidence distribution in each phase. In addition, we simulated x's for each partner in each session under the constraint that their choice accuracy was within 1% of the target choice accuracy for each coherence level (see stimulus calibration). In phase 4, and in the scan session, we did not simulate responses. To calculate how often a subject's individual decision had been selected as the group decision, we first computed the probability of a subject's decision being selected given their confidence report and the partner's confidence distribution on a given trial and then averaged these probabilities across trials.

Behavioural analysis

Regression analysis

We used ordinal regression to predict subjects' trial-by-trial confidence reports (**Figure 2**). The model included (contrast-coded) coherence and context as predictors of interest and (log-transformed) choice reaction time, choice, motion direction and marker starting position as predictors of no interest. We z-scored all variables before analysis. We excluded the trials in which the partner's identity was hidden. We performed a separate regression for each subject and tested the group-level significance of a predictor by comparing the coefficients across subjects to 0 (one-sample *t*-test).

Confidence model

We used a previously established approach (Bang & Fleming, 2018) to construct a model of trial-by-trial subjective confidence for fMRI analysis (**Figure 4**). We fitted an ordinal regression model with six predictors to a subject's confidence reports: (1) z-scored, contrast-coded coherence, (2) z-scored, log-transformed choice reaction time and (3-6) a dummy variable for each explicitly signalled context. We then applied the model to the data, setting the fitted context coefficients to zero. This approach yields a prediction about the level of confidence that would have been reported on each trial in the absence of contextual (i.e. partner-specific) modulation. The prediction is a probability distribution over possible responses (e.g., '1' has a 10% probability, '2' has a 20% probability and so on). We used the expectation over possible responses as our estimate of subjective confidence.

fMRI analysis

Acquisition

MRI data were acquired on a 3T Siemens Prisma scanner with a 64-channel head coil. T1-weighted structural images were acquired using a 3D MPRAGE sequence: 1x1x1 mm resolution voxels; 176 sagittal slices, 256x224 matrix; TR = 2530ms; TE = 3.34ms; TI = 1100ms. 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 mm^2 with 3x3 mm^2 resolution.

Preprocessing

MRI data were pre-processed using SPM12 (Wellcome Trust, London). The first 4 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 (Andersson, Hutton, Ashburner, Turner, & Friston, 2001). 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 (Ashburner & Friston, 2005). 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. The motion correction parameters estimated from the realignment procedure and their first temporal derivatives – 12 'motion' regressors in total – were included as confounds in the first-level analysis for each subject.

Physiological monitoring

Peripheral measurements of a subject's 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 a 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 (Hutton et al., 2011). Models for cardiac and respiratory phase and their aliased harmonics were based on RETROICOR (Glover, Li, & Ress, 2000) and a similar, earlier method (Josephs, Howseman, Friston, & Turner, 1997). 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 (Birn, Diamond, Smith, & Bandettini, 2006; Birn, Smith, Jones, & Bandettini, 2008) and heart rate (Chang & Glover, 2009). This procedure yielded a total of 14 'biophysical' regressors which were sampled at a reference slice in each image volume. The regressors were included as confounds in the first-level analysis for each subject.

Regions of interest

We focused on three a priori ROIs highlighted by previous research on decision confidence. The dACC mask was an 8-mm sphere around the peak coordinates (MNI coordinates [*x y z*] = [0 17 46]) identified in a previous study of confidence (Fleming et al., 2012). The pgACC mask was defined using the second-level *t*-map (coherence x distance; cluster-defining threshold, $P<0.001$) from Bang & Fleming (Bang & Fleming, 2018). The FPl mask was defined using the right-hemisphere atlas developed by Neubert et al. (Neubert, Mars, Thomas, Sallet, & Rushworth, 2014) and mirrored to the left hemisphere to create a bilateral mask.

Contrast estimates

Contrast estimates for the different factors of our task design (**Figure 3**) were obtained using SPM12 (Wellcome Trust, London). We constructed an event-related GLM (GLM1) with three condition regressors. First, the context screen when the partner was signalled (explicit, 1s boxcar). Second, the context screen when the partner was hidden (implicit, 1s boxcar). Third, the update screen informing subjects how often their decision had been selected as the group decision for each partner (update, 15s boxcar). We then parametrically modulated the explicit condition regressor using our task factors: (1) K, contrast-coded coherence, {-1.5,-.5,.5,1.5};

(2) K^2, contrast-coded coherence squared; (3) C, contrast-coded context, {-1.5,-.5,.5,1.5}; (4) C^2, contrast-coded context squared. Parametric modulators were not orthogonalized. We excluded trials in which subjects' choice reaction times were 2.5 SD below or above their grand mean reaction time within a scan run (0-2 trials per subject per run). In addition to the condition regressors, we added motion and biophysical parameters 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 cutoff) was applied to remove low-frequency drifts. We then extracted mean contrast estimates for our parametric modulators within ROIs from first-level contrast images and assessed group-level significance by applying one-sample *t*-tests against zero to extracted contrast estimates.

Activity time courses

We used activity time courses to study the neural encoding of trial-by-trial confidence and assess functional coupling between ROIs (**Figure 4**). We transformed each ROI mask from MNI to native space and extracted preprocessed BOLD time courses as the average of voxels within each mask. For each scan run, we regressed out variation due to head motion, applied a high-pass filter (128 s cut-off) to remove low-frequency drifts, and oversampled the BOLD time course by a factor of ~23 (time resolution of .144s). For each trial, we extracted activity estimates in a 12s window time-locked to our event of interest. We applied a linear regression to each time point and then, by concatenating beta-weights across time points, created a time course for each predictor of a regression model. We performed this step separately for each subject and pooled beta time courses across subjects for analysis**.**

Representational similarity analysis

Representational similarity analysis (RSA) of task representations (**Figure 5**) was performed using SPM12 (Wellcome Trust, London) and the RSA toolbox (Nili, Wingfield, Walther, Su, Marslen-Wilson, & Kriegeskorte, 2014). To estimate voxel activity patterns, we constructed an event-related GLM (GLM2) with a condition regressor locked to the context screen (1s boxcar) for each coherence × (explicit) context (i.e. 4 x 4 = 16 regressors per scan run). As for GLM1, we included implicit and update regressors as well as motion and biophysiological regressors as additional nuisance variables. 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 cutoff) was applied to remove low-frequency drifts. Finally, we extracted, separately for each scan run, a vector of voxel activity estimates for each task condition within an ROI from unsmoothed first-level beta images.

The exemplar discriminability index (EDI) is a test of whether an ROI carries information about a set of conditions – in our case the 16 conditions of the coherence × context task space. The hypothesis is that, if an ROI represents the conditions as distinct, then the ROI voxel activity pattern should be more stable across scan runs *within* conditions than *between* conditions. This hypothesis is tested by first constructing a split-data representational dissimilarity matrix (sdRDM) for each scan run, which contains the Mahalanobis distance (i.e. multivariate noise normalisation and Euclidean distance) between the voxel activity pattern in scan run i and the voxel activity pattern averaged across scan runs $j \neq i$ for all condition pairs, and averaging the run-specific sdRDMs across scan runs. An EDI metric can then be computed as the average dissimilarity across off-diagonal elements (blue in **Figure 5A**) minus the average dissimilarity across diagonal elements (green in **Figure 5A**). A positive EDI indicates that the voxel activity pattern in an ROI is more stable within conditions than between conditions and therefore that the ROI carries information about the conditions. As the EDI should theoretically not be below zero, we assessed group-level significance using a one-tailed sign-rank test.

DATA AND CODE

Anonymized behavioural and neural data and code supporting main analyses are available on GitHub: [*add link before publication*].

AUTHOR CONTRIBUTIONS

DB, HN and SMF conceived the study. DB designed the task. DB and SE collected the data. DB analysed the data under supervision of HN and SMF. DB, SE, HN and SMF interpreted the results. DB wrote the manuscript. HN and SMF provided critical comments. All authors approved the final manuscript.

ACKNOWLEDGEMENTS

The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). DB is supported by a Sir Henry Wellcome Postdoctoral Fellowship funded by the Wellcome Trust (213630/Z/18/Z). SMF is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206648/Z/17/Z).

REFERENCES

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10).

Allen, M., Glen, J. C., Müllensiefen, D., Schwarzkopf, D. S., Fardo, F., Frank, D., Callaghan, M. F., & Rees, G. (2017). Metacognitive ability correlates with hippocampal and prefrontal microstructure. *NeuroImage*, *149*(May 2016), 415–423.

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage*, *13*(5), 903–919.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200.

Badre, D., & D'Esposito, M. (2007). FMRI evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*, 1–18.

Badre, D., Hoffman, J., Cooney, J. W., & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nature Neuroscience*, *12*(4), 515–522.

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188.

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6).

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organising knowledge for flexible behaviour. *BioRxiv*, 365593.

Birn, R. M., Diamond, J. B., Smith, M. A., & Bandettini, P. A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *NeuroImage*, *31*(4), 1536–1548.

Birn, R. M., Smith, M. A., Jones, T. B., & Bandettini, P. A. (2008). The Respiration Response Function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *NeuroImage*, *40*(2), 644–654.

Chang, C., & Glover, G. H. (2009). Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. *NeuroImage*, *47*(4), 1448–1459.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds.* New York, NY: W. W. Norton.

Duncan, K., Doll, B. B., Daw, N. D., & Shohamy, D. (2018). More than the sum of its parts: a role for the hippocampus in configural reinforcement learning. *Neuron*, *98*(3), 645-657.e6.

Eichenbaum, H. (2017a). Memory: organization and control. *Annual Review of Psychology*, *68*(1), 19–45.

Eichenbaum, H. (2017b). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117–6125.

Fleming, S. M., Van Der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617–624.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relatiing Introspective Accuracy to Individual Differences in Brain Structure. *Science*, *1541*(September 2010), 1541–1544.

Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic Bulletin and Review*, *24*(2), 557–565.

Gherman, S., & Philiastides, M. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *ELife*, *7*, 1–28.

Glover, G. H., Li, T.-Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine*, *44*(1), 162–167.

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154.

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, *8*(1), 1–12.

Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd better think twice: post-decision perceptual confidence. *NeuroImage*, *99*, 323–331.

Hutton, C., Josephs, O., Stadler, J., Featherstone, E., Reid, A., Speck, O., Bernarding, J., & Weiskopf, N. (2011). The impact of physiological noise correction on fMRI at 7T. *NeuroImage*, *57*(1), 101–112.

Ibañez, A., & Manes, F. (2012). Contextual social cognition and the behavioral variant of frontotemporal dementia. *Neurology*, *78*(17), 1354–1362.

Josephs, O., Howseman, A. M., Friston, K. J., & Turner, R. (1997). Physiological noise modelling for multi-slice EPI fMRI using SPM. In *Proceedings of the 5th Annual Meeting of ISMRM* (p. 1682). Vancouver, Canada.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, *321*(5890), 806–810.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167.

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience*. Nature Publishing Group.

Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J., & Rushworth, M. F. S. (2014). Comparison of Human Ventral Frontal Cortex Areas for Cognitive Control and Language with Areas in Monkey Frontal Cortex. *Neuron*, *81*(3), 700–713.

Nili, H., Walther, A., Alink, A., & Kriegeskorte, N. (2016). Inferring exemplar discriminability in brain representations. *Manuscript in Preparation*, 080580.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4).

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553.

Penn, D. L., Ritchie, M., Francis, J., Combs, D., & Martin, J. (2002). Social perception in schizophrenia: the role of context. *Psychiatry Research*, *109*(2), 149–159.

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: the role of structure and attention. *Trends in Cognitive Sciences*, *23*(4), 278–292.

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078–5087.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. F. S. (2016). Self-other mergence in the frontal cortex during cooperation and competition. *Neuron*, *91*(2), 482–493.

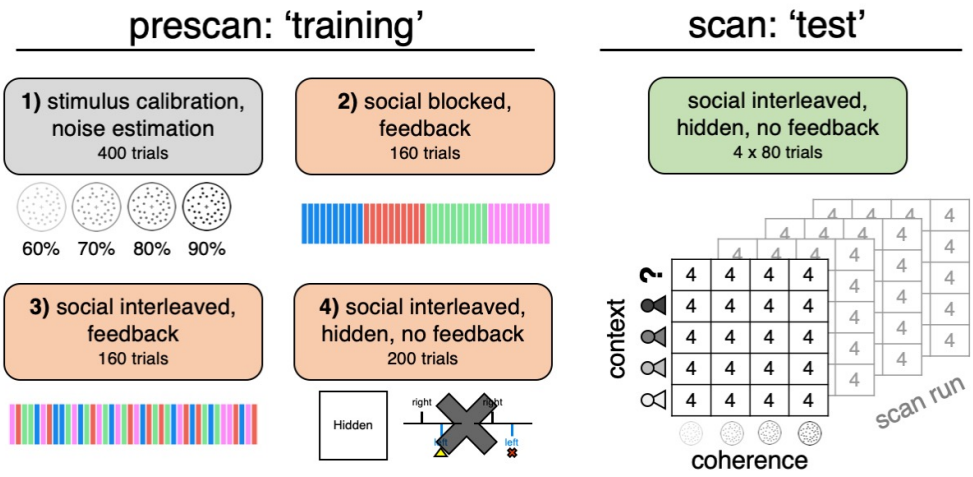
Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., Horie, K., Sato, S., Kawashima, R., & Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience Research*, *68*(3), 199–206.

Supplementary Materials

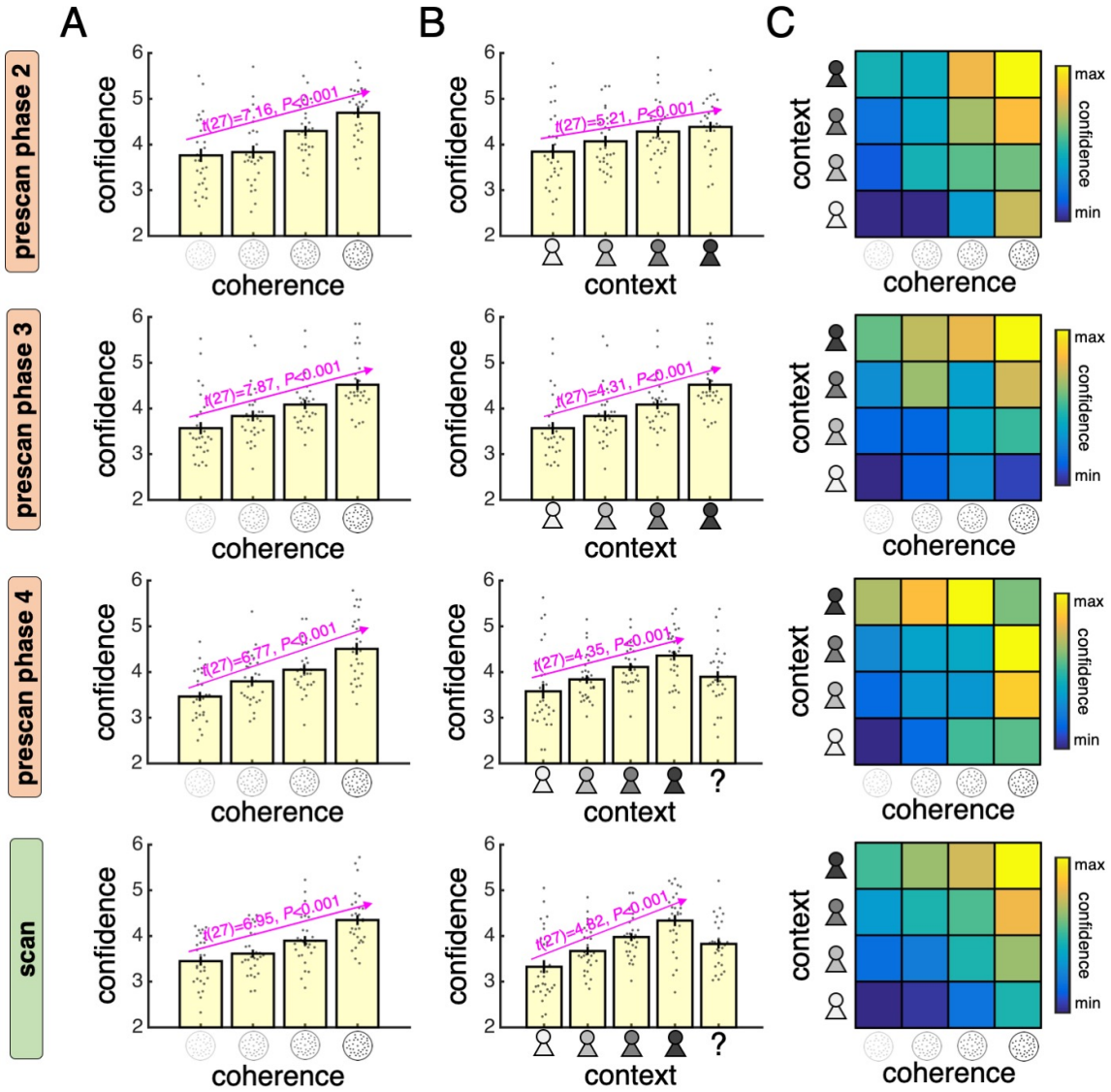
Private-public mappings in human prefrontal cortex

Dan Bang*, Sara Ershadmanesh, Hamed Nili and Stephen M. Fleming

*Correspondence: danbang.db@gmail.com



Supplementary Figure 1. Study procedure. Subjects participated in separate behavioural and fMRI sessions on the same day. The prescan session involved 4 phases. In phase 1, we calibrated the levels of coherence so as to achieve target levels of choice accuracy. In phases 2-4, we trained subjects on the social task. In phase 2, subjects were paired with the partners in a block-wise manner (here each partner is indicated by a unique colour). There were 4 cycles of blocks of 10 trials per partner (e.g., A-B-C-D-A-B-C-D-A-B-C-D-A-B-C-D). The context screen was shown before each block of 10 trials but not after a perceptual decision. In phase 3, subjects were paired with the four partners in an interleaved manner, with the partner's identity revealed after each perceptual decision. In phase 4, the 'showdown' stage was played out in the background. In addition, we introduced a condition where the social context was hidden. The scan session involved four runs, using the same design as in phase 4 of the prescan session. We matched the distribution of conditions (coherence × context) across scan runs.



Supplementary Figure 2. Analysis of behaviour in prescan and scan sessions. Rows 1-3 correspond to phases 2-4 of the behavioural session. Row 4 corresponds to the fMRI session and is included for comparison. **(A)** Mean confidence reported for each level of coherence. **(B)** Mean confidence reported when playing with each partner. The question mark indicates a condition in which the partner's identity was hidden. **(C)** Heat map over mean confidence in each condition of our factorial design aggregated across subjects (confidence was z-scored separately for each subject before aggregation). Warmer colours indicate higher confidence. In **(A)** and **(B)**, data are represented as group mean ± SEM. Each dot is a subject. Statistical testing (see pink text) was performed using ordinal regression as described in the Methods (we excluded trials in which the partner's identity was hidden).

Supplementary Figure 3. Confidence distributions used to generate four partners who differ in mean confidence. The distributions were constructed so as to have maximum entropy for a given level of mean confidence (see Bang et al., 2017, for details on calculation).

contrast	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	peak voxel MNI coordinates			
K: positive	striatum	1021	5.48	<.001	-20	8	10	L
	striatum	852	5.18	<.001	18	14	0	R
	posterior cingulate	419	4.49	<.001	-4	-34	32	LR
	cerebellum	479	4.45	<.001	-28	-62	-40	L
	cerebellum	281	4.41	.003	22	-72	-28	R
	parietal	196	4.30	.019	-36	-60	62	L
	parietal	462	4.11	<.001	42	-52	44	R
K: negative	----------							
K^2: positive	----------							
K^2: negative								
C: positive	occipital	222	5.10	.015	20	-80	-10	R
C: negative	occipital	277	3.71	.005	-6	-84	-14	L
C^2: positive	parietal (incl. TPJ)	1668	5.30	<.001	-40	-58	36	L
	parietal (incl. TPJ)	1656	4.82	<.001	50	-42	44	R
	lateral frontal (incl. FPl)	465	4.49	<.001	-38	56	8	L
	posterior cingulate	388	4.44	.001	-2	-24	34	LR
	dorsal medial frontal (incl. dmPFC)	692	4.29	<.001	4	36	32	LR
	dorsolateral frontal	330	4.25	.002	36	32	40	R
	lateral frontal (incl. FPl)	598	4.05	<.001	26	60	-8	R
	precuneus	366	3.90	.001	10	-70	44	R
C^2: negative	amygdala	196	4.54	.029	18	-8	-20	R

Supplementary Table 1. Summary of significant whole-brain activations for GLM1, using a 1s boxcar locked to the onset of the context screen and parametric modulators as specified in table. All reported activations are significant at $P<.05$, FWE-corrected for multiple comparisons. Cluster-defining threshold: $P<0.001$, uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. K: coherence. C: context. ^{2}: squared.