Mnemonic introspection in macaques is dependent on dorsolateral prefrontal but not orbitofrontal cortex

Sze Chai Kwok [1,2,3] *, Yudian Cai [1] and Mark J. Buckley [4]

[1] *Shanghai Key Laboratory of Brain Functional Genomics, Key Laboratory of Brain Functional Genomics Ministry of Education, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China*

[2] *Shanghai Key Laboratory of Magnetic Resonance, East China Normal University, Shanghai 200062, China*

[3] *NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai 200062, China*

[4] *Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom*

* To whom correspondence may be addressed. Email: sze-chai.kwok@st-hughs.oxon.org

ABSTRACT

Metacognition refers to the ability to be aware of one's own cognition. The anterior prefrontal cortex has been associated with meta-perceptual but not meta-memory decisions. A recent study has challenged this notion showing that neural activation in macaques' prefrontal areas 9 and 9/46d is associated with metamemory of recognition of items. Here, we verified the critical role of sub-regions of prefrontal cortex in the domain of spatial recognition memory. We contrasted performance of monkeys with superior dorsolateral prefrontal lesion with orbitofrontal lesioned monkeys and unoperated controls in spatial recognition memory tasks. We show that monkeys with dorsolateral lesions are impaired in meta-accuracy, but not in recognition performance, in comparison to orbitofrontal-lesioned and control monkeys. Together with the observation that the same orbitofrontal-lesioned monkeys were impaired in updating rule-value in a Wisconsin Card Sorting Test analog, we provide causal evidence towards functional specialisation between dorsolateral and ventromedial prefrontal cortices underpinning the introspection ability in relation to memory recognition in primates.

Metacognition refers to the ability to be aware of one's own cognition (e.g., knowledge of one's accuracy or knowing what one does or does not know). The anterior prefrontal cortex (aPFC) in the human brain is associated with this ability when the meta-judgement is based on perceptual but not on memory decisions [1]. This belief is supported by evidence coming from functional neuroimaging [2,3], structural neuroimaging measures such as white matter fiber tracking [4], microstructural measures of white-matter concentration [5], grey matter volume [6], and neuropsychological testing [7,8]. Nonetheless, a recent study showed that neural activation in specific regions within aPFC in the nonhuman primate brain is functionally associated with metamemory of recognition [9], implying that primate aPFC subserves metacognition for memory above and beyond perceptual processing. Using pharmacological intervention, Miyamoto et al. provided a partial parcellation of the aPFC by delineating two parallel neural streams supporting metamemory, one for temporally remote items in prefrontal area 9 (or 9/46d) versus another for more recent items in area 6, further corroborating the notion that first-order tasks (e.g., memory recognition) and second-order tasks (i.e., metacognitive processes such as meta-recognition) are dissociable [9]. It follows that in theory we could induce deficits in metamemory of recognition by lesioning these dorsolateral prefrontal (dlPFC) regions (lateral area 9) without changing other aspects such as type 1 memory task performance [10].

Furthermore, a recent neuroimaging study revealed that metacognitive control processes are neurally housed in anterior lateral (frontopolar) regions in humans, whereas other metacognitive processes underlying decision-making per se are independently computed by dissociable neural systems located more posteriorly to the frontopolar cortex [11]. One of the known hubs for decision-making within PFC is the orbitofrontal cortex (OFC). The OFC supports decision-making through updating changes in value [12,13] and inferring the consequences of potential behavior [14]. It has been proposed that decision confidence is represented in the OFC [15] such that lesions to the OFC would change the behavioral manifestation of decision confidence without affecting first-order task performance [16]. Since confidence estimation is a fundamental component of decision-making, and the OFC has been implicated in goal directed decisions that require the evaluation of predicted outcomes [17], such

representation of decision confidence signals in the OFC suggests the OFC might be causally required to support the computation of some of the elements of metacognition dissociable from those computed by antero-lateral PFC [15,16].

Taking the above into account, and in light of the findings that second-order metacognitive processes could be separated from confidence per se [18], we therefore set out to verify the causal roles of these two subregions (dlPFC vs. OFC) of the aPFC in memory using two variants of a spatial recognition memory paradigm. We hypothesized that the dlPFC is causally required for accurate memory introspection. Specifically, we contrasted the first-order memory and second-order metamemory performances of monkeys with superior dorsolateral PFC (sdlPFC) lesion (i.e., lateral area 9) (n=3) or with OFC-lesioned monkeys (n=3) to unoperated controls (n=7) (FIG. 1) in two delayed-matching-to-position spatial recognition tasks (FIG. 2). We specifically chose sdlPFC (i.e., cortex superior to the principal sulcus up to the midline as opposed to more ventral dlPFC areas 46 and 9/46 in and around the principal sulcus) and OFC because we anticipated (on the basis of a wide-ranging PFC lesion study literature review in macaque monkeys) that neither lesion would likely impair first-order spatial recognition per se, thereby making, if our hypotheses supported, any changes on second-order recognition easier to discern.

Three different but complimentary indices were used to quantify individual animal's metacognitive ability ("type II sensitivity"), as defined as the ability to accurately link confidence with performance. Here, we calculated the *meta-d′/d′*, a metric for estimating the metacognitive efficiency (level of metacognition given a particular level of performance or signal processing capacity), which enables a model-based approach to the computation of type II sensitivity that is independent of response bias and type I sensitivity (*d′*) on the primary task [7]. We also computed metacognitive efficiency using a hierarchical Bayesian estimation method, which can avoid edge-correction confounds and enhance statistical power [19]. Both *meta-d′* and *d′* measures assume that the variance of the internal response takes a Gaussian distribution, and that the distributions associated with the two type 1 responses respectively are of equal variance. To ensure our results were not due to any idiosyncratic violation of the assumptions of SDT, we additionally calculated the Phi coefficient (Φ), which does not make

these parametric assumptions [20]. We found that sdlPFC-lesioned monkeys – but not the OFC-lesioned ones – are impaired in meta-accuracy in the high spatial memory demand variant without showing any impairments in spatial recognition performance itself. We further established that these putative metacognitive deficits were specific to spatial recognition memory rather than to other confounds such as rule-learning, reward evaluation, or general representation of task information with results arising from analyses of some extant data of the same sdlPFC-lesioned monkeys when previously tested on a Wisconsin Card Sorting Test analog [13].

RESULTS

Meta-deficits in dlPFC-lesioned group in spatial recognition

Consistently with the two meta-indices, we revealed a significant main effect of "Group" in the spatial-variant task with a one-way ANOVA on SDT *meta-d'/d'*: F (2, 9) = 5.464, P = 0.028, post hoc test: CON vs. sdlPFC, one-tailed Dunnet P = 0.034, CON vs. OFC, P = 0.968; and on hierarchical-model *meta-d'/d'*: F (2, 9) = 6.524, P = 0.018, post hoc test: CON vs. sdlPFC, one-tailed Dunnet P = 0.020, CON vs. OFC, P = 0.964. The sdlPFC monkeys were impaired in meta-accuracy in spatially demanding recognition, whereas the OFC group did not show any meta-deficit in either of the tasks. In contrast, in the temporal-variant task, we found no main effect of Group in meta-accuracy, *meta-d'/d'*: F (2, 10) < 1; hierarchical-model *meta-d'/d'*: F (2,10) < 1. To add credibility to these results, we replicated these findings with the Phi coefficient (Φ) in both tasks, that is in the spatial-variant, Group effect: F (2, 9) = 4.904, P = 0.036, post hoc test: CON vs. sdlPFC, one-tailed Dunnet P = 0.026, CON vs. OFC, P = 0.904, and in the temporal-variant, Group effect: F (2, 10) < 1. These results using three indices convergently revealed severe impairment in metamemory of recognition in the sdlPFC lesioned monkeys (but not in OFC group), confirming that metacognitive ability was impaired in the spatially-demanding spatial recognition task, but not the temporally-demanding task (FIG. 3A-C).

These meta-indices in principle refer to how meaningful a subject's confidence is in distinguishing between correct and incorrect responses. We accordingly ran two mixed-design repeated-measures ANOVAs on percentage correct with "Group" as a between-subjects variable and "Confidence" as a within-subjects variable for the two tasks separately and obtained a significant interaction with the spatial-variant task F (2, 9) = 5.416, P = 0.029, but not with the temporal-variant task F (2, 10) = 0.355, P = 0.710. Percentage correct in high-confidence trials is usually higher than low-confidence trials, $P < 0.01$ for both CON and OFC monkeys in both tasks, but such effects were disrupted in the sdlPFC monkeys in the spatial-variant task P = 0.696, indicating the sdlPFC monkeys were unable to keep track of the efficacy of confidence during memory judgement (FIG. 3D-E). Correspondingly, one-way ANOVAs having "Group" as a between-subjects variable on *meta-d'* (a sensitivity measure quantifying the ability to discriminate between correct and incorrect judgments) for the two tasks separately also revealed that a significant main effect of "Group" in the spatial-variant task on SDT *meta-d'*: F (2, 9) = 5.701, P = 0.025, post hoc test: CON vs. sdlPFC, one-tailed Dunnet P = 0.015, CON vs. OFC, P = 0.867, but not in the temporal-variant F (2, 10) < 1.

In order to ascertain that these lesion effects were task-specific (Task: spatial-variant/temporal-variant), we ran three separate mixed-design repeated-measures ANOVAs considering only the CON and sdlPFC groups with "Group" as a between-subjects variable and "Task" as a within-subjects variable and confirmed a marginally significant "Task × Group" interaction on SDT *meta-d'/d'*: F (1, 7) = 4.194, P = 0.080 and on hierarchical-model *meta-d'/d'*: F (1, 7) = 4.599, P = 0.069, as well as a slightly weaker effect on Phi coefficient (Φ): F (1, 7) = 1..939, P = 0.206.

sdlPFC and OFC lesions did not result in recognition impairment

Given that metacognition is quantified by the correspondence between confidence and type 1 task performance, it is theoretically important to establish that the task (first-order) performances were matched between the groups in order to argue for the presence of a true difference in metacognition caused by the sdlPFC lesion. Despite the deficits in metamemory

accuracy in the sdlPFC group, importantly, we further established that there were not any memory deficits in their type I performance. In two mixed-design repeat-measures ANOVAs we entered the percentage correct or RT with one between-subjects factor "Group" and one within-subjects factor "Condition" and found neither a main effect nor interaction effects with "Group" in the two tasks, temporal-variant: % correct: F (2, 10) = 0.284, P = 0.759, RT: F (2, 10) = 0.932, P = 0.425, no "Group × Condition" interaction all Ps > 0.05; spatial-variant: % correct: F (2, 9) = 3.868 P = 0.061, RT: F (2, 9) = 0.794, P = 0.481, no "Group × Condition" interactions all Ps > 0.05. The ANOVAs also showed that performance decreased with Delay, % correct: F (4, 40) = 26.964, P < 0.001, RT: F (4, 40) = 9.918, P < 0.001, and with Separation, % correct: F (3, 27) = 33.960, P < 0.001, RT: F (3, 27) = 0.121 P = 0.105 for all three groups (FIG. 4). These analyses point to the fact that neither sdlPFC nor OFC lesion resulted in any type 1 recognition memory impairment.

Meta-deficits could not be explained away by speed-accuracy trade off

Here we have utilized reaction time as a proxy for memory decision confidence [18]. The meta-deficit effects observed here might be confounded by some speed-and-accuracy trade-off strategy adopted by the monkeys towards maximizing the time-spent per unit of reward (correct) ratio. Speed-and-accuracy trade-off taps into the monitoring of the current state of mind as regards the uncertainty properties of the judgement [21], whereas RT-indexed metacognition – defined as an introspective evaluation process – taps into the higher-level function. We thus analyzed the ratio between percentage correct and RT for each monkey for both tasks. In two mixed-design repeat-measures ANOVAs we entered the inverse efficiency (RT / % correct) with one between-subjects factor "Group" and one within-subjects factor "Condition". We found neither a main effect nor interaction effects with "Group" in the two tasks, all Ps > 0.05 (FIG. 5). We conclude that the putative lesion-related meta-deficits were not resultant from any speed-and-accuracy trade off.

No meta-memory deficit following sdlPFC lesion in short-term abstract rule memory in Wisconsin Card Sorting Test (WCST)

While the sdlPFC monkeys were impaired in metamemory for spatial recognition, we have not been able to ascribe the effects specifically to spatial recognition per se. Is this deficit uniquely ascribable to the metamemory in temporo-spatial recognition, or more generally to the metamemory of learning abstract rules, or other higher cognitive processes? Considering performance supporting WCST demands multi-processes such as memory and acquisition of abstract rules, as well as reward-value evaluation, we thus analyzed some extant data obtained from WCST to test for metacognitive deficits specifically in the sdlPFC monkeys. In contrast to the spatial recognition task, no meta-deficits were found with WCST in the sdlPFC group in comparison with the CON group in one-way ANOVAs on SDT *meta-d′/d′*: F (1, 10) = 0.677, P = 0.430; hierarchical-model *meta-d′/d′*: F (1, 10) = 0.666, P = 0.433; and Phi coefficient (Φ): F (1, 10) = 1.132, P = 0.312 (FIG. 6). These analyses confirm that meta-deficits caused by sdlPFC lesion were highly specific for spatial recognition and ruled out the explanation that such meta-deficits were attributable to processes involved in the maintenance of abstract rules or general representation of knowledge.

DISCUSSION

As expected, neither sdlPFC nor OFC lesions impaired spatially-complex or temporally-complex first-order recognition memory performance; yet importantly and consistent with our hypotheses we established that sdlPFC lesions induced selective deficits in a second-order meta-recognition within a recognition memory paradigm taxing recent spatial memory. No such change in metacognitive ability was observed after OFC lesions both affirming a critical functional role of the sdlPFC in supporting metamemory and showing evidence for functional specificity within prefrontal cortex for elements of metacognition. Our findings are robust because we assessed multiple measures of metamemory (both SDT and hierarchical model *meta-d′/d′* and Phi coefficient (Φ)) and found consistent significance across measures. Our findings are important because they provide causal evidence towards refining functional specialization of sub-regions of anterior part of the primate PFC underpinning the introspection

203 ability during memory recognition, and to date such causal evidence is very limited.

204

205 **Critical role of dlPFC in meta-recognition memory**

206 The role of aPFC in meta-evaluation of visual perception has been relatively well
207 established, by evidence observed in functional activation during post-decision evaluation [22,23],
208 various structural and connectivity profiles [4,24], and neuromodulation studies [25]. In contrast, the
209 neural basis of metamemory remains largely unknown. Previously, monkeys with combined
210 lesions to mid-dlPFC lesion (areas 46 and 9/46) and superior part of the mid-dlPFC (lateral area
211 9) were found to be unimpaired on standard first-order recognition memory maintenance for
212 recently presented objects (in contrast to lesions to ventrolateral PFC and OFC which do impair
213 first-order object recognition [26,27]) but were severely impaired in executive processes of
214 monitoring visual working memory information in a self-ordered version of the task [10];
215 moreover, in the same study monkeys with more restricted lesions within lateral area 9, a similar
216 region to our sdlPFC lesions in this study were impaired in the self-ordered task. Together with
217 other lesion findings that patients with more diffuse frontal lobe pathology exhibit impaired
218 feeling-of-knowing in the absence of amnesia [28], our pattern of results corroborate the extant
219 evidence that a dissociation between type 1 vs. type 2 performance in recognition is critically
220 dependent on the dorsolateral frontal cortices. This dissociation also aligns with the recent
221 distinction stipulating the dlPFC's putative role in metacognitive control as opposed to
222 decision-making per se [11].

223 Importantly, and in light of recent evidence from the macaque showing that there exist
224 dissociated networks between anterior and posterior dorsal PFC underlying metacognition for
225 remote versus recent recognition of visual objects respectively [9], our results significantly extend
226 the causal evidence for understanding the functional neuroanatomy of metacognition in PFC in
227 several key ways. Firstly, we extend causal evidence for metacognition into the spatial memory
228 domain. Secondly, we dissociate the functional neuroanatomy of first-order spatial recognition
229 memory from second-order spatial meta-recognition within PFC. The dlPFC region just dorsal
230 to the principal sulcus is believed to contain a spatial memory map as circumscribed muscimol

injections to different regions within this area impair visually-guided saccades in a visuospatial working memory task in a topographical manner [29]; further evidence that this dlPFC region but not the more superior dlPFC region impairs first-order spatial recognition comes from previous surgical lesions in monkeys [30]. Our sdlPFC lesion here fails to encroach much on this first-order region and accordingly fails to impair first-order spatial recognition; at the same time it clearly does impair meta-recognition in the same task, therefore demonstrating functional dissociation between the crucial sites required to be intact in order to mediate first-order and second-order spatial recognition within PFC. Taken together with Miyamoto et al, this study confirms that sdlPFC neither contributes exclusively to object nor spatial meta-recognition rather it contributes to both. Interestingly we found dissociation between meta-recognition deficits after sdlPFC lesions in a spatially demanding but not temporally demanding version of our task. The main difference between these two versions is that in the former the spatial difficulty is intentionally varied between trials; given it is a spatial recognition task we postulate efficacious metacognitive monitoring of performance will therefore be in flux and continually challenged necessitating a dynamic signal (from sdlPFC) to be input into the wider neural system in support of metacognitive computation. By contrast, in the temporal-variant the spatial difficulty is constant across trials so any metacognitive evaluation signal from sdlPFC would likely be a less important parameter to the system and either absent or less liable to be disrupted by the sdlPFC lesion accordingly, in accordance with our behavioural observations.

OFC represents value/confidence but not supports self-introspection

At the first glance, the OFC is the interpreter of specific values especially in terms of sorting and representation of inferred information [31,32], could it be the cornerstone of meta-appraisal towards one's own memory performance? Despite proposals that the orbitofrontal cortex is a key part of continuous decision-making under uncertainty [33], related to the explicit manifestation of decision confidence [15,16] and various aspects of decision-making [34], its role in metacognition in the present experimental context appears to be none as evident in the total absence of meta-impairment in the OFC group. One likely possibility is that value-assignment

[35], valuation of inferred information [31], and decision-confidence [16] per se are fundamentally distinct/dissociable from metamemory introspection. Indeed, meta-decision processes – as measured by the SDT and hierarchical-model *meta-d'* here – are in principle "bias-free" and are immune to any bias due to "confidence" [20], suggesting that the computation performed by the OFC underlying confidence signals [15,16] does not necessarily equate to the same neurobiological prerequisite for meta-cognitive computation. Relatedly, in our tasks there was no explicit requirement for reporting confidence, in which case the memory response need not to be bound with any explicit value valuation processes [36] or reward-based updating [13]. The introspection following memory decision was thus based entirely on some self-generated space, without any feedback or input exerted externally on their decision confidence or monitoring of degree of uncertainty [33]. This task feature discrepancy might explain the lack of meta-awareness deficits even when the OFC was surgically obliterated.

An alternative but not mutually exclusive explanation is that given that the frontal vs. parietal neural basis/correlates of metacognition is known to be domain-specific [6], it remains possible the OFC's contribution to metacognition might analogously be domain-specific. Previous studies tapping into the OFC role in meta-related processes were all on perceptual decisions, such as odor discrimination judgement [15,16], whereas at present the tasks in question concern mnemonic decisions. Some recent work on humans have evinced such specificity for perceptual vs. memory metacognition [1,37], and at present the differential effects of sdlPFC lesion on spatial vs. temporal-variant (see above) are also highly indicative for such specificity (see also Miyamoto et al., 2017).

A wider theoretical implication afforded by the current study is that this same group of sdlPFC monkeys – despite their impaired memory self-appraisal in the delayed-matching-to-position task – were completely intact in all aspects of a rule-guided memory Wisconsin Card Sorting Test analog [13]. This constitutes a stark contrast to the OFC-lesioned monkeys, who were impaired in updating rule-value representation in the WCST analog [13], but not in their introspection in the present delayed-matching-to-position tasks. These results taken together constitute a double dissociation between dorsolateral and ventromedial PFC regions in

differentially supporting two related, yet perhaps distinct, higher-order processes, providing compelling evidence suggestive of functional specialization of dual supervisory, self-monitoring abilities between dorsolateral vs. ventral parts of the PFC.

Supporting Information

Animals. Data were acquired from 16 adult macaque monkeys (eleven *Macaca mulatta*, three *Macaca fuscata*, and two *Macaca fascicularis*): Three monkeys had orbitofrontal lesion (OFC, consisting of two *M. mulatta* and one *M. fuscata*), another 3 monkeys had superior dorsolateral prefrontal cortex lesion (sdlPFC, consisting of one *M. mulatta* and two *M. fuscata*), 7 served as unoperated controls (CON, consisting of five *M. mulatta* and two *M. fascicularis*) for the main tasks, and 3 further CON were included only for the WCST analog analysis. All but six macaque monkeys were trained, operated and tested in Oxford, UK, the other six (three *Macaca fuscata* and three *Macaca mulatta*) in RIKEN Brain Science Institute, Wako, Japan. All animal training, surgery and experimental procedures were the same in both laboratories. Those conducted in the UK were licensed in compliance with the UK Animals (Scientific Procedures) Act 1986, and those in Japan were done in accordance with the guidelines of the Japanese Physiological Society and approved by RIKEN's Animal Experiment Committee.

Surgery. The operations were performed in sterile conditions with the aid of an operating microscope and the same surgeon performed all operations in both laboratories. The monkey was sedated on the morning of surgery with ketamine (10 mg／kg) or with ketamine (10 mg／kg), xylazine (0.5 mg／kg) and／or midazolam (0.25 mg／kg), i.m. Once sedated, the monkey was given atropine (0.05 mg／kg) to reduce secretions, antibiotic (amoxicillin, 8.75 mg／kg) for prophylaxis of infection, opioid (buprenorphine 0.01 mg／kg i.v, repeated twice at 4- to 6-h intervals on the day of surgery, i.v. or i.m.) and nonsteroidal anti-inflammatory (meloxicam, 0.2 mg／kg, i.v.) agents for analgesia, and an H2 receptor antagonist (ranitidine, 1 mg／kg, i.v.) to protect against gastric ulceration as a side-effect of the combination of steroid and nonsteroidal anti-inflammatory treatment. The head was shaved and an intravenous cannula put in place for intraoperative delivery of fluids (warmed sterile saline drip, 5 mL／h／kg). The monkey was moved into the operating theater, intubated, placed on isoflurane anaesthesia (1–2.75%, to effect, in 100% oxygen) and then mechanically ventilated. Adjustable heating

blankets allowed maintenance of normal body temperature during surgery. Monkeys were maintained in a state of deep anesthesia by monitoring pulse rate, blood oxygenation, body temperature, and peripheral reflexes throughout surgery. The monkey was placed in a head-holder and the head cleaned with alternating antimicrobial scrub and alcohol and draped to allow a midline incision. The skin and underlying galea were opened in layers. The temporal muscles were retracted to expose the skull surface and a bone flap was turned to allow access to the desired lesion site with the craniotomy extended with rongeurs as necessary. The dura was cut and reflected to expose the cortex and the lesion was made in the intended site by aspiration with the aid of an operating microscope. When the lesion was complete, the dura was drawn back or sewn, the bone flap replaced and held with loose sutures, and the skin and galea were closed in layers. The monkey was removed from the head-holder and anaesthesia discontinued. Nonsteroidal anti-inflammatory analgesic (meloxicam, 0.2 mg／kg, oral) and antibiotic (8.75 mg／kg, oral) treatment continued following surgery in consultation with veterinary staff, typically for 5 days.

Histology. After the conclusion of the experiments the animals with ablations were sedated, deeply anesthetized, and then perfused through the heart with saline solution (0.9%), which was followed by formol saline solution (10% formalin in 0.9% saline solution). The brains were blocked in the coronal stereotaxic plane posterior to the lunate sulcus, removed from the skull, allowed to sink in sucrose formalin solution (30% sucrose,10% formalin), and sectioned coronally at 50μm on a freezing 10 microtomes. Every 10th section through the temporal lobe was stained with cresyl violet and mounted. When referring to cytoarchitecturally defined regions in the lesion description below we have adopted the nomenclature and conventions of Petrides and Pandya [38] and have reconstructed lesion extents on standard drawings based upon those provided by the Laboratory of Neuropsychology at NIMH. Detailed description of the surgical procedure has been reported [13].

Superior dorsolateral prefrontal cortex (sdlPFC) lesion: The intended extent of the sdlPFC lesion was designed to include the cortex on the dorsolateral aspect of the PFC extending up to midline (i.e., lateral area 9 and the dorsal portions of areas 46 and 9/46) but excluding ventrally situated dlPFC cortex; the lesion excluded posteriorly located premotor areas 8A, 8Bd, and 8Bv, nor did it extend anteriorly into area 10. Fig. 1 (left panel) depicts coronal sections through the area of the intended lesion in the three sdlPFC animals (sdlPFC1 to sdlPFC3) in addition to drawing of reconstruction of the actual lesion extent on drawings of standard views of the lateral surfaces of the macaque brain. All three of the sdlPFC lesions were as intended.

Orbitofrontal cortex (OFC) lesion: The intended extent of the OFC lesion included at its lateral extent, the cortex in the medial bank of the lateral orbital sulcus; the lesion included all of the cortex between the medial and lateral orbital sulci, and also extended medially until the lateral bank of the rostral sulcus. The anterior extent of the lesion was an imaginary line drawn between the anterior tips of the lateral and medial orbital sulci, and the posterior extent was an imaginary line drawn just anterior to the posterior tips of these two sulci. The intended lesion therefore included areas 11, 13 and 14 of the orbital surface and did not extend posteriorly into the agranular insula. Fig. 1 (right panel) depicts coronal sections through the area of the intended lesion in the three OFC animals (OFC1 to OFC3) in addition to drawing of reconstruction of the actual lesion extent on drawings of standard views of the ventral surface of the macaque brain. None of the OFC lesioned animals sustained any bilateral damage outside the area of the intended region; two animals sustained extremely slight unilateral damage beyond the intended lateral boundary of the lesion OFC2 and OFC3; in all three animals the lesions did not extend as far medially as intended.

Metamemory quantification. *Signal detection theoretic and hierarchical Bayesian estimation meta-index (meta-d'/d')*: Using a Type II SDT toolbox [39], which has been extensively used for evaluation of metacognitive ability [1,40], it is possible to compute a measure of metacognitive accuracy that is unconfounded by type I performance directly from the empirical type II

receiver operating characteristic (ROC) curve. The type II ROC curve reflects the relationship between the accuracy of type I and the observer's confidence rating. This approach exploits the link between type I and type II SDT models to express observed type II sensitivity at the level of the type I SDT model (termed *meta-d'*). Maximum likelihood estimation is used to determine the parameter values of the type I SDT model that provide the best fit to the observed type II data. A measure of metacognitive ability that controls for differences in type I sensitivity is then calculated by taking the ratio of *meta-d'* and the type I sensitivity parameter *d'*: meta efficiency, computed as *meta-d'/d'*. The most straightforward approach to computing meta efficiency involves an equal variance SDT model in which the variances of internal distributions of evidence for "target" and "foil" in the type I model are assumed to be equal. We thus quantified metacognitive sensitivity with the SDT-based measure *meta-d'*. Based on type II signal detection theory, meta-efficiency (in terms of *meta-d'/d'*) reflects how much information, in signal-to-noise units, provides a response-bias free measure of how well confidence ratings track task accuracy. The toolbox for the SDT-based *meta-d'/d'* estimation was available at http://www.columbia.edu/~bsm2105/type2sdt/. Of note, the standard type II SDT toolbox is designed for 2AFC tasks, in which S1 and S2 are always constant in the left or right, but our target and foil are randomly presented at the screen and we only recorded the separation of the two probes and did not track the specific position of the target and the foil. Since the target and the foil were presented randomly on the screen, and the SDT algorithm only requires the distribution of those four kinds of trials, we divided the number of those trials equally to S1 trials and S2 trials in a random manner considering the animals would not have any preference to any given side/location of the screen. In addition, we have also replicated the analyses using a variant of metacognitive efficiency (H-*meta-d'*) with a hierarchical Bayesian estimation method (https://github.com/smfleming/HMeta-d), which can avoid edge-correction confounds and enhance statistical power [19].

Phi coefficient (Φ): In order to ensure our results were not due to any idiosyncratic violation of the parametric assumptions of SDT, we additionally calculated the phi coefficient index, which

does not make the SDT assumptions. The phi coefficient is a contingency index of preference for optimal choice [41,42] and was calculated according to the following formula using the number of trials classified in each case [n(case)]:

$$\text{phi coefficient } (\Phi) = \frac{n(Correct\ High)\times n(Incorrect\ Low) - n(Correct\ Low)\times n(Incorrect\ High)}{\sqrt{n(Correct)\times n(Incorrect)\times n(High)\times n(Low)}}$$

.

The Φ coefficient evaluates how optimally each trial was assigned for high or low confidence based on performance in the preceding cognitive judgment, reflecting the correlation between the two binary variables. Note that despite differences in their mathematical assumptions, the three meta-cognitive metrics are highly correlated with each other (FIG. S1).

For the computation for SDT *meta-d'/d'*, hierarchical-model *meta-d'/d'*, and Phi coefficient, four types of trials and their distribution are required. The computation performed here are based on the premise that confidence is computed in a retrospective manner [43]. Using a summary of the decision process, trials that are responded fast are judged as more of higher certainty [44]. Following this logic, we accordingly used trial-specific reaction times (RT) as a proxy for confidence [18]. We collapsed all trials per monkey and classified the trials within-monkeys by the median of all RT crossing with correct/incorrect responses into four kinds: correct/high confidence (fast RT), incorrect/low confidence (slow RT), correct/low confidence (slow RT), incorrect/high confidence (fast RT). For each of the final analyses, each monkey had one single value for the measurement of meta-ability. The advantage of RT-indexed metacognition is that it does not suffer from any training-induced associations, which could contaminate true introspection.

Behavioral tasks and pre-analysis. *Spatial recognition tasks (delayed-matching-to-position, or DMP)*. A temporally-demanding DMP task and a spatially-demanding DMP task were performed by the monkeys. In both variants, each trial consisted of an encoding phase in which a spatial position ("sample") was indicated by a red cross. After the monkey touched the sample,

a blue square ("distractor") appeared in the center of an imaginary (i.e., invisible) circle whose circumference transected the centre of the red cross. A touch to the blue square initiated a variable delay interval (i.e., the manipulation of delay for the temporal-variant) and then a choice phase consisting of two identical red crosses in different positions, both located on the circumference of the aforementioned imaginary circle (hence equidistant from the blue square distractor just touched), albeit one positioned in the same (i.e., spatial-match) position as the first red cross and the other (i.e., non-match) positioned some angle (with respect to the centre of the imaginary circle) away from the spatial-match along the invisible circumference. From trial to trial we could vary the angle of separation along the circumference to allow for easy trials (i.e., large angle and accordingly large spatial separation) and harder trials (i.e., smaller angles and accordingly smaller spatial separation) (cf. the manipulation of separation between two probes for the spatial-variant). As mentioned above, one of the crosses appeared in the same position as the sample (target; S+) and the other one in a different position (foil; S-). A touch to the S+ resulted in a delivery of a reward pellet, removed the S-, and the S+ remained alone for a further 1 s for positive feedback. The screen would then be blanked for an ITI of 6 s before the next trial. A touch to the S- removed both S+ and S- from the screen and the screen would be blanked for an ITI of 12 s. There was no time constraint imposed on responses made to the choices and therefore there were no missed trials. No repetition correction routines were implemented following an error response, each trial was new, and independent of the outcome of the preceding trial. In terms of sizes of visual stimuli, the sample subtended a visual angle of 9° in task acquisition and the temporal-variant task, or 6.8° in the spatial-variant task; the distractor subtended a visual angle of 4.6° in all tasks (FIG. 2).

In the temporal-variant DMP, there were five trial-types with differing delay intervals (either 1, 2, 4, 8, or 16 s) between the distractor and probes. Trials within a session were divided into five trial-types with differing intervals of delay between the distractor and probes. The trial-type order was randomized within each successive set of five trials (with one trial of each trial-type per set) so that the delay changed unpredictably from one trial to another. The two probe

choices were separated by a visual angle of 21.7°. In the spatial-variant DMP, the separation between two red crosses varied across trials; there were four different trial-types with differing spatial separations (visual angles of either 4.8°, 8.6°, 15.2°, or 21.7°; equivalent to 5, 9, 16, or 23 cm on screen) between probe choices. Delays were fixed at 1 s for the spatial-variant DMP. In the final testing, each animal accrued 200 rewards to complete the temporal-variant (across two daily sessions) and 150 rewards (one session) to complete the spatial-variant. Since the animals accrued varying numbers of errors to complete the tasks, the mean total numbers of trials were 271.2 and 209.8 trials (averaged across groups) respectively for the two tasks. One CON did not complete the spatial-variant so only 12 monkeys were analyzed in the spatial-variant task.

For the formal analyses, trials with RT longer than 20 s and shorter than 100 ms in the memory judgement were discarded (< 0.5%). Moreover, we also removed trials with touch-distractor RT longer than 1000 ms (15.0% and 20.2% trials discarded respectively for temporal-variant task and spatial-variant task). There were no differences in touch-distractor RT between the groups in either of the tasks following this trial removal procedure, with a one-way ANOVA on temporal-variant: F (2, 10) = 1.27, P = 0.322; and on spatial-variant: F (2, 9) < 1. The requirement to touch the distractor before the memory task was deliberately added to gauge whether the monkeys were distracted and/or less willing/ready prior to initiating the memory judgement. Such stringent selection of good trials on which the monkeys were attentive is crucial for the metacognition analysis. The whole set of main results did not differ if we chose to use other touch-distractor RT cut-off criteria of either 800 ms, 900 ms, 1100 ms, or 1200 ms.

Wisconsin Card Sorting Test (WCST) analog. Given that the DMP tasks involve multiple processes which might confound our main results, we therefore analyzed extant data obtained from a WCST analog, which is a validated rule-guided task taxing multi-processes such as perception (involved in matching stimuli), memory and acquisition of abstract rules, and reward-value evaluation [13,45]. We accordingly used some WCST data to rule out that the putative

meta-deficits observed here were not attributable to these perceptual and reward-value evaluation processes. The WCST analog paradigm is summarized as follows: on each trial, a randomly selected sample (a square, a circle, or a cross of different colors) is displayed alone on the center of the touch screen, and when the sample is touched, three additional choice items immediately appear (one matching in color, one matching in shape, and one not matching in either dimension), with their positions randomly chosen. If the animal's choice is correct (i.e., the animal selects the choice item that matches according to the currently reinforced rule, which changes unannounced every time the animal attains 85% in 20 consecutive trials), then a reward pellet is delivered, and the correct choice remains on the screen for 1 s to provide visual feedback; if the animal makes an incorrect choice, then no reward is given, and the stimuli are removed and replaced by an error signal (white circle), which is presented on the screen for 1 s instead.

We analyzed WCST data from 12 monkey data points (9 CON vs. 3 sdlPFC). Six out of the nine CON monkey data points here were from the pre-lesion data of the six lesioned monkeys (3 sdlPFC and 3 OFC). We included 3,000 trials (acquired from ten 300-trial daily sessions) per monkey data point. We collapsed all trials and classified the trials into four types of trials for the computation for the meta-efficiency and Phi coefficient. Since the type II SDT toolbox was designed for 2AFC tasks, and the WCST task contained three stimuli, we ran three separate sets of computation, each one discarding only either the bottom, left, or right choice, for each of the three meta-indices. For each monkey, we then computed the mean of these three values as his meta-score to enter into the meta-indices calculation. In this analysis, we did not include the OFC monkeys because the OFC monkeys were severely impaired in the WCST type I task, thus making any analyses on meta-ability invalid (their chance level implies they did not know how to make correct judgments, violating the prerequisite for the meta-assessment of their judgement).

Preliminary training. All monkeys completed preliminary training and task acquisition before

performing the two main tasks and WCST described above. We conducted the spatial-variant task immediately after the temporal-variant task without any additional training. The monkeys performed one session per day, 6–7 d per week. For the lesioned animals, the task was administered post-operatively (on average 22 months post-lesion). For the two DMP tasks, during task acquisition the monkeys were trained until they reached ≥ 90% performance level within a 100-reward session. All trials in this stage consisted of a short delay interval (1 s), and a wide separation between choice positions (21.7°, or 23 cm) to make the trials "easy" to acquire. Upon reaching criterion, the three groups were not different in the number of errors accrued, $F < 1$ and number of rewards received, $F < 1$, indicating that the groups of lesioned monkeys learned to perform these spatial recognition problems as well as controls.

Apparatus. The tasks were performed in an automated test apparatus. The subject sat, unrestrained, in a wheeled transport cage fixed in position in front of a touch-sensitive screen on which the stimuli could be displayed. The animals could reach out between the horizontal or vertical bars (spaced approx. 45 mm apart) at the front of the transport cage to touch the screen. An automated pellet delivery system delivered banana flavored reward pellets (190 mg supplied by Noyes Company Inc. and Neuroscience Inc.) into a food well (approx. 80 mm in diameter) positioned beneath and to one side of the screen, in response to correct choices made by the subject to the touch screen. Pellet delivery was accompanied by an audible click. A spring-loaded lunchbox (length 200 mm, width 100 mm, height 100 mm) was positioned beneath and to one side of the subject; this opened immediately with a loud crack on completion of the testing session and contained the subject's daily diet of wet monkey chow, primate pellets, nuts, raisins, and a slice of apple, banana, and orange (water was provided in the home cage *ad libitum*). An infrared camera allowed the subject to be observed while it was engaged in the task. The entire apparatus was housed in an experimental cubicle that was dark apart from the background illumination from the touch screen. A computer, with a millisecond accuracy timer-card to record reaction times, controlled the experiment and data acquisition. Identical software controlled the tasks in both laboratories to ensure that the tasks were replicated exactly.

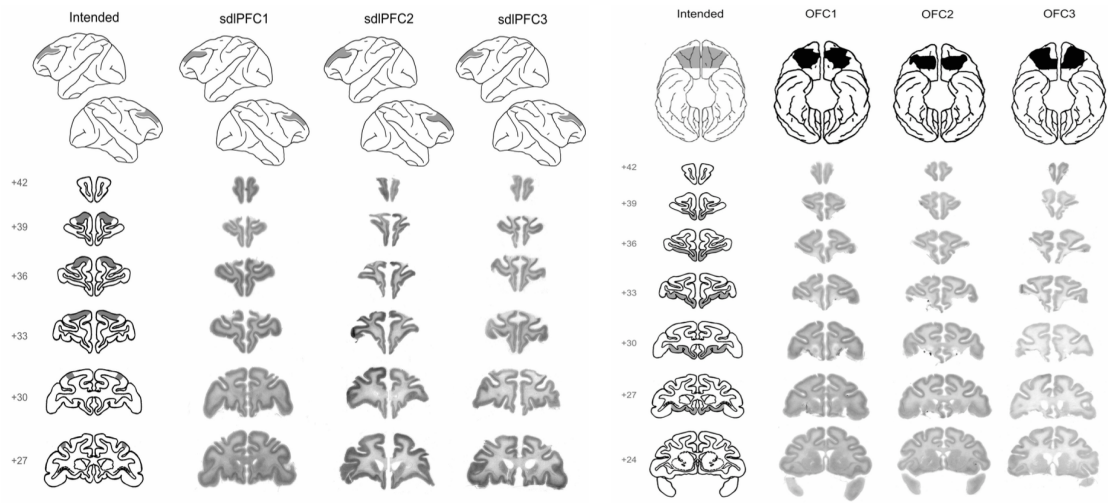

Figure 1. Histology. Photomicrographs of stained coronal sections through the intended lesion in the three animals with sdlPFC and OFC lesions (left panel: sdlPFC1 – 3, right panel: OFC1 – 3). Top row in each panel shows reconstruction of the area lesioned on drawings of representative lateral surfaces. Numerals: distance in mm from the interaural plane.

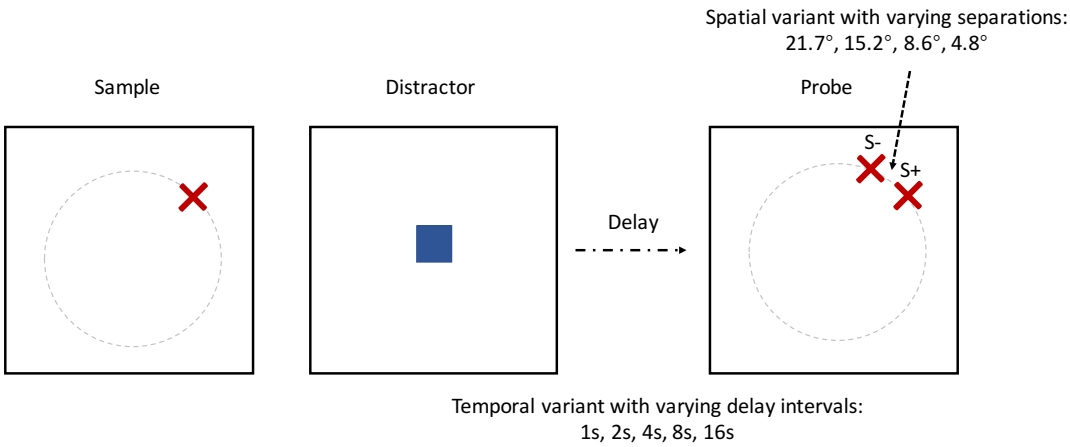

Figure 2. Delayed-matching-to-position tasks. Each trial consisted of a sample (red cross), a distractor (blue square), a delay, and a probe (2 red crosses). Temporally-taxing DMP: five levels of delay interval between distractor and probes (either 1, 2, 4, 8, or 16 s); spatially-taxing variant DMP: four levels of separation between two red crosses in probe (visual angles of either 4.8°, 8.6°, 15.2°, or 21.7° which are equivalent to 23, 16, 9, and 5 cm on screen; all delay fixed at 1s). S+ denotes the target, S- the foil. The grey dotted circle in the figure is invisible to the animal.

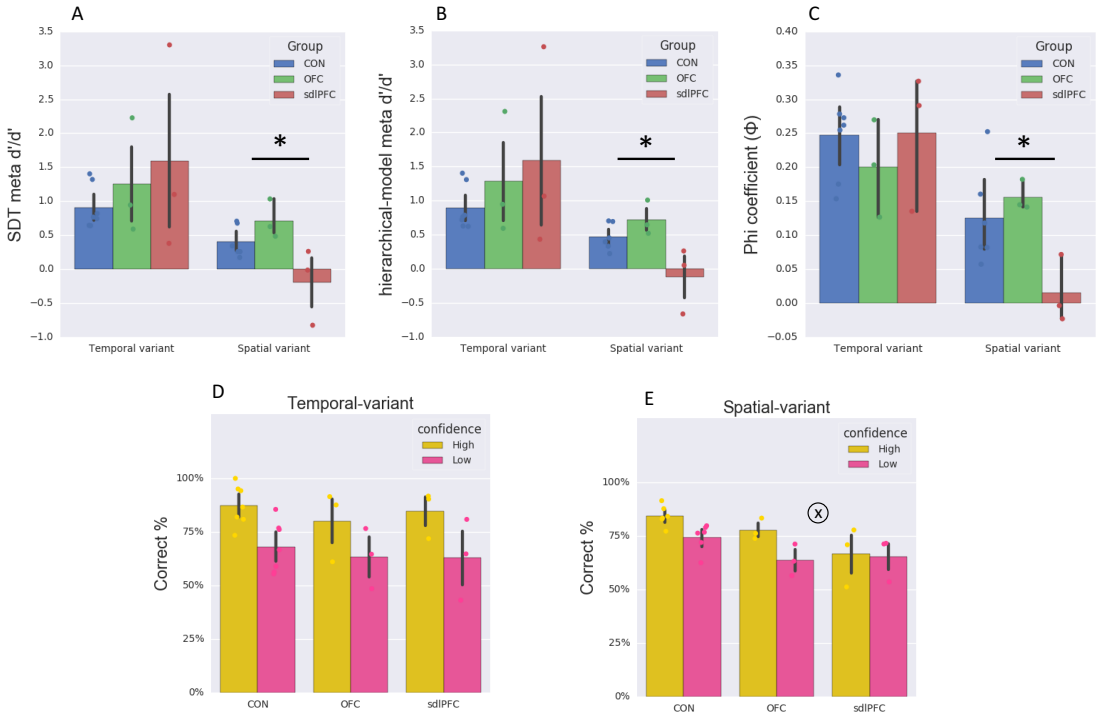

Figure 3. Differential deficits in meta-indices and accuracy x confidence interaction in sdlPFC group in spatially-demanding recognition. Meta performance for the 3 monkey groups (OFC, sdlPFC, CON). Metacognitive accuracy in sdlPFC group was lower than CON group for spatial-variant task, but not for the temporal-variant task: (A) SDT *meta-d'/d'*, (B) hierarchical-model *meta-d'/d'*, and (C) Phi coefficient (Φ). Horizontal axes represent the two spatial recognition tasks (temporal-variant; spatial-variant). Vertical axes represent the three meta indices. (D – E) Accuracy in high-confidence trials is usually higher than low-confidence trials (for both CON and OFC monkeys in both tasks) but such effects were disrupted in the sdlPFC monkeys especially in the spatial-variant task. ⊗ indicates significant "Group x Confidence" interaction $P < 0.05$, * $P < 0.05$. Colored dots depict individuals. Error bars indicate 90% confidence interval around the estimate computed by a bootstrapping procedure.

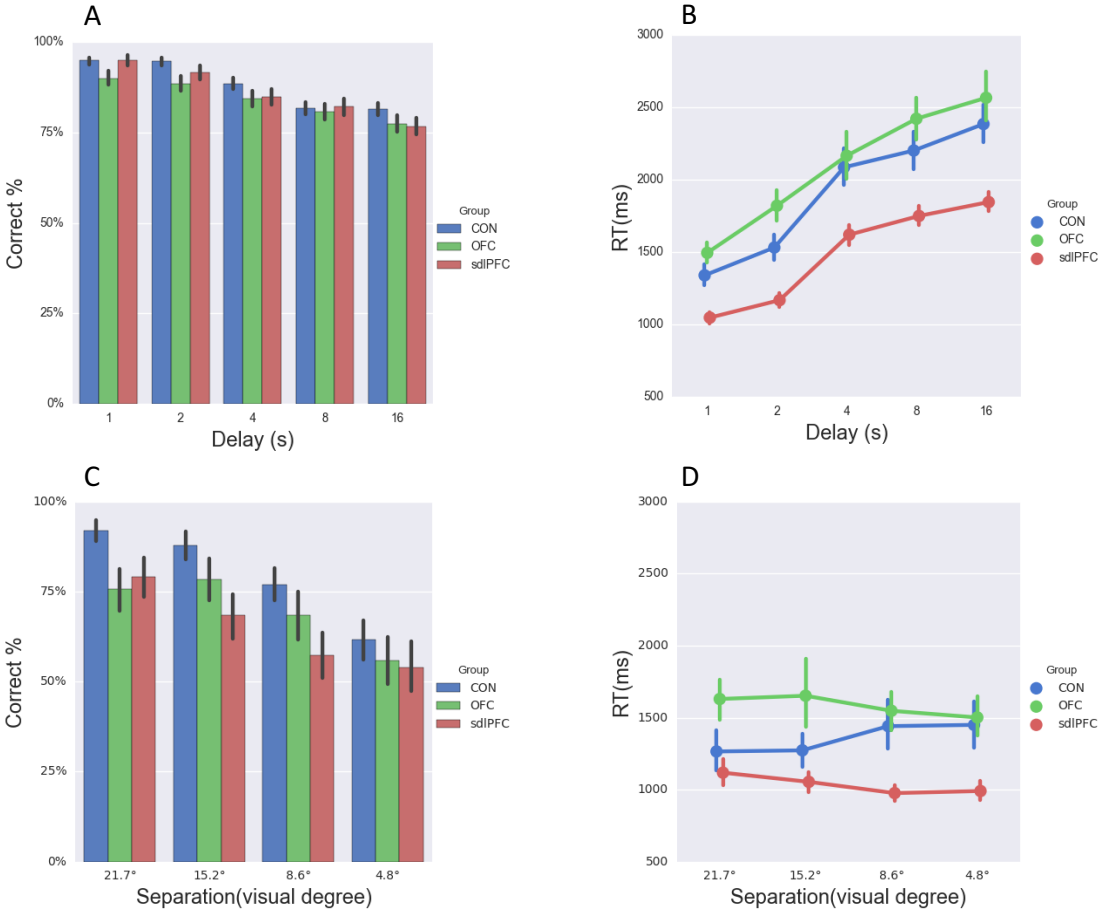

Figure 4. sdlPFC and OFC lesions did not result in recognition impairment. Memory task performance was intact in both tasks: (A) temporal-variant percentage correct, (B) temporal-variant reaction time, (C) spatial-variant percentage correct, and (D) spatial-variant reaction time. Error bars indicate 90% confidence interval around the estimate computed by a bootstrapping procedure.

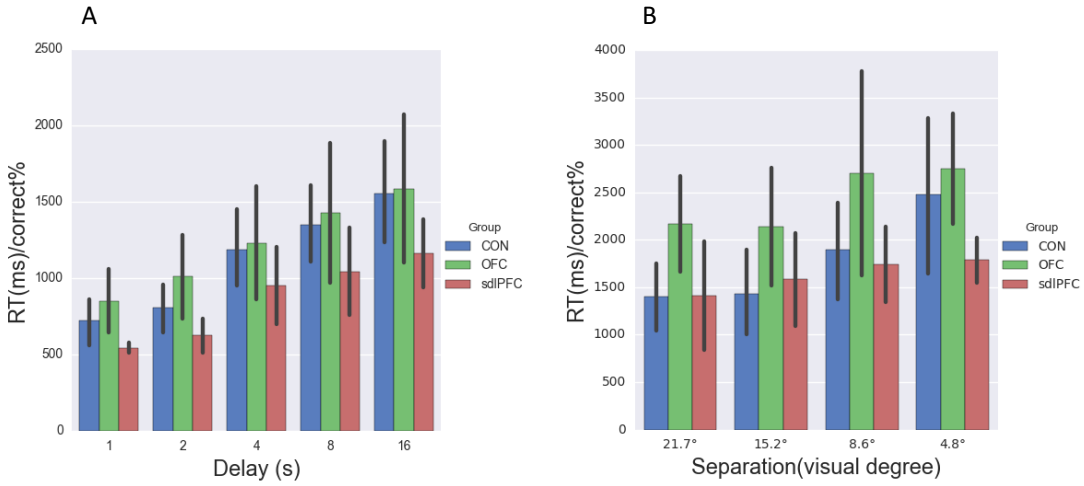

Figure 5. Meta-memory deficits could not be explained away by speed-accuracy trade off. Inverse efficiency (reaction time / percentage correct) shows no main effect of "Group" in (A) temporal-variant task and (B) spatial-variant task. Error bars indicate 90% confidence interval around the estimate computed by a bootstrapping procedure.

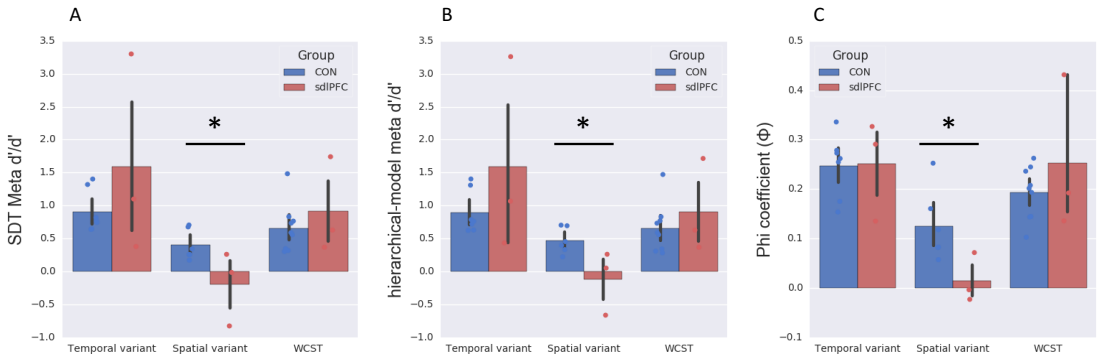

Figure 6. No meta-memory deficit following dlPFC lesion in WCST analog. Meta performance for the 2 monkey groups (sdlPFC, CON) in the two spatial recognition tasks (temporal-variant; spatial-variant) and WCST analog. Vertical axes represent the three meta indices: (A) SDT *meta-d'/d'*, (B) hierarchical-model *meta-d'/d'* and (C) Phi coefficient (Φ). Metacognitive accuracy in sdlPFC group was lower than CON group for spatial-variant task (see also FIG. 3A-C), but not for WCST analog or temporal-variant task. Error bars indicate 90% confidence interval around the estimate computed by a bootstrapping procedure. * $P < 0.05$. Colored dots depict individual monkeys.

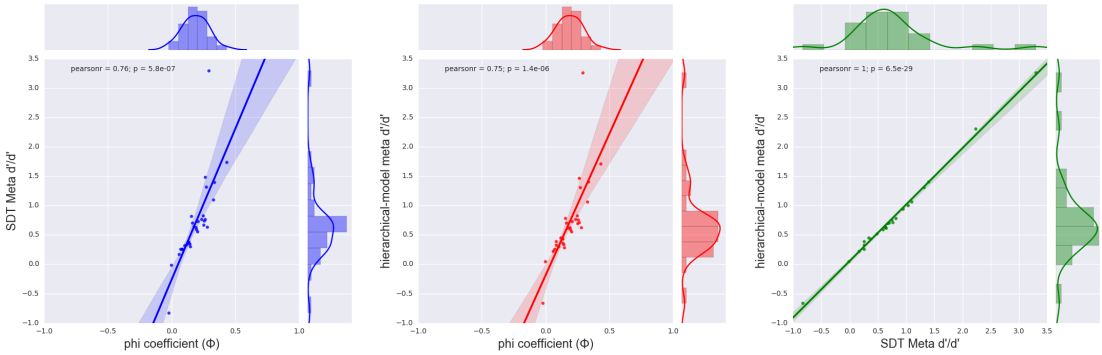

Figure S1. Strong correlations among the three meta-cognitive metrics. Pearson correlations computed among the three indices were all statistically significant (all *P*s < 0.001). Colored dots depict individual data points collapsed across monkey groups, with each monkey shown twice (temporal- and spatial-variants).

Acknowledgements: This work is supported by the Ministry of Education of PRC Humanities and Social Sciences Research grant 16YJC190006, STCSM Shanghai Pujiang Program 16PJ1402800, STCSM Natural Science Foundation of Shanghai 16ZR1410200, Fundamental Research Funds for the Central Universities 2018ECNU-HWFW007, NYU Shanghai and NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai (S.C.K.); Royal Society University Research Fellowship, and U.K. Medical Research Council (MRC) project grant G0300817 (M.J.B.).

Author Contributions: S.C.K. and Y.C. conceived research. S.C.K. and M.J.B. designed and performed experiments. S.C.K. and Y.C. analyzed data. S.C.K. wrote the manuscript with input from Y.C. and M.J.B.

Conflict of Interest: The authors declare no competing financial interests.

REFERENCES

1 Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811-2822 (2014).

2 Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**, 6117-6125, doi:10.1523/JNEUROSCI.6489-11.2012 (2012).

3 Morales, J., Lau, H. & Fleming, S. M. Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *The Journal of Neuroscience* **38**, 3534-3546, doi:10.1523/JNEUROSCI.2360-17.2018 (2018).

4 Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541-1543, doi:10.1126/science.1191883 (2010).

5 Allen, M. *et al.* Metacognitive ability correlates with hippocampal and prefrontal microstructure. *Neuroimage* **149**, 415-423, doi:10.1016/j.neuroimage.2017.02.008 (2017).

6 McCurdy, L. Y. *et al.* Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience* **33**, 1897-1906, doi:10.1523/JNEUROSCI.1890-12.2013 (2013).

7 Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E. & Lau, H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience* **1**, 165-175, doi:10.1080/17588921003632529 (2010).

8 Shekhar, M. & Rahnev, D. Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *The Journal of Neuroscience* **38**, 5078-5087, doi:10.1523/jneurosci.3484-17.2018 (2018).

9 Miyamoto, K. *et al.* Causal neural network of metamemory for retrospection in primates. *Science* **355**, 188-193 (2017).

10 Petrides, M. Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. *The Journal of Neuroscience* **20**, 7496-7503 (2000).

11 Qiu, L. *et al.* The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS biology* **16**, e2004037 (2018).

12 Baltz, E. T., Yalcinbas, E. A., Renteria, R. & Gremel, C. M. Orbital frontal cortex updates state-induced value change for decision-making. *eLife* **7**, e35988, doi:10.7554/eLife.35988 (2018).

13 Buckley, M. J. *et al.* Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science* **325**, 52-58, doi:10.1126/science.1172377 (2009).

14 Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human orbitofrontal cortex

663 represents a cognitive map of state space. *Neuron* **91**, 1402-1412 (2016).
664 15 Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates,
665 computation and behavioural impact of decision confidence. *Nature* **455**, 227-
666 231 (2008).
667 16 Lak, A. *et al.* Orbitofrontal cortex is required for optimal waiting based on
668 decision confidence. *Neuron* **84**, 190-201 (2014).
669 17 Rudebeck, P. H. & Murray, E. A. The orbitofrontal oracle: cortical mechanisms
670 for the prediction and evaluation of specific behavioral outcomes. *Neuron* **84**,
671 1143-1156 (2014).
672 18 Dotan, D., Meyniel, F. & Dehaene, S. On-line confidence monitoring during
673 decision making. *Cognition* **171**, 112-121,
674 doi:https://doi.org/10.1016/j.cognition.2017.11.001 (2018).
675 19 Fleming, S. M. & Daw, N. D. Self-Evaluation of Decision-Making: A General
676 Bayesian Framework for Metacognitive Computation. *Psychological Review*
677 **124**, 91-114, doi:10.1037/rev0000045 (2017).
678 20 Fleming, S. M. & Lau, H. How to measure metacognition. *Frontiers in Human*
679 *Neuroscience* **8**, 443 (2014).
680 21 Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U. & Nieuwenhuis, S. The
681 neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences* **33**, 10-16
682 (2010).
683 22 Wan, X., Cheng, K. & Tanaka, K. The neural system of postdecision evaluation
684 in rostral frontal cortex during problem-solving tasks. *eNeuro* **3**, 1-24,
685 doi:10.1523/eneuro.0188-16.2016 (2016).
686 23 Ochsner, K. N. *et al.* The neural correlates of direct and reflected self-
687 knowledge. *Neuroimage* **28**, 797-814 (2005).
688 24 Fleming, S. M. & Dolan, R. J. The neural basis of metacognitive ability. *Phil.*
689 *Trans. R. Soc. B* **367**, 1338-1349 (2012).
690 25 Meiron, O. & Lavidor, M. Prefrontal oscillatory stimulation modulates access
691 to cognitive control references in retrospective metacognitive commentary.
692 *Clinical Neurophysiology* **125**, 77-82 (2014).
693 26 Kowalska, D. M., Bachevalier, J. & Mishkin, M. The role of the inferior
694 prefrontal convexity in performance of delayed nonmatching-to-sample.
695 *Neuropsychologia* **29**, 583-600 (1991).
696 27 Bachevalier, J. & Mishkin, M. Visual recognition impairment follows
697 ventromedial but not dorsolateral prefrontal lesions in monkeys. *Behavioural*
698 *brain research* **20**, 249-261 (1986).
699 28 Janowsky, J. S., Shimamura, A. P. & Squire, L. R. Memory and metamemory:
700 Comparisons between patients with frontal lobe lesions and amnesic patients.
701 *Psychobiology* **17**, 3-11 (1989).
702 29 Sawaguchi, T. & Iba, M. Prefrontal cortical representation of visuospatial
703 working memory in monkeys examined by local inactivation with muscimol.
704 *Journal of Neurophysiology* **86**, 2041-2053 (2001).

30 Levy, R. & Goldman-Rakic, P. S. Association of storage and processing functions in the dorsolateral prefrontal cortex of the nonhuman primate. *The Journal of Neuroscience* **19**, 5149-5158 (1999).

31 Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953-956 (2012).

32 Wallis, J. D. Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience* **30**, 31-56 (2007).

33 Kennerley, S. W., Behrens, T. E. & Wallis, J. D. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nature Neuroscience* **14**, 1581-1589 (2011).

34 Izquierdo, A. Functional heterogeneity within rat orbitofrontal cortex in reward learning and decision making. *The Journal of Neuroscience* **37**, 10529-10540 (2017).

35 Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223-226 (2006).

36 Rolls, E. T. & Grabenhorst, F. The orbitofrontal cortex and beyond: from affect to decision-making. *Progress in Neurobiology* **86**, 216-244 (2008).

37 Ye, Q., Zou, F., Lau, H., Hu, Y. & Kwok, S. C. Causal evidence for mnemonic metacognition in human precuneus. *The Journal of Neuroscience* **38**, 6379-6387 (2018).

38 Petrides, M. & Pandya, D. Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *European Journal of Neuroscience* **11**, 1011-1036 (1999).

39 Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition* **21**, 422-430 (2012).

40 Baird, B., Smallwood, J., Gorgolewski, K. J. & Margulies, D. S. Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience* **33**, 16657-16665 (2013).

41 Middlebrooks, P. G. & Sommer, M. A. Metacognition in monkeys during an oculomotor task. *Journal of Experimental Psychology: Learning Memory and Cognition* **37**, 325-337 (2011).

42 Kornell, N., Son, L. K. & Terrace, H. S. Transfer of Metacognitive Skills and Hint Seeking in Monkeys. *Psychological Science* **18**, 64-71, doi:10.1111/j.1467-9280.2007.01850.x (2007).

43 Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review* **117**, 864-901 (2010).

44 Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329-1342 (2014).

45 Mansouri, F. A., Buckley, M. J., Mahboubi, M. & Tanaka, K. Behavioral

consequences of selective damage to frontal pole and posterior cingulate cortices. *Proceedings of the National Academy of Sciences* **112**, E3940-E3949 (2015).