The ghost in the machine: Self-reflective consciousness and the neuroscience of metacognition.

Janet Metcalfe

Columbia University

Bennett L. Schwartz

Florida International University

To appear in J. Dunlosky & S. Tauber (Eds). *Oxford Handbook of Metamemory*.

Correspondence concerning this chapter should be addressed to Janet Metcalfe or Bennett Schwartz. Janet Metcalfe, Psychology Dept. 401B Schermerhorn 1190 Amsterdam Avenue, MC: 5501, New York, NY 10027, USA. Email may be sent to jm348@columbia.edu

Bennett Schwartz, Department of Psychology, DM 283 Florida International University, University Park, Miami, Florida, 33199, USA. E-mail may be sent to bennett.schwartz@fiu.edu.

Key terms: metacognition, metamemory, consciousness, anterior cingulate cortex, BA 10

Abstract

Although metacognition is considered to be the highest human cognitive function, the capability that separates humans from other animals, the source of self-knowing consciousness, and the crucial self-reflective function that allow us to have free will and to make self-determined choices, sorting out where this modern pineal gland resides in the brain is an enterprise fraught with peril. The literature on the neural basis of human metacognition, although numerically modest, is, nevertheless, confused and confusing. Searching for metacognition in the brain is like searching for the Holy Grail: It always seems to be in the next valley. A primary reason for the confusion, we argue, is that basic-level and meta-level processes are often conflated. Here we direct our investigation, insofar as possible, at only the meta-level processes of monitoring and control. Furthermore, two additional considerations are of prime importance. First, metacognitions are conscious. They spontaneously occur when something goes wrong, and a conflict-based "feeling state" is manifest. We argue that, at least when metacognitive feelings are spontaneous, that feeling state is adaptive because it triggers action (be it mental or physical) needed to resolve the conflict. The conscious feeling state is, therefore, related to the control function of metacognition. Second, metacognitive feelings, in humans at least, are self-referential. They refer to the core person and indicate that

the conflict that is being experienced is a potential threat to the self. These two considerations drive our search for key neural activations that are related, in a central manner, to metacognition.

Introduction

Metacognition may be thought of in terms of a system in which underlying basic level processes--both cognitive and emotional-- give rise to reflective self-relevant phenomenal experiences that comment on these underlying processes. The phenomenal experiences, in concert with the fact that they are considered to be focally relevant to the person experiencing them, are poised to give rise to corrective action, whether that action is overt bodily movement or covert cognitive process. The fact that the person consciously perceives these states[1] is important in allowing the ensuing action to be freely determined and to potentially produce a

[1] The state that we are arguing is conscious is the phenomenological feeling inherent to the conflict experienced upon monitoring. This feeling may or may not have a verbal label, but if it does, it is not the label that is the essence of the consciousness that interests us. We are not proposing that metacognition implies that the basic-level representation that is monitored is necessarily conscious. Indeed, in several classic cases, such as TOTs, and the feeling of imminence preceding an 'aha' experience, the individual does not have conscious access to the sought-after definition or to the insight solution. Nevertheless, the person is conscious of the metacognitive feeling of being in a TOT. Indeed, those feelings impel the subject to rally the effort needed to bring the unconscious referent into consciousness. Our view that the feeling state related to the metacognitive monitoring is conscious does not imply, then, that the representation that is being so monitored need itself reportable or access-conscious (see, Charles, Van Opstal, Marti, & Dehaene, 2013, for an interesting further discussion of the relation between metacognition and consciousness of the referent or basic-level representation).

change in the individual's current knowledge or behavior. This result of metacognitive monitoring is usually called control.

Spontaneous metacognitive awareness happens when something changes, when something is unexpected, or, when something goes wrong. This can occur under various circumstances: when the answer does not come smoothly to mind as it usually does; when one was sure one was right and finds out that one was wrong; when comprehension breaks down; when a discrepancy is found; when the feeling state is mismatched to the cognitive knowledge state. In such situations, the conflict imposes itself on consciousness as metacognitive phenomenology that urges a change in action. Otherwise control is not needed, and the system continues in a business-as-usual manner. So, if one were riding a bike along the path smoothly, or reading a text with full comprehension, or effortlessly retrieving from memory, these processes could continue to occur without any kind of metacognitive awareness or intervention. But when something goes wrong, the individual spontaneously becomes metacognitively aware, a mental state that may initiate corrective control.

Interestingly, although we posit that self-reflective metacognitive capabilities may have arisen evolutionarily to detect threats and provide a means for self-control under these circumstances, the laboratory data suggest that this same reflective system can be recruited under more mundane circumstances, to provide judgments even in the absence of salient irregularities. We appear to be able to use this reflective system not only to monitor and control what we do not know, but also

what we do know (see Dunlosky & Metcalfe, 2009). Nonetheless, when discrepancies occur, conscious metacognitive states arise *spontaneously*, occurring without the explicit experimental instructions that are often given in laboratory experiments needed to elicit metacognitive judgments. Examples of this include tip-of-the-tongue states, noticing that one's mind is wandering, and states such as déjà vu (e.g., Schwartz & Cleary, this volume). Although some of what researchers have studied as metacognition does not fit this description, it is possible to interpret most metacognition data within this framework. We suggest that consciously-accessible strong judgments of learning, high comprehension, or even emotional understanding of a situation may rarely surface spontaneously in the real world, although one can request and readily receive such judgments in the lab. These "everything is fine, so carry on as usual" states may be the default, and require no action or change. The states urging action that arise spontaneously, we suggest, are only those that indicate that the knowledge base is wrong, that one has misunderstood, that one does not know what one thinks one knows, or one is unable to smoothly and fluently retrieve the solution one needs.

In this chapter, we review where these self-reflective processes might occur in the brain. Because the basic-level processes on which metacognition operates may be quite varied, including perception, memory, emotional understanding, and problem solving, 'metacognitive' tasks often involve multiple neural networks throughout the brain that are convergent only at a high level. It is this divergence of the underlying tasks that makes the study of the neural bases of metacognition so

complex. It is difficult to distinguish the components that involve the monitoring and control from the basic-level non-reflective cognitive/emotional processes. One of the chief functions of metacognition, though, is to detect irregularities, errors, threats or discrepancies in ongoing cognition/emotion. We propose that the anterior cingulate cortex (ACC), an area of the brain implicated in many monitoring processes and involved in the detection of conflict is important in the neural underpinnings of metacognitive qualia. We also suggest that the medial prefrontal cortex (and most focally, Brodman Area 10) ,as well as the precuneus and perhaps the insula, is crucially involved in high level monitoring/control and, in particular, in the self-referential aspect of human metacognition (see http://en.wikipedia.org/wiki/Brodmann_area for an illustration of these areas). This self-referential aspect of thinking about thinking, of course, goes all the way back to Descartes' (1637, 1641; and see 1998 translation) meditations. We hypothesize that this particular reflective system is an indicator of a kind of mental experience that alerts us to the need to change whatever we are doing, because (a) things are not going well (ACC), and (b) the self (BA10) is at risk.

Machines and human primates

In the 1984 movie, "The Terminator," the eponymous cyborg played by Arnold Schwarzenegger, "sees" the world on an internal screen, which lists in print format the options as to what to say in any particular circumstance. In responding to an angry hotel manager, the cyborg selects the most aggressive line to speak from a list of generated lines, to the delight of male teenagers ever since. Though not the

concern of the movie, this scene exemplifies many of the philosophical problems with artificial intelligence (see Brown & Decker, 2009). How does the cyborg see the list of options – *who* is watching that internal screen, if, indeed, anyone is? How does the cyborg "know" what is the most aggressive thing to say? And who is doing the deciding anyway? Is there a conscious being in the cyborg or is he simply following mechanistic protocols? What would indicate a feeling self inside the machine? One of the reasons that philosophers have been interested in metacognition is that it has been thought that the ability to engage in this kind of reflective thought implicates a self or a mind; its presence might provide a litmus test for personhood. The conscious feeling states of conflict--so often accompanying metacognitive states--appears to have the purpose of stimulating corrective action in the service of protection of the self, and so this feeling state, in the presence of reflective ideation, may indicate some kind of self-awareness. Indeed, Humphrey (2006, 2011) has argued that the internalized reification of a self--which perhaps only humans and a few other animals have in a truly conceptual form--evolved to further the protection of our bodily selves. A concern with and desire to protect the so-called 'self'-- in a conscious and premeditated way--has obvious survival advantages.

The behavior of the "Terminator", in the first film, at least, indicated no such 'self'-protective behavior whatsoever. The Terminator carried on, unflinchingly and inflexibly, toward its murderous goal despite grotesque amputations of most of its physical self, in a manner reminiscent of the digger wasps (*Philanthus triangulum*),

reported by Tinbergen (1975). Their fixed action patterns are unrelentingly directed by a signaling stimulus determined by instinct, just as the Terminator's rigidly determined behavior appeared to be attributable to mindless preprogramming that precluded metacognition, choice, and flexibility. But does metacognition-- the ability to know what we know and do not know-- in and of itself, imply a self? If so, then we may have to avouch selfhood to rhesus monkeys insofar as they have demonstrated metacognitive capabilities (see, Hampton, 2001; Kornell, Son & Terrace, 2007). Some dolphins, scrub jays, orangutans, and elephants might also qualify, if we are lenient about the criteria (Kornell, 2014). On such grounds, we should probably also grant selfhood to Watson the IBM computer who resoundingly beat his human competitors at the game of Jeopardy using what is patently a metacognitive strategy (see Metcalfe & Son, 2013). We make the argument here, though, that mere monitoring of basic level cognition--a function readily captured in computer programs, and that may exist, at least in an elementary form in some non-human animals-- is not enough. Additionally *experiencing* the metacognitive states (i.e., the qualia) and referencing them to a *self* concept in an effort to promote the well-being of that self, is what is crucial in *human* metacognition (Aizawa, 2010).

Our starting point for understanding the neural basis of metacognition, and for teasing apart truly metacognitive processes from other cognitive/perceptual/emotional processes--as alluded to above and following Nelson and Narens (1990)--is that metacognition depends on but is not co-extensive

with lower level or 'object-level' processes (see Dunlosky, Mueller, & Thiede, this volume). Examples of object-level processes include retrieval of a word from lexical memory, recognition matching processes, perceptual detection processes, or computations that indicate the expressive valence of a particular perceived face. Monitoring processes, which are at the meta-level, then "observe" those object-level outputs. Such observation need not imply a homunculus. It is enough that a decision or commentary can be made about the object-level outputs or representations. A metacognitive monitor provides information about whether the object-level process was successful or not and whether, if it was not successful, it is likely to later be successful. That kind of information can serve a purpose, allowing the individual to alter its own learning and behavior. This internally generated control potentially frees the individual from being entirely stimulus bound and allows him or her to be self-determining.

Separating object-level and meta-level processing

We propose that a primary reason that experiments investigating the brain basis of metacognition have seemed so confusing and difficult to parse is that many if not most of these studies have conflated first order, or basic level, and second order, or meta-level, processes. Thus, much of what has been imputed as providing the neural signature of metacognition has, in fact, been the neural signature of the underlying basic cognitive processes (such as memory, semantic knowledge, perception, etc.). The activations attributable to these more basic processes have all too often obscured those of the monitoring and control processes that are

distinctive to metacognition proper. In the analysis that follows we attempt to isolate the monitoring and control processes.

That first order (object level) and second order (meta-level) judgments can be dissociated provides support for the idea that metacognition represents unique processes. Research, starting with Shimamura and Squire's seminal (l986) study on Korsakoff patients, has found that frontal impairments can impact metacognitive judgments independently of object level performance. Shimamura and Squire found that temporal-lobe amnesiacs showed spared metacognitive performance when feeling-of-knowing accuracy was measured, but that Korsakoff patients', who were also amnesic, exhibited feeling-of-knowing judgments with no predictive value whatsoever. Other dissociations have been observed. For example, Izaute and Bacon (2006) reported object and meta-level dissociations resultant upon the administration of the amnestic drug lorezapam. Rounis, Maniscallco, Rothwell, Passingham and Lau (2010) showed that TMS (transcranial magnetic stimulation) to prefrontal cortex selectively impaired metacognition of visual awareness. Fleming, Weil, Nagy, Dolan, and Rees (2010) demonstrated individual differences in competence at the basic and meta-level (also see Fleming & Dolan, 2014).

In an elegant experiment isolating the neural correlates of the meta-level, as distinct from basic level processes, Fleming, Huijgen, and Dolan, (2012) had participants conduct a perceptual judgment task in which on two-thirds of the trials they were asked to make a retrospective confidence judgment about whether they had been correct or not on the just completed perceptual decision. On one-third of the trials they did the perceptual task, as before, but then just moved the cursor to

mimic the motor movements they would have had to make in the metacognitive task. The contrast between making and not making a metacognitive judgment revealed increased activity in BA10 (specifically right lateral PFC), ACC, and right posterior parietal cortex. Interestingly, activity increases in these regions were greater for low than for high confidence trials, as if the low confidence judgments were more demanding of reflective processes. Additionally, the authors found that the magnitude of the anterior frontal activation was correlated with individual differences in the goodness (e.g., accuracy of predicting performance) of participants' metacognitive judgments. In essence, Fleming et al., (2012) separated the effect of underlying cognitive performance from metacognitive differences. Moreover, they also quantified cortical mass and observed that the size of mPFC (BA10) positively correlated with the accuracy of participants' metacognitive monitoring.

McCurdy, Maniscalco, Metcalfe, Liu, de Lange, and Lau (2013) replicated Fleming et al's findings concerning visual metacognition —again pointing to mPFC as a core metacognitive area. In addition, though, they conducted a parallel metamemory test. In one phase of the experiment, as in that of Fleming et al (2012), participants had to decide which of two patterns contained a Gabor grating and then make a confidence judgment on their decision. In a second phase, though, McCurdy et al. (2013) had participants perform a metamemory task. The structure of the task was similar to the perceptual metacognitive task: participants made a 2-alternative forced choice verbal memory decision about just-studied words. This memory

decision was followed by a meta-level confidence rating concerning the basic-level decision. They found, as had Fleming et al, that there was a correlation between the volume of the frontal polar region and the goodness of metacognition in the perceptual task. Furthermore, this correlation was found with the metamemory task. In addition, though, memory metacognition was positively related to precuneus volume. Interestingly, the authors noted that there was a correlation across participants between the volume of the medial prefrontal cortex and the volume of the precuneus, suggesting, perhaps, coevolution of these linked areas.

A number of other behavioral and imaging studies have also dissociated the object-level content and process (such as memory retrieval) from the meta-level content and process, such as TOT experiences (see Chua et al, 2014). For example, in an fMRI study, Do Lam et al. (2012) found a dissociation between the processes underlying monitoring and predictions, which were located in the mPFC, and those associated with object-level memorial processes, which were located in the medial temporal lobes. Maril, Simons, Mitchell, Schwartz, and Schacter (2003) found that the retrieval of previously studied items (i.e., basic-level memory processes) was associated with activity in the hippocampi and medial temporal lobes but that differences in the magnitude of feeling-of-knowing judgments (meta-level processes) were associated with a variety of activity in different areas of the prefrontal lobe. Unfortunately they did not have a no-judgment condition, so they were not able to directly evaluate the effect of monitoring versus not monitoring. Nor did they examine the neural correlates of differences in the goodness of

monitoring (see Table 1 for an overview of the research).

Metacognition and Phenomenal Experience

Certain spontaneous metacognitive states are accompanied by distinctive and recognizable conscious feelings. Examples of such metacognitive experiences include, for example, the feeling of imminence that precedes the classic "aha" moment in which an individual realizes--against all expectations-- that the problem is solved. The unsettling feeling we undergo when we experience déjà vu comprises another spontaneous metacognitive state. The nagging tip-of-the-tongue state that we endure when retrieval does not smoothly produce the answer we seek provides another well-known and universally experienced example of a metacognitive experience (Schwartz, 2006). The processes needed, for example, for a déjà vu experience to occur, include several conflicting sources of information, each with its own neural correlates. A visual scene must be familiar; the person must infer through a conscious decision-making process that in fact the scene is new; and, the individual must strongly believe both of these conflicting sources of evidence are accurate (Cleary et al., 2012, and see Schwartz & Cleary, this volume). When these processes co-occur, they converge to create the déjà vu experience. Similarly, the tip-of-the-tongue experience occurs when there is a conflict: an item that one expects to be able to retrieve effortlessly and quickly from memory is not forthcoming. The nagging unpleasant aspect of the phenomenology results from the conflict between the person's expectations about the ease of retrieval and the unanticipated difficulty in retrieval in the particular case. This conflict between

knowing that one knows and being unable to produce the concrete evidence that one knows in the form of the correct answer urges the subject to further attempts at retrieval (or perhaps to look up the answer on our iPhones).

The anterior cingulate cortex (ACC) component and detection of discrepancy

A preponderance of research supports the view that the ACC responds to mismatches in both the physical and internal environment. In both perceptual and cognitive domains, it becomes active when the individual expects one outcome, but another outcome occurs. The classic situation is when an error has been committed as compared to when no error occurred (Bush et al, 2000; McGuire & Botvinick, 2010). Conflict between expectations and perceptual events give rise to ACC activation. For example, it is the ACC that becomes active when color word names are presented in a mismatched color in the Stroop task (Posner & DiGirolamo, 1998). McGuire and Botvinick (2010) showed that the ACC is more active when decisions have to be made that have greater costs associated with them relative to when the same decisions are made without the great cost. When a trained response must be overridden because penalties are now associated with the original response, the ACC becomes active. The ACC is also activated on a trial-by-trial basis as participants make subtle adjustments to continue doing a task successfully. When we look to the neuroimaging work done on creativity, we find that the ACC is correlated with creative problem-solving, which may also involve the perception of a mismatch (Fox & Christoff, 2014).

The ACC has also been linked in a variety of ways to conscious experience,

especially emotional conscious experience (Gray, Bargh, & Morsella, 2013; see Shushruth, 2013 for a discussion of the role of ACC in anesthesia). First, damage to the ACC can result in a condition called akinetic mutism, in which patients will not move or talk, and exhibit extreme apathy. They are, apparently, registering events but are not consciously aware of a motivating feeling state. Interestingly, in this condition, not being consciously aware of a feeling state--presumably of conflict-- is associated with a lack of movement, motivation, or 'will'. Holroyd and Yeung (2012) describe a patient who upon regaining ACC function described her state during akinetic mutism as having nothing in “mind” and no “will” to do anything.

Of course, this feeling of 'no will' may refer to the lack of a feeling of discrepancy or conflict that would normally be driving one to seek resolution (rather than to a lack of a motor plan necessary to direct willful action, which may be the role of the pre-supplementary motor area). This lack of feeling of 'caring' or 'lack of caring about pain' has led some to conjecture that this kind of consciousness is related to motivation, and when the ACC is damaged, motivation is impaired as well. The most prominent clinical symptom of bilateral cinguloctomy is apathy--people do not care about what goes on around them. In support of the motivation view of ACC, Stuss et al. (2005) showed that damage to the ACC results in slowed responses in tasks in which motivation is required.

The ACC is also linked to emotional regulation, the experience of emotion, and particularly to emotional conflict (see Efklides, this volume). For example, Sturm et al. (2013) looked at the neural correlates of embarrassment. Healthy

controls and patients with frontotemporal dementia engaged and then later viewed themselves performing an embarrassing karaoke task. Sturm et al. compared the fMRI pattern of participants watching a videotape of themselves singing to a control task, in which they watched a sad segment from a movie. In both cases they watched video – but in one case it was of them in another case, it was of others. In both healthy controls and the patients, the ACC was the area that showed the greatest correlation to self-conscious embarrassment, that is, to watching the self-karaoke film than while watching the control film. This finding supports three independent contentions about the ACC – its link to negative emotion, its link to detecting surprising situations, its link to pain, and the fact that people are conscious of these feelings when the ACC is activated.

So with these studies in mind, we consider the role of the ACC in metacognition. There have been few neuroimaging studies examining metacognition, but, even so, ACC activity is salient among them. Jing, Niki, and Philips (2004) examined the cognitive processing that people engaged in prior to an "aha" experience, the feeling one gets when one suddenly, after considerable conflict and unease, understands the solution to a problem. Activity in the ACC was correlated with the difficult and challenging processing required before the experience of the delightful "aha" reaction but not with other more mundane and less conflict-ridden mental states during routine problem solving. These data support the conclusion that the ACC is involved in metacognition and signals conflict (Holroyd & Yeung, 2012).

Metcalfe, Butterfield, Habeck, and Stern, (2012) investigated confidence judgments in semantic memory question answering. Previous researchers (e.g., Butler, Karpicke, & Roediger, 2008; Butterfield & Metcalfe, 2001) observed what is called a 'hypercorrection effect', that is, that there is greater recall of correct answers to questions that were initially answered incorrectly with high rather than low confidence. This finding contrasts with the predictions of many theories of memory that indicate that answers produced with high confidence should be exceedingly difficult to overwrite or update. In the experiment, participants were presented with general-information questions, such as "What is the highest mountain in Europe?" (The answer to which, in this case is Mount Elbrus, in Russia). Participants answered the questions outside of the scanner, giving their confidence about each answer, but received feedback about their answers later while being scanned by fMRI. Then, later, a retest was given outside of the scanner.

During the feedback phase, items were selected from the pool of questions such that some high confidence but incorrect answers were always included (e.g., Mt. Blanc, high confidence, but incorrect). Brain activation was contrasted between the feedback event to questions that had been answered erroneously but with low confidence (which would evoke relatively little conflict) and to wrong answer given with high confidence (which would evoke considerable conflict). In addition, the authors investigated brain activation during the feedback provided to low confidence correct answers (which show a mismatch between expectation and feedback) as well as during feedback to high confidence corrects (in which no conflict would, presumably, be experienced). The study revealed that ACC

activation (in both the left and right hemispheres) was correlated with the mismatched conditions--during feedback to high confidence errors and to low confidence corrects (see Figure 1). The ACC was much less active when recall was correct and confidence was high, or when recall was incorrect and confidence was low. Thus, the ACC data do not reflect either high confidence or correctness, but rather the mismatch between the two. In addition, this mismatch was also correlated with activation in the medial frontal gyrus. Like the ACC, the medial frontal gyrus is an area thought to be involved in the conscious monitoring of emotional states (Phan, Wager, Taylor, & Liberzon, 2002), in metacognitive states (Fleming et al., 2010), and, as we shall elaborate shortly, when the processing at hand is self-relevant. The right dorsolateral prefrontal cortex was associated with error suppression, consistent with its role in metacognition (Schwartz & Bacon, 2008). The fMRI data from this study showed patterns of activity consistent with the idea that the ACC is important in monitoring and eliciting conscious awareness of surprising events, and underlies some of the control aspects of the metacognition/control function.

Figure 1. goes here. Figure 1 from Metcalfe et al (2012).

We turn now from retrospective confidence to a prospective metamemory judgment, in this case, judgments of learning. Judgments of Learning (JOLs) are judgments that predict future retrieval of items currently under study. JOLs may either be made when viewing both the cue and the target or when seeing the cue alone. Typically, predictive accuracy is higher when the target is not present, as

participants can use current recall as a basis for predicting future recall (Dunlosky & Metcalfe, 2009; see Rhodes, this volume). JOLs have been examined using fMRI technology in a few studies, and the results conform to the pattern already described – that is JOLs are associated with activity in the medial pre-frontal cortex. Moreover, we see ACC activity when there is an opportunity to experience conflict between the judgments and the prospect of recall. Do Lam et al (2012) conducted a JOL study that looked at face-name associations and then tested later when showing only the faces, thus allowing for the possibility of perceived conflict between a high judgment of learning coupled with an inability to recall. Participants viewed photographs of faces, randomly paired with gender-appropriate names. Immediately after study, the face was presented alone and the participants made a JOL about later recall of the target name. Recall was then assessed 4 seconds after the JOL. All phases were done while in the scanner. Do Lam et al found that memory performance was correlated with bilateral activity in the hippocampi. JOLs, though, were correlated with activity in the ACC, as well as the orbitofrontal cortex and medial prefrontal cortex (see Figure 2)

Figure 2 goes here: from Do Lam et al (2012)

Although Do Lam et al (2012) did not address this issue in the study, we suspect that the ACC activity was mediated by the surprise participants may have experienced when, during the JOL phase, they could not recall a name that they had just seen seconds earlier. If this were the case, the ACC activity was registering the conflict in the task, consistent with the hypotheses of this chapter. This view may

also explain why not all JOL studies show ACC activity.

Kao, Davis, and Gabrielli (2005) also examined JOLs in the scanner, but the task was to make a JOL concerning the potential recognition of a picture that would be provided later. In this study, participants viewed images and rated - while the images were still present - whether they would recognize the images later when they were seen among new images. Insofar as the images were present when participants made their judgments they did not have the opportunity to find that they were unable to retrieval items they thought would be retrievable, and be surprised by their failure to do so. Thus, in the immediate JOL paradigm, illustrated by this experiment, no conflict is experienced between the person's knowledge that s/he 'should' be able to retrieve and their failure. During the judgment phase, Kao et al. (2005) found activity in other areas of the prefrontal lobe (as well as in the posterior parietal lobe), but not in the ACC.

TOTs are feelings of future retrievability, confounded by frustration in the present lack of recall. We have argued elsewhere that TOTs are metacognitive feelings that monitor our potential knowledge and drive us to further retrieval efforts (Schwartz & Metcalfe, 2011; 2014). Here we advance the notion that phenomenologically TOTs reflect conflict detection-- the conflict between the lack of recall and the confidence that recall is imminent. In this framework, TOTs are metacognitive experiences that notify us of this mismatch. Because we become aware of the conflict, TOTs can then direct us to control retrieval behavior in appropriate ways. Given that the TOT state illustrates an internal conflict that

propels behavior, we would expect activity in the ACC.

Maril et al. (2005) gave participants cues such as "Carmen, composer," and the participants were expected to generate the target (e.g., "Bizet"). This recall test was given during fMRI scanning – so responses were limited to indicating that (1) they had recalled the answer, (2) they did not know the answer, or (3) they were in a TOT for the target. Accuracy of recalled answers was verified later outside the scanner. Maril et al compared the brain activity across these three conditions. They found that the areas of the brain uniquely activated during TOTs were mostly in the right prefrontal lobe including, prominently, the ACC, the right dorsolateral prefrontal cortex, and the right inferior prefrontal cortex, similar to the areas seen by Metcalfe et al (2012) though using a different methodology (see Figure 3).

Figure 3 From Maril et al, 2005

TOTs are sometimes considered to stem from problems with retrieval from lexical memory (Bacon, Schwartz, Paire-Ficout, Izaute, 2007; Brown, 2012). As such, it is noteworthy to consider TOTs in connection to language as well as metacognition. In this vein, we find that there is converging evidence from the neuroscience of language that also supports the view that the ACC is involved in conflict monitoring in metacognition. This evidence comes from recent studies on the frontal aslant pathway (Catani et al., 2013). The frontal aslant pathway is a newly-discovered white-matter tract that goes directly from Broca's area to the ACC. Damage to the frontal aslant pathway is associated with dysfluency but not grammatical impairment in aphasics (Catani et al., 2013). Although there has not

yet been a study looking at changes to the frontal aslant pathway during TOTs, we suggest that damage to this pathway might explain the neuroimaging TOT correlations. The ACC may receive distorted signals, which may increase TOTs, even as retrieval is impaired by the damage. We wonder if changes in this pathway cause the nearly constantly frustrating TOT feelings in aphasics, particularly anomic aphasics, who cannot retrieve sought-for words, but have high confidence that they know them (Funnell, Metcalfe, & Tsapkini, 1996). We can only speculate on the pathway here, but the potential relation between the frontal aslant pathway and TOTs is intriguing.

In contrast to the ACC activations associated with being in a tip-of-the-tongue state, feeling-of-knowing judgments do not elicit ACC activity (Chua et al., 2014; but see Maril et al, 2003 for an exception). Feeling-of-knowing judgments are typically predictions of future recognizability, often of items not currently recalled. Although TOTs and feeling-of-knowing judgments are superficially similar judgments, the phenomenology of the two is different, and behavioral studies have dissociated them (Schwartz, 2006; 2008; see Thomas, this volume). For example, divided attention lowers the number of TOTs, but does not affect feeling-of-knowing judgments (Schwartz, 2008). This counterintuitive dissociation between TOTs and feeling-of-knowing judgments is supported by a difference in neural patterns. In feeling-of-knowing judgments activity in medial prefrontal cortex is seen (as are basic-level activations), but ACC activity is not a consistent correlate of feeling-of-knowing judgments (Chua et al, 2009; Kikyo, Ohki, & Miyashita, 2002; Kikyo & Miyashita, 2004). This may be because feeling-of-knowing judgments are a more

analytic judgment, made based on predictions of future recognition, but not drawing on the mismatch between confidence in the existence of knowledge in semantic memory paired with the inability to express that knowledge.

These studies support the view that ACC activity occurs across a variety of metacognitive judgments. We propose here that the ACC plays a role in metacognition in which it signals a mismatch between parallel streams of cognition. The mismatch leads to a metacognitive experience, which leaps into consciousness to alert us that our confidence is not matched to our performance. This allows us to engage in behaviors to resolve the conflict, the putative function of this system. For TOTs, that resolution comes from continued search either internally (that is, from memory) or externally (finding outside sources, such as Google that might know the answer). For mismatched confidence, we initiate a hypercorrection process that allows us to recall the correct answer to previous high-confidence incorrect answer (Metcalfe et al., 2012).

The ACC's function of detecting inconsistencies between different channels of information extends beyond the metacognitive and cognitive domain. For example, the ACC is involved in pain perception, in the regulation of the somatosensory system, in emotional regulation, and in basic autonomic functioning (Davis et al., 1999). Indeed, with respect to pain, research shows that the damage to the ACC does not eliminate the pain, but it eliminates our concern or worry about the pain (Price, 2000). On the surface, these functions may have little to do with the experience of mismatch seen in the metacognition studies. But we propose that

what all of these seemingly disparate functions have in common is the causation of conscious experience that alerts us to conflict and spurs us to action. For example, pain may be a low-level sensation primarily triggered by the activation of free nerve endings in the skin, but concern with pain allows it to function to induce the organism to seek cover or remove itself from a dangerous stimulus. Nothing gets us to the dentist faster than worry about a painful toothache. Similarly, when we are highly confident that Sydney is the capital of Australia and then are told it is actually Canberra, this is surprising and potentially important knowledge. It may also be embarrassing, or somewhat painful to be wrong. Thus, a conscious experience of unpleasant surprise about our current state of knowledge may help us to rally to recode our semantic memory. Moreover, experiencing a TOT results in the action of continued retrieval attempts--we try harder (Litman et al; 2005; Schwartz & Metcalfe, 2013). The conscious experience is the inducement to get us moving towards these goals.

This view of ACC function is consistent with recent accounts of the role of ACC in cognition, even though those models usually do not emphasize its role in formally defined metacognitive judgments (Botvinick, 2007; Holroyd & Yeung, 2012). For example, Holroyd and Yeung place the ACC at the top of a network for both monitoring performance and planning action (i.e., control).. They then hypothesize the ACC communicates information to the dorsolateral prefrontal cortex, which is responsible for executing intended actions. Then other areas of the brain, such as the orbitofrontal cortex sustain actions until the ACC gives it feedback that the action is no longer necessary. Holroyd and Yeung advanced this model to

reconcile neuropsychological finds that show that damage to the ACC can slow cognitive processes, leading to akinetic mutism with the neuroimaging data which show that the ACC is involved in cognitive control and conflict monitoring. Like Holroyd and Yeung, we also see the ACC as near the top of a complex neural network, being involved in monitoring conscious experience and triggering cognitive action.

The central involvement of the self in metacognition

The notion that the self --the source of consciousness as some would have it-- is focally involved in metacognition is, in large part, what makes pursuit of the neural correlates of metacognition in humans so fascinating. Although virtually all neuroscientists disagree with Descartes' dualism as implying a mind/body separation and a self that has no physical instantiation --that " has no need of place nor depends on any material thing " (Descartes, 1637, tr. Cress, 1998, p. 19), we might nevertheless take seriously his observation that something that we might characterize as 'the self' or the 'I' is focally involved when we reflect upon or contemplate our memories, capabilities, traits and limitations, as he did in his famous meditations. It is easy to see that the contemplation and commentary on memories and other mental capabilities--the reflection upon cognition-- is a central part of what all researchers agree to define as metacognition. It is controversial whether judgments indicating uncertainty are necessarily at the meta-level or just refer to object-level processes. But thoughts that are specifically about thinking itself (see Nelson & Narens, 1990), or judgments about internal representations at the cognitive level (i.e., second order assessments of first order representations or

processes) are accepted by all as being metacognition. We will focus, in this section, on attempts to narrow down the neural correlates of these meta-level processes, and attempt to determine whether self-reference is entailed in such judgments.

The medial prefrontal component in metacognition

The medial prefrontal cortex (mPFC), with a particular emphasis on Brodmann area 10 (Baird, Smallwood, Gorgolewski, & Margulies, 2013; Fleming et al., 2010; Fleming & Dolan, 2012), along with the precuneus and perhaps the insula, appear to be focally implicated in metacognitive processing. We hypothesize that this brain system--so often seen in studies of metacognition-- is associated with the self-referential characteristics of metacognition. A number of studies support this view, but before we get to the involvement of the sense of self in metacognition, we will briefly summarize the neuroimaging data that indicate a correlation between mPFC activation and metacognition.

Many of the same studies that show ACC involvement in metacognition also show mPFC involvement. For example, Do Lam et al (2012) examined neural correlates of judgments of learning. They found that mPFC was more active during the making of judgments of learning than it was during study or retrieval. Similarly, in the already-discussed studies on the neural correlates of TOTs, mPFC was implicated (Maril et al, 2001, 2005). Interestingly, we see mPFC activity even in metacognitive judgments, such as feeling-of-knowing judgments, that do not also activate the ACC. For example, Maril et al (2003) found that mPFC was active during high feeling of knowing relative to “don’t know” states. Similarly, Maril et al (2005)

found heightened ACC activity only during TOTs, but not during feeling-of-knowing judgments. But both metamemory judgments were correlated with mPFC activity. Similarly, Schnyer, Nicholls, and Verfaellie, 2005, also showed that mPFC was active during high feeling-of-knowing judgments in the absence of ACC activity, as did Kikyo et al, 2002, and Kikyo and Miyashita, 2004. Clearly then, mPFC activity--taken broadly-- is involved in metacognition.

Moreover, in other studies already discussed, Fleming et al (2010) found that structural differences in Brodmann Area (BA) 10 in the mPFC was associated with individual differences in the accuracy of metacognitive judgments. Similarly, Fleming et al's (2013) perceptual metacognition experiment pointed to BA10 (specifically rlPFC, as the current discussion indicates), ACC (as expected from the preceding section of this chapter) and also right posterior parietal cortex (which was probably involved in basic level aspects of the task), as being selectively implicated when the person was making a metacognitive judgment. Notably, the extent of the anterior frontal activation was central and was related directly to the accuracy of the judgments. Furthermore, McCurdy et al (2013) replicated Fleming et al's findings that individual differences in the structural size of anterior frontal areas was related to the goodness of the judgments people made when performance on the object level task was equated.

The Medial PFC and the self

Interestingly, these medial prefrontal areas that are often activated in metacognitive judgments, are also prominent in studies investigating self-attributions. The possibility that introspective meditation about our cognition or

emotion is linked to self-reflective consciousness finds some support in other paradigms not devised to particularly assess metacognition. People's attribution of particular characteristics to the self, their comparisons to the self, and high level integrative and abstract processing which might be thought to implicate a unifying self-principle activate the same brain areas, as do metacognitive judgments

There have been many studies investigating the neural correlates of the 'self,' outside of the context of metacognition. For example, following Craik, Moroz, Moscovitch, Stuss, Winocur, and Tulving's (1999) pioneering positron emission tomography study showed selective frontal activation during self-relevant mental processing. Kelley, Macrae, Wyland, Caglar, Inati and Heatherton (2002) compared self attributions to semantic judgments. A self-attribution is something like, "I like to travel," whereas a semantic judgments is akin to "Italy's economy benefits from tourism." Kelley et al found mPFC activation specifically when people made self-relevant trait attributions --dissociating such self-related processing from non-self related semantic processing. Similarly, Johnson, Baxter, Wilder, Pipe, Heiserman and Prigatano (2002) found evidence for mPFC and posterior cingulate activation while people were answering questions (such as "I am a good friend" or "I have a quick temper") directed at their own self-awareness. Ochsner, Knierim, Ludlow, Hanelin, Ramachandran, Glover, and Mackey (2004) found that both self and other judgments activated mPFC. Other judgments also showed activation in more lateral regions of PFC and medial occipital cortex; self-judgments were also related to left temporal activation. As they noted, however, the mPFC was central in all of the attributions that were of *relevance* to the self.

Jenkins and Mitchell (2011) had participants contemplate various different aspects of themselves, such as their own personality traits, their current feeling states, and their physical attributes. The mPFC responded during *all* of these self-relevant judgments. Areas other than the mPFC responded selectively to particular self-relevant mental states, however. For example, the temporal parietal junction was also active when people made responses concerning their own transient mental states. The intraparietal sulcus and the caudate were activated when people considered their own personalities. Reflection on one's own physical attributes resulted in activation of the cerebellum. But in all cases, regardless of the type of self-relevant reflection, mPFC was activated.

Ochsner, Beer, Robertson, Cooper Gabrieli, Kihlstrom and D'Esposito (2005) investigated two ways in which the self can be known: through direct appraisals (i.e., an individual's own self-beliefs) and through what they called 'reflected' appraisals, which were the person's perception of how others viewed him or her. The authors contrasted self and other evaluations. Whether appraising the self or close others and regardless of whether making direct appraisals or reflected appraisals, though, they found that all self-relevant judgment activated mPFC. Direct appraisals of the self as compared to others more strongly recruited mPFC and also right rostrolateral PFC, whereas reflected appraisals recruited emotion-related and memory-related areas such as the insula, the orbitofrontal cortex and the temporal cortex along with mPFC.

Whereas the mPFC is consistently implicated in reviews of the neural correlates of self, some meta-analyses have pointed to other, more posterior medial

areas. Northoff, Heinzel, de Greck, Bermpohl, Dobrowolny and Panksepp (2006) conducted a meta-analysis of 27 PET and fMRI studies on self-related tasks (including some visual motor tasks, presumably included because they require motor control, perspective taking tasks, matching tasks, and other tasks that may have involved the self in some way but did not involve self-related judgments). They pointed to three clusters (including a mPFC/ACC cluster) that they call the Cortical Midline Structures system that, they argue, allows the transformation of the so-called proto-self into a core mental self that is foundational for continuous self-referential processing of the "stream of subjective experience" (p. 451).

Denny, Kober, Wager, and Ochsner (2012) performed a more recent meta-analysis of 107 imaging studies directed at self-relevant processing. They found that mPFC was involved both when people made judgments of self and relevant others. In addition, though, ventral mPFC, left ventrolateral PFC and left insula were activated by self judgments, whereas dorsal medial PFC, bilateral Temporal Parietal Junction and precuneus, tended to be activated by judgments about the other.

It would appear, from the foregoing discussion, that mPFC (especially, BA10) is implicated in metacognition, and that it is also implicated in just about everything having to do with the self. Although there have been many studies that have shown the role of mPFC in metacognition, only one study has directly contrasted simple metacognition (a reflection or judgment about the object level or what Metcalfe and Son, 2012, called 'noetic' metacognition) and *self-referential* metacognition or what Metcalfe and Son called *autonoetic* metacognition (e.g., Tulving, 2005). Miele, Wager, Mitchell and Metcalfe (2011) conducted an fMRI experiment in which people

played a computer game in which they were sometimes in complete control and sometimes not. After playing for a short period of time, the participants were asked to make judgments of agency or judgments of performance. The former is a direct measure of people's own feelings of control over and responsibility for their actions. It directly refers to their selves. As such, it is both self-referential and autonoetic. The latter judgment is about the goodness of performance and could be either self-referential or not. If the person interprets the judgments as referring to how well *he or she* performed--a query about them personally-- then it could be self-referential. However, if the person interprets it as being only about observable outcomes--what was performance like -- then it could be non-self referential. Interestingly, although judgments of performance are usually strongly correlated with judgments of agency, some populations --including people with schizophrenia and people with Asperger's syndrome-- exhibit excellent performance judgments, but show pronounced impairments in their judgments of agency (see, Metcalfe, et al., 2012; Zalla, et al, 2014). This dissociation indicates that the two kinds of judgments are separable. Miele et al. (2011) found that the difference in activation between these two kinds of judgments showed up in the mPFC (i.e., BA10). Whereas noetic metacognitive judgments activate mPFC, metacognitive judgments that are specifically self-referential "*hyper*-activate" that same area.

Conclusion

So let us return to Descartes, and his claim that a particular kind of thinking that we might consider to be metacognitive in nature (i.e., deliberations about his

own thinking and his certainty therein) irrefutably indicated that he has a Self. Descartes, after considering how his percepts, memories, and deductions could be faulty, claimed that the only thing that was certain and irrefutable was that "I think therefore I am." He elaborated: "From the very thought of doubting the truth of other things, it followed very evidently and very certainly that I existed.... From this I knew that I was a substance the whole essence or nature of which is simply to think." (Descartes, 1637, tr. Cress, 1998, p. 19).

But does this kind of thinking--in which there is a reflection or a judgment (in Descartes' case of doubting) concerning more basic level cognitions-- guarantee that there is an "I?" Bertrand Russell (1967) claims not. He notes that: " 'I think therefore I am' states rather more than is strictly certain. It might seem as though we were quite sure of being the same person today as we were yesterday, and this is no doubt true in some sense. But the real Self is as hard to arrive at as the real table, and does not seem to have that absolute, convincing certainty that belongs to particular experiences. When I look at my table and see a certain brown color, what is quite certain is not the "*I* am seeing a brown color" but rather "A brown color is being seen." ... It does not of itself involve that more or less permanent person whom we call 'I'. " (Russell, 1967, p. 8).

If Russell is right, then metacognition--reflection upon object level cognition-- does not necessarily imply an enduring self. The case is most clear if we return to the example of the Terminator cyborg, or Watson the Jeopardy-playing computer who wins by using a metacognitive strategy. Does the demonstration of the ability

to monitor thought processes necessarily indicate that either of these machines had a self? We think not. Despite Descartes' contention, there is no *necessary* connection between one process monitoring another and the existence of a self. Indeed, it is not even necessary that the monitor be internal to the machine. The hardware underlying the monitoring process in a cyborg in Los Angeles could be in Toledo, Ohio[2]. Similarly, non-human animals, such as monkeys, appear to be able to make some simple metacognitive judgments, a formidable accomplishment that has been demonstrated by Kornell et al (2007), and by Hampton (2001). But if Russell is right, doing so does not, in and of itself, indicate that they also have a 'Self.'

So why was Descartes confused? We think that it is highly likely that Descartes' medial prefrontal cortex was flashing off like a firecracker when he was engaged in his metacognitive meditations. As we have seen from the studies described above, BA10 is co-activated by metacognition and by self-relevant ideation. The two are conflated in the human brain. It is not a logical necessity that it be so. They could occur in entirely different regions of the brain (or one could be in Toledo). But, in point of fact, they co-occur. This conflation of self-relevant processing and metacognition is, we think, an empirical fact about the human brain but it is not a logical necessity inherent to any metacognitive processing. So, we suggest, Descartes got the phenomenology--relating the existence of a self and metacognitive processing such as doubting-- absolutely right. And he wrote it down

[2] Descartes appeared to appreciate this, and noted that the monitor could be quite separate from the body. Indeed, he, the ultimate dualist, went even further to argue that the monitor could not only be external to the body but need have no physical existence.

beautifully and compellingly. His only mistake was positing that the phenomenology-- thinking that having a clear and distinct impression of self-relevance, or as we might rephrase it, undergoing strong and interpretable brain activation in a particular area-- was tantamount to irrefutable truth.

When typical humans do metacognitive processing, then, it appears that they do activate the region of the brain that is also used for self-referential processing. We take our metacognition personally! What about non-human primates? Do they, too, have a self that is related to metacognition via co-activation in their brains? Behaviorally, they have been shown to be able to make some metacognitive judgments, whereas other animals, such as pigeons, are unable to do so (Shettleworth, 2012). Do the metacognitive judgments that the monkeys make activate the monkey analogue to BA10, and do their self-relevant thoughts also activate this region?

To date, only one study has compared resting brain states in humans and monkeys using fMRI, and the main differences between the human and monkey was in BA 10 (Neubert, Mars, Thomas, Sallet, & Rushworth, 2014). Furthermore,, we do know that BA10 is the region that is least developed in other primates, relative to humans. But they do have BA10.

Although other regions of the brain, and even other regions of the frontal cortex, scale smoothly from other primates to humans; BA10 is the exception. It is prominent in humans but, relatively, much smaller in other primates, suggesting that it was a focus of late human brain evolution (Semendeferi et al, 2001). Furthermore, using tractography, Thiebaut de Schotten, Dell'Acqua, Valabregue and

Catani (2012) found an absence of the inferior frontal-occipital fasciculus (which is prominent in humans and projects to BA10) in the monkey brain. They suggested that this tract may be unique to the human brain and that the projections of the inferior fronto-occipital fasciculus to BA 10, in humans, might explain the larger relative size of BA10 in humans than other primates. In addition, Allman, Hakeem, and Watson (2002, and see Allman et al, 2012) reviewed two anatomical specializations of the brains of apes and humans--namely a morphologically distinct cell type--the spindle (or von Economo) neuron in the ACC, and BA10 in the mPFC. They suggested that the spindle cells relay the motivation to act to other parts of the brain but particularly to BA10. Interestingly, although apes (but not other animals) show some development of both ACC spindle cells and BA10, there is a large difference between humans and apes with the order being the same for both spindle neurons and BA10: humans > bonobos > chimps > gorillas > orangutans > gibbons.

The processes involved in metacognition are complex, and even the neural circuitry underlying all of the processing contributing to even the most straightforward of judgments about cognition is complicated. Every study conducted on the neuroscience of metacognition has revealed multiple areas activated by metacognitive evaluations (Chua et al., 2014). But, even so, when the processes involved in what is inevitably a combination of object-level and meta-level processes are analyzed in an effort to isolate those processes that are distinctively at the meta-level, two components converging on ACC and mPFC tend to emerge. The first is a component related to expectation violation phenomenology,

prominent in spontaneous metacognitive feeling states such a TOT states, deja vu states, and hypercorrection phenomena. The second, the medial prefrontal cortex centering on BA10--an area that implicates self-relevant processing--is consistently activated when people are doing metacognitive reflection. Together, these, along with the object-level processing, which varies depending on the object-level task , appear to form an evolutionarily relevant system underlying the integrative, self-relevant, conscious, cognitive and emotional processes used by people in metacognitive reflection.

References

Aizawa, K. (2010). Consciousness: Don't Give up on the Brain. In Pierfrancesco, B., Kiverstein, J., & Phemister, (Eds.) *The Metaphysics of Consciousness: Royal Institute of Philosophy Supplement, 6*. (pp. 263-284).

Allman, J. Hakeem, A, & Watson, K. (2002). Two phylogenetic specializations in the human brain, *The Neuroscientist, 8*, 335-346.

Allman, J. M., Tetreault, N A., Hakeem, A. Y., Manaye, K. F., Semendeferi, K., Erwin, J. M., Park, S., Goubert , V. & Hof, P. F. (2012). The von Economo neurons in the frontoinsular and anterior cingulate cortex, *Annals of the New York Academy of Sciences, 1225,* 59–71.

Bacon, E., Schwartz, B. L., Paire-Ficout, L., & Izaute, M. (2007). Dissociation between the cognitive process and the phenomenological experience of the TOT: Effect of the anxiolytic drug lorazepam on TOT states. *Cognition and Consciousness, 16*, 360–373.

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

Botvinick, M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective and Behavioral Neuroscience, 7,* 356-366.

Brown, A. S. (2012). *Tip of the tongue states*. New York: Psychology Press.

Brown, R., & Decker, K. S. (2009). *Terminator and Philosophy: I'll be back. Therefore I am.* Wiley: New York.

Bush G., Luu P., & Posner M.I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Science, 4*, 215 – 222.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 918–928.

Butterfield, B. & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491-1494.

Catani, M., Mesulam, M. M., Jackobsen, E., Malik, F., Martensteck, A., Wieneke, C., Thompson, C. K., Thiebaut de Schotten, M., Dell'Acqua, F., Weintraub, S., & Rogalski, E. (2013). A novel frontal pathway underlies verbal fluency in primary progressive aphasia. *Brain, 136,* 2619 – 2628.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. Neuroimage, 73, 80 – 94.

Chua, E.,F., Pergolizzi, D., & Weintraub, R. R. (2014) The cognitive neuroscience of metamemory monitoring: Understanding metamemory processes, subjective levels expressed, and metacognitive accuracy. In S. M. Fleming & C. D. Frith (Eds). *The Cognitive Neuroscience of Metacognition.* pp. 267 – 291. Springer: New York, NY.

Chua, E. F., Rand-Giovannetti, E., Schacter, D. L., Albert, M. S., & Sperling, R. A. (2004).

Dissociating Confidence and Accuracy: Functional Magnetic Resonance Imaging Shows Origins of the Subjective Memory Experience. *Journal of Cognitive Neuroscience 16,* 1131–1142.

Chua, E. F., Schacter, D. L., Rand- Giovannetti , E., & Sperling, R. A.(2006). Understanding metamemory: Neural correlates of the cognitive process and subjective level of confidence in recognition memory. *NeuroImage, 29,* 1150 – 1160

Chua, E. F., Schacter, D. L., & Sperling, R.A. (2009). Neural Correlates of Metamemory: A Comparison of Feeling-of-Knowing and Retrospective Confidence Judgments. *Journal of Cognitive Neuroscience 21*, 1751–1765.

Cleary, A.M., Brown, A.S., Sawyer, B.D., Nomi, J.S., Ajoku, A.C., & Ryals, A.J. (2012). Familiarity from the configuration of objects in 3-dimensional space and its relation to déjà vu: A virtual reality investigation. *Consciousness and Cognition, 21*, 969-975.

Cole, M. W., Yeung, N., Freiwald, W. A., & Botvinick, M. (2009). Cingulate cortex: Diverging data from humans and monkeys. *Trends in Neuroscience, 32*, 566 – 574.

Craik, F. I. M, Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E, & Kapur, S. (1999). In search of the self: A positron emission tomography study. *Psychological Science, 10*, 26 – 34.

Crowe, S. F., & Crowe, L. M. (2013). Does the presence of posttraumatic anosmia mean that you will be disinhibited? *Journal of Clinical and Experimental Neuropsychology, 35*, 298 – 308.

Davis, K. D., Taylor, S. J., Crawley, A.P., Wood, M. L., & Mikulis, D. J. (1997). Functional MRI of pain- and attention-related activations in the human cingulate cortex. *Journal of Neurophysiology, 77*, 3370–3380.

Denny, B., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*, 1742-1752.

Descartes, R. (1998). *Discourse on methods and meditations on first philosophy, fourth edition.* (Discourse on methods first publ. 1637; Meditations first publ. 1641). Translated by D. A. Cress, Hackett Publishing Company: Indianapolis/ Cambridge.

Do Lam, A. T.A., Axmacher, N., Fell, J. Staresina, B. P., Gauggel, S., Wagner, T., Olligs, J., & Weis, S. (2012). Monitoring the mind: The neurocognitive correlates of metamemory, *Plos One*, 7, 1 – 9.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

Elman, J. A., Marian, D. E., Klostermann, E., Verstaan, A., & Shimamura, A. P. (2012). Neural correlates of metacognitive monitoring during episodic and semantic retrieval. *Cognitive, Affective, and Behavioral Neuroscience, 12*, 599–609.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society, B, 367*, 1338–1349.

Fleming, S.M., Huijgen, J. & Dolan, R.J. (2012) Prefrontal contributions to metacognition in perceptual decision-making. Journal of Neuroscience, 32(18): 6117-25.

Fleming, S. M., & Dolan, R. J. (2014). The neural basis of metacognitive ability. In S. M. Fleming & C. D. Frith (Eds). *The Cognitive Neuroscience of Metacognition.* Elsevier Press. pp. 245 - 265.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329,* 1541–1543.

Fox, K. C.R., & Christoff, K. (2014). Metacognitive facilitation of spontaneous thought processes: When metacognition helps the wandering mind find its way. In S. M. Fleming & C. D. Frith (Eds). *The Cognitive Neuroscience of Metacognition.* pp. 293 – 319. Springer: New York, NY.

Funnell, M., Metcalfe, J., & Tsapkini, K. (1996). In the mind but not on the tongue: Feeling of knowing in anomic patient H.W. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 171–194). Hillsdale, NJ: Erlbaum.

Gray, J. R., Bargh, J. A., & Morsella, E. (2013). Neural correlates of the essence of conscious conflict: fMRI of sustaining incompatible intentions. *Experimental Brain Research, 229,* 453 – 465.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences, USA, 98*, 5359–5362.

Holroyd, C.B., & Yeung, N. (2012). Motivation of extended behaviors by anterior

cingulate cortex. *Trends in Cognitive Sciences, 16*, 122 – 128.

Humphrey, N. (2006). *Seeing Red: A Study in Consciousness*, Belknap Press/Harvard University Press, 2006.

Humphrey, N. (2011). *Soul Dust: the Magic of Consciousness*. Princeton University Press: NJ.

Izaute M. & Bacon E. (2006). Effects of the amnesic drug lorazepam on complete and partial information retrieval and monitoring accuracy. *Psychopharmacology*, *188*, 472-481.

Jenkins, A.C. & Mitchell, J.P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience, 6*, 211-218.

Jing, L., Niki, K., & Philips, S. (2004). Neural correlates of the 'Aha! reaction' *NeuroReport, 15*, 2013-2017.

Jing, L., Niki, K., Xiaoping, Y., & Yue-jia, L. (2004). Knowing that you know and knowing that you don't know: A fMRI study on feeling-of-knowing (FOK). *Acta Psychologica Sinica, 36*, 426–433.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2012). Neural correlates of self-reflection, *Brain,* 125, 1808 – 1814.

Kelley, W.T., Macrae, C.N., Wyland, C., Caglar, S., Inati, S. & Heatherton, T.F. (2002). Finding the self? An event-related fMRI Study. *Journal of Cognitive Neuroscience, 14*, 785-794.

Kikyo, H., & Miyashita, Y. (2004). Temporal lobe activation of "feeling-of-knowing" induced by face-name associations. *NeuroImage, 23*, 1348–1357.

Kikyo, H., Ohki, K., & Miyashita, Y. (2002). Neural correlates for feeling-of-knowing: An fMRI parametric analysis. *Neuron, 36*, 177–186.

Kikyo, H., Ohki, K., & Sekihara, K. (2001). Temporal characterization of memory retrieval processes: an fMRI study of the "tip of the tongue" phenomenon. *European Journal of Neuroscience, 14*, 887–892.

Kim, H., & Cabeza, R. (2007). Trusting our memories: dissociating the neural correlates of confidence in veridical versus illusory memories. *Journal of Neuroscience, 27*, 12190 – 12197.

Kornell, N. (2014). Where is the "meta" in animal metacognition. *Journal of Comparative Psychology*. 128, 143-149.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science, 18*, 64-71.

Maril, A., Simon, J. S., Mitchell, J. P., Schwartz, B. L., & Schacter, D. L. (2003). Feeling-of-knowing in episodic memory: An event-related fMRI study. *NeuroImage, 18*, 827–836.

Maril, A., Simons, J. S., Weaver, J. J., & Schacter, D. L. (2005). Graded recall success: An event-related fMRI comparison of tip of the tongue and feeling of knowing. *Neuroimage, 24,* 1130–1138.

Maril, A., Wagner, A. D., & Schacter, D. L. (2001). On the tip of the tongue: An event-related fMRI study of semantic retrieval failure and cognitive conflict. *Neuron, 31,* 653–660.

McCurdy, L.Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience, 33*, 1897-1906.

McGuire, J. T. & Botvinick, M. M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences, 107,* 7922-7926.

Metcalfe, J., Butterfield, B., Habeck, C., & Stern, Y. (2012). Neural correlates of people's hypercorrection of their false beliefs. *Journal of Cognitive Neuroscience, 24*, 1571-83.

Metcalfe, J., & Son, L. K. (2012). Anoetic, noetic, and autonoetic metacognition. In *The Foundations of Metacognition,* M. Beran, J. R. Brandl, J. Perner, & J. Proust (Eds.) Oxford University Press: Oxford, UK, pp. 289-301.

Metcalfe, J., Van Snellenberg, J. X., DeRosse, P., Balsam, P. & Malhotra, A. (2012). Action monitoring and metacognition of agency in participants with schizophrenia. *Philosophical Transactions of the Royal Society B, 376,* 1391-1400.

Miele, D. M., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience, 23*, 3620-3636.

Moritz, S., Gläscher, J., Sommer, T., Büchel, C., & Brause, D. F. (2006). Neural correlates of memory confidence. *NeuroImage, 33*, 1188–1193

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of learning and motivation: Advances in*

research and theory (Vol. 26, pp. 125–173). San Diego: Academic Press.

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 languagewith areas in monkey frontal cortex. Neuron, 81,, 700–713.

Northoff G., Heinzel A., de Greck M., Bermpohl F., Dobrowolny H., Panksepp J. (2006). Self-referential processing in our brain--a meta-analysis of imaging studies on the self. *Neuroimage, 31,* 440 – 457.

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Kihlstrom, J. F., D'Esposito, M & Gabrieli, J. D. E. (2005). The neural correlates of direct and reflected self-knowledge. *Neuroimage, 28,* 797-814.

Ochsner, K. N., Knierim, K., Ludlow, D., Hanelin, J., Ramachandran, T. & Mackey, S. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience, 16*, 1748-1772.

Pannu, J. K., & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychological Review, 15*, 105 – 130.

Phan, K. L., Wager, T. D., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16,* 331–348.

Posner M.I., & DiGirolamo G.J. (1998). Executive attention: Conflict, target detection, and cognitive control. In R. Parasuraman (Ed.), *The attentive brain.* Cambridge, Mass: MIT Press.

Price, D. D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science, 288,* 1769 – 1772.

Russell, B. (1967). *The Problems of Philosophy.* London: Oxford University Press.

Schnyer, D. M., Nicholls, L., & Verfaellie, M. (2005). The role of VMPC in metamemorial judgments of content retrievability. *The Journal of Cognitive Neuroscience, 17*, 832 – 846.

Schwartz, B. L. (2006). Tip-of-the-tongue states as metacognition. *Metacognition and Learning, 1*, 149 – 158.

Schwartz, B. L. (2008). Working memory load differentially affects tip-of-the-tongue states and feeling-of-knowing judgment. *Memory & Cognition, 36,* 9 – 19.

Schwartz, B. L., & Bacon, E. (2008). Metacognitive Neuroscience. In J. Dunlosky, & R. A. Bjork (Eds). *Handbook of Memory and Metamemory: Essays in Honor of Thomas O. Nelson.* Psychology Press: New York, New York. pp. 355 – 371.

Schwartz, B. L., & Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, Behavior, and Experience. *Memory & Cognition, 39*, 737 – 749.

Schwartz, B. L., & Metcalfe, J. (2013). Tip-of-the-tongue states and information seeking. Presented at the 52th annual meeting of the Psychonomics Society, Toronto, Ontario, Canada. November, 2013.

Schwartz, B. L., & Metcalfe, J. (2014). Tip-of-the-tongue (TOT) states: Mechanisms and metacognitive control. In B. L. Schwartz & A.S. Brown (Eds). *Tip-of-the-tongue states and related phenomena*. Cambridge University Press. in press.

Semendeferi K., Armstrong E., Schleicher A., Zilles K., & Van Hoesen G.W. (2001). Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology, 114*, 224 – 241.

Shettleworth, S. (2012). *Fundamentals of Comparative Cognition.* Oxford University Press: New York.

Shimamura, A. P., & Squire, L. R. (1986). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Leaning, Memory, and Cognition, 12*, 452-460.

Shushruth, S. (2013). Exploring the neural basis of consciousness through anesthesia. *The Journal of Neuroscience, 33*, 1757 – 1758.

Sturm, V.E., Sollberger, M., Seely, W. W., Rankin, K. P., Ascher, E. A. Rosen, H. J., Miller, B. L., & Levenson, R. W. (2013). Role of right pregenual anterior cingulate cortex in self-conscious emotional reactivity. *Social, Cognitive, and Affective Neuroscience, 8,* 468 - -474.

Stuss, D.T., Alexander M.P., Shallice, T., Picton, T.W., Binns, M.A., Macdonald R., Borowiec, A., & Katz, D.I. (2005). Multiple frontal systems controlling response speed. *Neuropsychologia, 43*, 396–417.

Thiebaut de Schotten, M., Dell'Acqua, F., Valabregue, R., Catani, M. (2012). Monkey to human comparative anatomy of the frontal lobe association tracts. *Cortex, 48*, 82-96.

Tinbergen, N. (1975) *The Animal and its World Vol. 1*, Massachusetts: Harvard University Press. pp. 76-78, 103-105.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace & J. Metcalfe (Eds.), The missing link in cognition: Origins of self-reflective consciousness (pp. 3–56). New York, NY: Oxford University Press.

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*, 199 – 206.

Table 1. Neuroimaging studies, metacognitive judgments, and brain region uniquely correlated with metacognitivejudgment (based on but updated from Schwartz & Bacon, 2008).

Study	Metacognitive measure	Region of cortex
Kao et al (2005)	JOL	ventromedial PF, lateral and dorsomedial PF
Do Lam et al (2012)	JOL	ACC, medial PF, orbitofrontal cortex
Kikyo et al (2001)	TOT	ACC, dorsolateral PFC, inferior PFC
Maril et al (2001)	TOT	ACC, dorsolateral PFC, inferior PFC
Maril et al (2005)	TOT	ACC, dorsolateral PFC, inferior PFC
Maril et al (2003)	FOK	Inferior PFC
Kikyo et al (2002)	FOK	Inferior PFC, Medial PFC, insula
Kikyo & Miyashita (2004)	FOK	Medial PFC, dorsolateral PFC, ventromedial PFC
Jing et al (2004)	FOK	Inferior PFC
Schnyer et al (2005)	FOK	Ventromedial PFC
Chua et al (2009)	FOK	ventromedial PF, dorsolateral PF, parietal lobe
Elman et al (2012)	FOK	medial PFC, ventromedial PFC, dorsolateral PFC,

Moritz et al (2006)	RCJ	ACC, parietal lobe
Chua et al (2006)	RCJ	parietal lobe
Kim and Cabeza (2007)	RCJ	lateral PFC, parietal lobe
Chua et al (2009)	RCJ	ventromedial PFC, dorsolateral PFC, parietal lobe
Yokoyama et al (2010)	RCJ	fronto-polar cortex
Fleming et al (2010)	RCJ	BA10 (medial PFC), ACC, right parietal cortex
McCurdy et al (2013)	RCJ	medial PFC, precuneus
Jing et al (2004)	Aha	ACC, insula, lateral PFC
Miele et al (2011)	JOA	BA 10 (medial PFC)

JOL = judgment of learning; TOT = tip of the tongue state; FOK = feeling of knowing; RCJ = retrospective confidence judgment; Aha = "aha" reaction. JOA = Judgment of Action.

Figures

Figure 1. Contrast map of high confidence errors and low confidence errors. Red areas were more active for high confidence errors, whereas green areas were more active for low confidence errors; both thresholded at an uncorrected threshold of $p < .001$. Yellow and blue areas, respectively, denote areas of higher and lower activation for high confidence errors at a family-wise-corrected p level of .05. These more stringently thresholded areas are small and barely visible.

Reprinted from Metcalfe et al, 2012, Figure 1. P. 1577.

Figure 2. Figure 2. Statistical activation maps, and bar graphs depicting the parameter estimates per condition. Activation maps are overlaid onto the mean anatomical image across participants. Regions of interest (ROIs) defined from (a) JOLs following successful memory formation (JOL_SM) were located in the left MTL; (b) JOLs predicting memory formation were located in the ACC; (c) JOLs predicting memory formation (JOL_PM) masked with successful recall (REC_SM) was located in the mPFC. Coordinates are presented in Table 1.

doi:10.1371/journal.pone.0030009.g002

Reprinted from Do Lam et al (2012). Figure 2. P. 5.

Figure 3. Five regions demonstrated TOT-selective activation: (a) anterior cingulate (6, 18, 36), (b) right DLPFC (42, 39, 33), (c) right inferior PFC (42, 18, 3), (d) bilateral anterior frontal cortex (left: 33, 54, 9; right: 30, 54, 21). Displayed are sections through each region, and averaged event-related responses associated with each retrieval outcome. Coordinates are in MNI space.

Reprinted from Maril et al (2005). Figure 2. P. 1135.

Acknowledgements.

We thank John Dunlosky, Uma Tauber, Matthew Sutherland, and Steve Fleming for comments on an earlier draft of this paper. We thank Jack Frazier for his help with illustrations.