Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging

Marion Rouault¹, Andrew McWilliams¹,²,³, Micah G. Allen¹ and Stephen M. Fleming¹,⁴

¹Wellcome Centre for Human Neuroimaging, University College London, London, UK, ²Department of Child and Adolescent Psychiatry, Institute of Psychiatry, Psychology and Neuroscience, London, UK, ³Great Ormond Street Hospital for Children NHS Trust, London, UK and ⁴Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

Abstract

Metacognition is the capacity to evaluate and control one’s own cognitive processes. Metacognition operates over a range of cognitive domains, such as perception and memory, but the neurocognitive architecture supporting this ability remains controversial. Is metacognition enabled by a common, domain-general resource that is recruited to evaluate performance on a variety of tasks? Or is metacognition reliant on domain-specific modules? This article reviews recent literature on the domain-generality of human metacognition, drawing on evidence from individual differences and neuroimaging. A meta-analysis of behavioral studies found that perceptual metacognitive ability was correlated across different sensory modalities, but found no correlation between metacognition of perception and memory. However, evidence for domain-generality from behavioral data may suffer from a lack of power to identify correlations across model parameters indexing metacognitive efficiency. Neuroimaging data provide a complementary perspective on the domain-generality of metacognition, revealing co-existence of neural signatures that are common and distinct across tasks. We suggest that such an architecture may be appropriate for “tagging” generic feelings of confidence with domain-specific information, in turn forming the basis for priors about self-ability and modulation of higher-order behavioral control.

Whether a mental process is domain-general (shares resources across many situations or tasks) or domain-specific is a broad question that is pertinent to many areas of psychology. For instance, it has long been debated whether intelligence relies on a single underlying resource (a g factor) or on independent components (Chiappe & MacDonald, 2005; Kanazawa, 2004; Kievit et al., 2017). In cognitive neuroscience, Duncan et al. have proposed that a “multiple demand” system supports executive functions across many different tasks (Duncan, 2010). In this article, we focus on recent research on the domain-generality of neurocognitive substrates supporting metacognition.

Metacognition is defined as cognition about cognition—the ability to reflect on, monitor and control another cognitive process (Dunlosky & Metcalfe, 2008; Nelson & Narens, 1990). In the laboratory, as we will see in more detail below, metacognition can be assessed by recording individuals’ judgments of their performance on a particular task, such as their confidence in a decision or a judgment of whether learning will be successful (a “second-order” judgment). Because metacognition is by definition second-order to other cognitive processes, it may operate across multiple “domains” of cognition—for instance, one might engage in metacognition about percepts, about memories, about decisions, and so forth. Progress has been made on understanding the neural basis of metacognition (see Fleming & Dolan, 2012 for a review), which will be considered at more length below. However, it remains poorly understood as to whether metacognition relies on a domain-general resource that is “applied” to the task at hand, or whether different metacognitive processes are engaged when evaluating performance in different domains (Figure 1a). This article reviews and critically appraises progress on this issue.

Measures of metacognition

In order to assess the relationship between metacognition across domains, we require metrics of metacognitive ability that are robust and comparable across tasks. Here we focus on objective measurement of metacognition from behavioral tasks rather than self-report...
find that metacognitive sensitivity is highly correlated across two
domains, it is important to ensure that estimation of these components of metacognition is not con-
founded by first-order task performance. For instance, we might describe them as having a high degree of metacognitive
ability. In general, metacognition is said to be accurate when correct decisions are held with high confidence and incorrect
decisions are held with lower confidence—in other words, metacognitive accuracy refers to the correlation between task
performance and confidence. The various approaches for characterizing this correlation have been comprehensively reviewed
elsewhere (Fleming & Lau, 2014). It is useful to distinguish two aspects of metacognitive judgments—their sensitivity and bias. These are illustrated in the
cartoon in Figure 1b. Each panel shows example probability densities of confidence ratings conditional on correct and incorrect
task performance. If these distributions are cleanly separated, this implies the subject is able to recognize accurate from inaccurate performance using the confidence scale and we would describe them as having a high degree of metacognitive sensitivity. In contrast, metacognitive bias refers to an overall level of confidence averaging over performance. These aspects of metacognition are theoretically independent. For instance, someone who has low overall confidence (low bias) may still be sensitive on a trial-by-trial basis to fluctuations in performance (high sensitivity). By applying a modification of signal detection theory (SDT), known as “type 2” SDT, it is possible to quantify sensitivity and bias of ratings with respect to objective performance (Clarke, Birdsaill, & Tanner, 1959; Fleming & Lau, 2014; Galvin, Podd, Drga, & Whitmore, 2003).

However, when making inferences about processes that are shared or distinct across domains, it is important to ensure that estimation of these components of metacognition is not con-
founded by first-order task performance. For instance, we might find that metacognitive sensitivity is highly correlated across two
unrelated tasks, but this correlation would be less interesting if it were simply a consequence of first-order performance also being correlated between tasks. Indeed, several measures of metacognitive sensitivity (such as area under the type 2 Receiver Operating Characteristic curve [AUROC2] and confidence-accuracy correlations) are themselves affected by first-order performance (Galvin et al., 2003; Masson & Rotello, 2009)—the same individual will likely show greater metacognitive sensitivity on an easy task compared with a hard task. If performance is not matched or accounted for between conditions, erroneous conclusions may be drawn, for instance that a patient group has a deficit in metacognition when such a deficit is instead explained by a difference in first-order performance (Figure 2).

One elegant solution to the problem of controlling for performance confounds is the meta-d’ framework developed by Maniscalco and Lau (2012). This approach posits a generative model of confidence data within a SDT framework. Fitting the model to data returns a parameter, meta-d’, that reflects the level of first-order performance (known as d’) that would have led to the observed confidence rating data under an ideal observer model. Meta-d’ can then be compared with actual d’ (for instance by computing the ratio meta-d’/d) to give a measure of metacognitive efficiency, which quantifies the level of metacognitive sensitivity relative to first-order performance. By using metacognitive efficiency as our measure of metacognition we can meaningfully compare scores across individuals or task domains. Alternatively, if simpler measures of metacognitive sensitivity are

---

**Figure 1.** (a) It remains debated whether metacognition operates as a domain-general resource applied over cognitive domains (left) and/or whether metacognition itself relies on domain-specific components that operate over corresponding cognitive domains. (b) Metacognitive bias and metacognitive sensitivity are two independent aspects of metacognition. Metacognitive bias corresponds to an overall tendency to rate confidence higher (right panels) or lower (left panels), irrespective of performance. Metacognitive sensitivity quantifies the extent to which correct and error trials can be discriminated (adapted from Fleming & Lau, 2014).

---

**Figure 2.** Top panel: differences in task performance might produce spurious differences in metacognitive sensitivity between groups or task domains. Bottom panel: if task performance is matched between domains, differences in metacognitive sensitivity are likely to reflect true differences in metacognition.
employed, it is important to ensure that any potential confounds due to differences in first-order performance between conditions are examined and accounted for (Figure 2).

**Domain-generality in metacognitive ability (1): individual differences**

A classical approach to studying domain-generality of mental processes is examining patterns of individual differences to ask whether variance is shared or distinct across tasks. Assuming that our metrics are reliable and free of confounds, cross-correlations between domains indicate shared constraints on a particular ability. For instance, if we find that across individuals, faster choice response times are strongly predictive of IQ scores, we might conclude that greater processing speed contributes to both decision time and intelligence (Ratcliff, Thapar, & McKoon, 2010; Ritchie, Bates, Der, Starr, & Deary, 2013). Correlations of metacognitive measures with other stable individual differences (such as personality or mental health) may also reveal domain-general aspects of metacognition. In this section we review studies that have taken this approach to investigate domain-general and domain-specific aspects of metacognitive efficiency and bias, and provide a formal meta-analysis to quantify behavioral evidence for domain-generality.

In the perceptual domain, metacognitive sensitivity (measured as AUROC2) has been found to be correlated across individuals for contrast and orientation discrimination tasks (Song et al., 2011), despite perceptual thresholds (first-order performance) in each case being uncorrelated. Similar results were found when examining metacognitive efficiency (meta-\(d/d')\) correlations across visual, auditory, and tactile modalities (Fairvre, Filevich, Solovey, Kühn, & Blanke, 2018). Tactile metacognitive sensitivity (measured using AUROC2) was found to be uncorrelated with metacognitive sensitivity on cardiac and respiratory discrimination tasks, despite the latter correlating with each other (Garfinkel et al., 2016). Ais, Zylibzer, Bartfield, and Sigman (2016) found strong correlations between metacognitive bias (average confidence levels) across several perceptual tasks (auditory, luminance and contrast discrimination tasks, and a "partial report" task which required identification of a letter in a briefly flashed array). However, they found correlations in metacognitive sensitivity (AUROC2) only between auditory and luminance tasks. In addition, this study identified similar confidence "profiles" for a given individual, indicating idiosyncratic and stable patterns of confidence ratings across tasks (Ais et al., 2016).

Although these studies explored inter-relationships between metacognition in different perceptual modalities, it could be argued that all such tasks belong to a broader perceptual domain, but are further in task space from other cognitive domains such as memory. Within the memory domain, metacognitive bias, but not metacognitive sensitivity (assessed by the degree of match between confidence and recall performance), was found to be correlated across face and word recall tasks (Bornstein & Zickafoose, 1999; Sadeghi, Ekhhtiari, Bahrami, & Ahmadabadi, 2017; Thompson & Mason, 1996; West & Stanovich, 1997) and across a variety of JOL tasks (Kelemen, Frost, & Weaver, 2000)—albeit in these studies first-order performance was not matched across tasks. More recent studies have compared metacognition for perception and memory while also matching performance: In this case, metacognitive efficiency was correlated with perceptual and memory domains (McCurdy, Maniscalco, Metcalfe, Liu, de Lange & Lau, 2013; Palmer, David, & Fleming, 2014). Samaha and Postle also found evidence of domain-generality in metacognitive sensitivity for perceptual discrimination and visual working memory (measured using performance-confidence correlations and AUROC2), but whether this result reflects generalization beyond perception is unclear because perceptual resources may also be required for short-term memory of visual orientation (Samaha & Postle, 2017). In contrast, other studies found no correlation between metacognitive efficiency across memory and perceptual tasks (Baird, Cieslak, Smallwood, Grafton, & Schooler, 2015; Baird, Smallwood, Gorgolewski, & Margulies, 2013; Morales, Lau, & Fleming, 2018). These mixed findings may be due to differences in metacognition metrics (AUROC2 in Baird et al. vs. meta-\(d/d'\) in McCurdy et al. and Morales, Lau, & Fleming, 2018), and/or differences in task requirements (two-alternative-forced choice [2AFC] vs. Yes/No) (Ruby, Giles, & Lau, 2017), as we discuss further below. Finally, no correlation was found between metacognitive sensitivity (measured using AUROC2) on a visual discrimination task and a task involving mentalizing and reasoning (Valk, Bernhardt, Böckler, Kanske, & Singer, 2016) or between metacognitive sensitivity on perception, memory and error awareness tasks (Fitzgerald, Arvaneh, & Dockree, 2017).

The above studies can be clustered into two main groups: those that examine inter-correlations between metacognitive sensitivity in different perceptual discrimination tasks and those studying correlations between metacognition of recognition memory and perception. To assess evidence for domain-generality we conducted a meta-analysis of cross-domain correlation coefficients for these two study categories (Figure 3). From a literature search we identified studies that fell into one of these categories and employed signal-detection theoretic measures of metacognition (M-ratio or AUROC), revealing 12 manuscripts and 19 total independent experiments. Effect size (\(r\)), sample size (\(n\)), and types of cross-domain correlation (memory-perception vs. perception-perception) were hand coded. In cases where multiple modalities were probed within a sample (e.g., audition, touch, and vision), the "multi-modal" \(r\)-value was calculated as the average \(R\) across modalities. Meta-analytic effect sizes (cross-domain \(r\)) were calculated using a Fisher R-to-Z random effects model, implemented in the metafor R-package, version 3.3.2 (Viechtbauer, 2010). Specifically, we performed three meta-analyses—one of overall cross-domain correlations (\(n = 19\), one on cross-domain correlations within the perceptual modality (\(n = 9\)), and another of memory-perception cross-domain correlations (\(n = 10\)).

We found that across all studies, cross-domain correlations were significantly greater than zero (meta-analytic \(r = 0.27, 95\%\) confidence interval \([CI] = [0.13, 0.41]\), \(p < .001\)) and exhibited significant heterogeneity across effect sizes (\(Q = 50.46, df = 18, p < .001, I^2 = 69.7\%\)). This result was primarily driven by medium to strong cross-perceptual correlations (meta-analytic \(r = 0.55, 95\%\) CI = \([0.34, 0.76]\), \(p < .001\)). Restricting our analysis to perceptual effect sizes, we did not observe significant heterogeneity (\(Q = 14.49, df = 8, p = .07, I^2 = 45.4\%\)). In contrast, cross-domain correlations between memory and perception-based tasks were not significant (\(r = 0.09, 95\%\) CI = \([-0.02, 0.21]\), \(p = .10\)), and did not show effect heterogeneity (\(Q = 14.76, df = 9, p = .09, I^2 = 38.3\%\)). These results suggest that cross-task correlations in metacognitive ability are primarily obtained when examining tasks tapping into the same functional modality, that is perception (Figure 3).
Specifically, previous studies have documented that meta-confidence ratings in the two cases (Ruby, Giles, & Lau, 2017). The right column indicates the Fisher’s $z$ transformed correlation coefficient. CI = confidence interval.

**Meta–Perception Only**

Samaha & Postle, 2017. Exp. 1. 0.83 [0.35, 1.30]
Samaha & Postle, 2017. Exp. 2. 0.81 [0.34, 1.29]
Samaha & Postle, 2017. Exp. 3. 0.52 [0.05, 1.00]
Garfinkel et al., 2016. 0.18 [−0.13, 0.50]
Ais et al., 2016. 0.13 [−0.31, 0.57]
Faivre et al., 2016. Exp. 1. 0.44 [−0.16, 1.03]
Faivre et al., 2016. Exp. 2. 0.48 [−0.11, 1.08]
Faivre et al., 2016. Exp. 3. 0.85 [0.37, 1.32]
Song et al., 2011. 0.89 [0.38, 1.39]

RE Model for Perception (Q = 14.49, df = 8, p < 0.001; $I^2$ = 45.4%)

**Meta–Memory vs Meta–Perception**

Morales et al., 2018. –0.08 [−0.51, 0.35]
Fitzgerald et al., 2017. 0.10 [−0.28, 0.48]
Ruby et al., 2017. Exp. 1. 0.32 [0.12, 0.52]
Ruby et al., 2017. Exp. 2. 0.07 [−0.14, 0.28]
Ruby et al., 2017. Exp. 3. 0.09 [−0.12, 0.30]
Baird et al., 2015. −0.07 [−0.38, 0.24]
Baird et al., 2014. Pre–training. 0.01 [−0.29, 0.31]
Baird et al., 2014. Post–training. 0.04 [−0.26, 0.34]
McCurdy et al., 2013. 0.51 [0.16, 0.86]
Baird et al., 2013. −0.13 [−0.41, 0.15]

RE Model for Memory (Q = 14.76, df = 9, p = 0.100; $I^2$ = 38.3%)

**Meta–Perception Only**

Samaha & Postle, 2017. Exp. 1. 0.55 [0.34, 0.76]
Samaha & Postle, 2017. Exp. 2. 0.36 [0.12, 0.60]
Samaha & Postle, 2017. Exp. 3. 0.32 [0.12, 0.52]
Garfinkel et al., 2016. 0.10 [−0.26, 0.46]
Ais et al., 2016. 0.01 [−0.29, 0.31]
Faivre et al., 2016. Exp. 1. 0.09 [−0.12, 0.30]
Faivre et al., 2016. Exp. 2. 0.07 [−0.14, 0.28]
Faivre et al., 2016. Exp. 3. 0.09 [−0.12, 0.30]
Song et al., 2011. 0.89 [0.38, 1.39]

RE Model for All Studies (Q = 50.46, df = 18, p < 0.001; $I^2$ = 69.7%)
counts for known \(d'/d\)' values using the metad_sim function from the HMeta-d toolbox (https://github.com/metacoglab/HMeta-d), keeping confidence rating criteria fixed across domains and subjects. The number of trials per subject differed between simulations (50 vs. 400). The model outputs a posterior belief distribution over the across-domain correlation coefficient. It can be seen that as the number of trials per subject increases (i.e., the certainty associated with individual \(d'/d\) estimates goes up, lower panel in Figure 4), we can be more certain about the presence of a domain-general correlation (narrower posterior density). We recommend applying such multi-level models when analyzing individual difference correlations to ensure this parameter uncertainty is appropriately taken into account.

Other studies have investigated relationships between metacognition and other aspects of personality and executive function which, if found, would lend support to a domain-general metacognitive contribution to metacognition. For instance, factors such as general intelligence or motivation to engage with a task could affect metacognition across multiple domains. However, if such general factors are significantly altered, it is unlikely that metacognitive processes would be selectively affected while also leaving first-order performance spared; for instance, an altered ability to follow task instructions following a prefrontal cortex (PFC) lesion is likely to affect both task performance and metacognitive evaluation. The integrity of these “global” factors can thus be seen as a necessary but not sufficient condition for enabling metacognition. Notably, however, despite both relying on aspects of higher cognition, we have found that over several datasets perceptual metacognitive efficiency is not related to measures of fluid intelligence (Fleming, Huijgen, & Dolan, 2012; Palmer, David, & Fleming, 2014), even when such correlations were examined in a large-scale dataset of ~1,000 individuals (Rouault, Seow, Gillan, & Fleming, 2018). Such independence may be due to fluid intelligence relying on posterolateral frontal and parietal “multiple demand” regions (Woolgar et al., 2010), whereas metacognition has been linked to anterior prefrontal regions, as considered in more detail below.

In summary, analyses of individual differences in metacognitive efficiency indicate the presence of domain-general contributions to confidence judgments across distinct perceptual discrimination tasks. Such variation in metacognition is isolated from variation in first-order performance. However, it remains unclear whether a shared resource supports metacognitive efficiency across more distant domains, such as recognition memory and perceptual discrimination. One important consideration in conducting such cross-domain correlation analyses is to ensure that uncertainty in estimation of metacognition within a particular domain is appropriately propagated to the analysis of between-domain correlations, which is now possible within hierarchical Bayesian frameworks. Taken together these findings also raise the issue of how to define a separation between domains, and whether the notion of domain should instead be considered as existing along a continuum or gradient (see “Computational processes” section below). Furthermore we should remain mindful that the architecture of metacognition (and therefore any shared variance between different tasks) may
well be organized along different lines than the cognitive processes being monitored, and which are typically compared in the laboratory (e.g., perception, memory).

**Domain-generality in metacognitive ability (2): neuropsychology**

The study of individual differences identifies shared variance in behavioral performance across a large number of healthy individuals. In contrast, neuropsychology seeks to identify dissociations between abilities induced by patterns of brain damage. Classic studies by Shimamura and colleagues revealed that metamemory abilities (such as feeling of knowing [FOK] or JOLs) are selectively impaired following frontal lesions (Janowsky, Shimamura, & Squire, 1989; Shimamura & Squire, 1986). Metamemory evaluation has itself been divided into distinct judgment types (see Chua, Pergolizzi, & Weintrob, 2014, for a review). A key distinction is that judgments can be either prospective, occurring prior to memory retrieval, or retrospective. Prospective judgments include FOK, the likelihood of recognizing an item that currently cannot be recalled, and JOL, a belief during learning about the success of subsequent recall. More recent studies indicated such lesion deficits may not apply to all forms of metamemory judgment. For instance, two independent studies found that damage to the medial PFC was associated with decreased prospective FOK accuracy but intact retrospective confidence judgments (Schnyer et al., 2004) and JOLs (Modirrousta & Fellows, 2008). The reverse dissociation was reported by Pannu and Kaszniak (2005), who found that deficits in retrospective confidence judgments were associated with lateral frontal lesions.

Fewer studies have taken a neuropsychological approach to ask whether metacognitive deficits are shared across multiple task domains. Fleming et al. (2014) studied three groups of subjects matched for age and IQ—a healthy control group, a group with aPFC lesions, and a group with temporal lobe lesions. Each participant completed both a recognition memory task with word stimuli and a perceptual discrimination task about the relative density of two dot patches. In both tasks retrospective confidence ratings were elicited on a trial-by-trial basis, allowing assessment of meta-\(d^\prime\) for each subject/task domain. A selective deficit in metacognitive efficiency (meta-\(d^\prime/d^\prime\)) for perceptual discrimination was observed in the aPFC group (Figure 5a) despite equivalent first-order performance and metacognitive bias. Such a result is consistent with a contribution of aPFC to metacognition of perceptual decision-making (Allen et al., 2017; Fleming, Weil, Nagy, Dolan, & Rees, 2010; Yokoyama et al., 2010), but suggests other brain regions may be sufficient to support intact metacognition of recognition memory (Baird et al., 2013; McCurdy et al., 2013). Notably, all patients were tested in a post-acute phase of their lesion, so it is possible that an early domain-general deficit may have been observed sooner after surgery. More generally, neuroplasticity and reorganization following lesions make it
Human Metacognition Across Domains

difficult to draw strong conclusions about the typical functional anatomy of metacognition from lesion studies alone (Lemaitre, Herbet, Duffau, & Lafargue, 2017).

**Domain-generality in metacognitive ability (3): neuroimaging**

Behavioral and neuropsychological data can inform on whether a mental process relies on a shared resource, but provide less insight into the mechanisms that underpin this resource. As noted above, a common resource may be involved but not be detected due to domain-specific unreliability in the measurement of metacognition. Conversely, a domain-general pattern may be driven by a third factor that affects domain-specific processes in equal measure, such as stress (Reyes, Silva, Jaramillo, Rehbein, & Sackur, 2015) or fatigue (Maniscalco, McCurdy, Odegaard, & Lau, 2017).

Several recent studies have focused on the neural basis of human metacognition, which have been reviewed at length elsewhere (Fleming & Dolan, 2012). Briefly, in concordance with the neuropsychological literature highlighted above, anatomical (Allen et al., 2017; Fleming et al., 2010; McCurdy et al., 2013) and functional (Baird et al., 2013; Cortese, Amano, Koizumi, Kawato, & Lau, 2016; De Martino, Fleming, Garrett, & Dolan, 2013; Fleck, Daselaar, Dobkins, & Cabeza, 2006; Fleming et al., 2012; Higenstock, Weiss, & Witte, 2014; Yokoyama et al., 2010) neuroimaging data indicate that a frontoparietal network contributes to metacognitive estimates of task performance across a range of tasks. Within this network, electrophysiological studies in humans have focused on the role of the posterior medial frontal cortex (pMFC) and associated error-related negativity in performance monitoring (Gehring, Goss, Coles, Meyer, & Donchin, 1993; Dehaene, Posner, & Tucker, 1994). More recently, aPFC has also been implicated in supporting explicit metacognitive judgments, leading Fleming and Dolan (2012) to propose that connectivity between interoceptive cortices (cingulate and insula) and aPFC may underpin the fidelity of explicit metacognition. Human neurophysiological work has also identified a key role for frontoparietal areas in confidence formation. In particular, recordings in monkey lateral intraparietal cortex indicate that variability in lateral intraparietal cortex firing rates is predictive of both decisions and decision confidence (Hanks, Mazurek, Kiani, Hopp, & Shadlen, 2011; Kiani and Shadlen, 2009). Furthermore, activity in rat orbitofrontal cortex carries signals related to decision confidence in a perceptual discrimination task (Kepecs, Corthell, & Shadlen, 2014), and expected value. Some studies of confidence explicitly modeled reaction times in their fMRI analysis (Fleck et al., 2006; Gherman & Philiastides, 2017; Lebretou, Abitbol, Daunizeau, & Pessiglione, 2015), whereas others did not (Heereman, Walter, & Heekeren, 2015; Morales, Lau, & Fleming, 2018), which are also observed in the memory domain (episodic retrieval) (Fleck et al., 2006). In pMFC and ventromedial prefrontal cortex (vmPFC), multivariate fMRI analyses revealed that it was possible to predict confidence in a memory task from patterns decoded in a perceptual task matched for stimulus and task requirements, and vice-versa (Figure 5C), suggesting that confidence covaries with task-independent neural representations (Morales, Lau, & Fleming, 2018). In contrast, right lateral aPFC instead showed significant decoding effects within- but not across-domain. This domain-specific neural representation of confidence suggests that lateral aPFC may “tag” metacognitive representations with task-specific information, which could be particularly relevant for future meta-level control decisions such as which task to engage in next.

In perceptual decision-making tasks, classical univariate analyses of functional magnetic resonance imaging (fMRI) BOLD signal reveal a negative parametric relationship between confidence reports and activity in pMFC (encompassing dorsal anterior cingulate cortex and pre-supplementary motor area) (Fleck et al., 2006; Heereman, Walter, & Heekeren, 2015; Morales, Lau, & Fleming, 2018), which are also observed in the memory domain (episodic retrieval) (Fleck et al., 2006). In pMFC and ventromedial prefrontal cortex (vmPFC), multivariate fMRI analyses revealed that it was possible to predict confidence in a memory task from patterns decoded in a perceptual task matched for stimulus and task requirements, and vice-versa (Figure 5C), suggesting that confidence covaries with task-independent neural representations (Morales, Lau, & Fleming, 2018). In contrast, right lateral aPFC instead showed significant decoding effects within- but not across-domain. This domain-specific neural representation of confidence suggests that lateral aPFC may “tag” metacognitive representations with task-specific information, which could be particularly relevant for future meta-level control decisions such as which task to engage in next.

We note that some observations of domain-generality in neural data may be due not only to confidence but other correlated variables, for instance, decision time, which has been proposed as a relevant input for a confidence computation (Kiani, Corthell, & Shadlen, 2014), and expected value. Some studies of confidence explicitly modeled reaction times in their fMRI analysis (Fleck et al., 2006; Gherman & Philiastides, 2017; Lebretou, Abitbol, Daunizeau, & Pessiglione, 2015), whereas others did not (Heereman, Walter, & Heekeren, 2015; Morales, Lau, & Fleming, 2018), and it is not straightforward to determine whether decision time should be treated as a confound or a relevant variable of interest for studies of the neural basis of metacognition. Notably, regions often implicated in encoding expected value such as ventral striatum and vmPFC (Clithero & Rangel, 2013) are also often found to scale with confidence in perception, value and memory domains (De Martino et al., 2013; Gherman &
Philistides, Lebreton et al., 2015; Morales, Lau, & Fleming, 2018), suggesting that being confident is valuable, and/or that when highly confident, subjects expect imminent reward. The fact that a majority of studies have not dissociated confidence from implicit expected value potentially explains these pervasive, domain-general activations. Future studies are required to directly investigate a putative commonality of confidence and value representations, and to separate the component inputs to confidence formation (Bang & Fleming, 2018).

Previous studies have suggested that involvement of the precuneus (medial parietal cortex) is specific to metacognitive judgments. Indeed individual metacognitive efficiency in a memory task, but not in a perceptual task, was found to correlate with gray matter volume in the precuneus (McCurdy et al., 2013), and resting-state functional connectivity revealed that better metamemory was associated with increased connectivity between medial aPFC and precuneus (Baird et al., 2013). In addition, transcranial magnetic stimulation (TMS) application over precuneus impaired metacognitive efficiency for memory but not perception, both measured as meta-$d'$ - $d'$ (Ye, Zou, Lau, Hu, & Kwok, 2018), and univariate fMRI activation in precuneus was selectively increased during metacognitive judgments of memory, but not perception (Morales, Lau, & Fleming, 2018). However, a relationship between precuneus grey matter volume and metacognitive efficiency has also been detected in a perceptual decision-making task, albeit at an uncorrected whole-brain threshold (Fleming et al., 2010). Moreover, the same region was found to correlate negatively with confidence level in a visual motion-discrimination task (Heereman, Walter, & Heekeren, 2015). Lastly, using multivariate decoding of fMRI activity, Morales et al. found that classification of high versus low confidence trials in precuneus generalized across memory and perception domains (Morales, Lau, & Fleming, 2018). Together these results suggest that precuneus involvement may not be specific to metamemory. Interestingly, recent results suggest that the vividness of episodic memory was tracked by precuneus activity over and above memory precision and retrieval success (Richter, Cooper, Bays, & Simons, 2016). To the extent to which similar appraisals of vividness feed into the formation of perceptual confidence, this may explain the domain-general nature of findings in this region. An interesting alternative possibility is that precuneus is engaged when subjects leverage prior beliefs about self-ability to compute confidence in perception, hence needing to retrieve global beliefs about past experience from memory (see also Figure 6).

In summary, neuroimaging studies indicate a more nuanced picture than studies of behavior or individual differences, which have tended to argue for either domain-specific or domain-general aspects of metacognition. It is possible to reconcile these perspectives by demonstrating that both domain-specific and domain-general signals co-exist in the human brain, and that there may exist a gradient in which some tasks (such as different types of perceptual judgment) are more likely to rely on shared circuitry for metacognitive evaluation than others. In the next section we attempt to formalize these ideas through the lens of computational modeling.

**Computational processes supporting metacognition across domains**

**Models of confidence formation**

The simplest first-order models of confidence formation (such as SDT) assume that the internal states supporting decisions and confidence estimates are identical. Such frameworks predict that any covariation between metacognitive efficiency across domains should be accompanied by covariation in lower-level performance, and struggle to accommodate the evidence reviewed above that confidence can be selectively altered or impaired independently of task performance (see also Cortese et al., 2016; Lak et al., 2014; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). An extension to first-order models of confidence introduces post-decisional processing, thus explaining additional variability in confidence estimates without altering the fidelity of first-order performance (Navajas et al., 2017; Pleskac & Busemeyer, 2010; van den Berg et al., 2016). A somewhat more elaborate but flexible model is a “second-order” computation of confidence (Fleming & Daw, 2017). In the second-order framework, the efficacy of actions is monitored based on higher-order knowledge of the reliability of the decision-making system.

A second-order account provides a natural perspective on findings of domain-generality in metacognition. Similar circuits for second-order inference may be engaged across different domains to the extent to which their lower-level states and actions are similar (the “inputs” and “outputs”; see Figure 6). To take a concrete example, suppose that we are comparing metacognitive efficiency for a visual and auditory discrimination task, both requiring a right-handed button press to indicate the first-order judgment. While in each task, “state” estimation may depend on distinct (visual and auditory) neural circuitry; actions are supported by a common output (the left motor cortex). This commonality in response mapping may be sufficient to induce commonalities in second-order inference, leading to the observations of domain-general confidence signals in neuroimaging data. Such a pattern was observed by Faiivre et al., who suggest, “the supramodality of metacognition relies on supramodal confidence estimates and decisional signals that are shared across sensory modalities” (2018).

However, current computational models of confidence are relatively narrow in scope, focusing on experimentally controllable states and actions. To the extent that other internal variables covary with expected success, these may also become relevant “inputs” for subjective confidence. For instance, response times can provide a proxy for decision time, and be subsequently...
available to the agent for computing confidence (Benjamin, Bjork, & Schwartz, 1998; Kiani et al., 2014). However, for response times to be informative, one needs to have an estimation of the expected level of performance on the task, for instance, in the form of a running average over decision accuracy in a given experimental condition. If expected performance is variable within a task, or if it varies significantly across tasks, response times might be less useful as proxies for inferring decision accuracy and hence confidence. For example, in situations where noise in the stimulus induces a deviation from the expected performance level, the expected mapping between confidence and decision time may break down (Fetsch, Kiani, Newsome & Shadlen, 2014; Rahnev, Maniscalco, Luber, Lau & Lisanby, 2011; Peters et al., 2017; Zylberberg, Fetsch & Shadlen, 2016). Similarly, interoceptive states may provide additional proxies when evaluating self-performance (Allen et al., 2016; Chua & Bliss-Moreau, 2016). This perspective on the formation of confidence in decision-making converges with established "inferential" or cue-based models of confidence formation in the metamemory literature, which suggest that cues such as accessibility (the degree of partial knowledge about the target) contribute to confidence estimates (Koriat and Levy-Sadot, 2001). Similarly, familiarity with the stimuli (De Martino et al., 2013), and volatility/variability in stimulus evidence (Meyniel, Schlunegger & Dehaene, 2015; Zylberberg, Fetsch & Shadlen, 2016) also inform confidence. To the extent that some cues, such as response time or fluency, are useful across many different tasks, they may provide domain-general inputs to confidence and metacognition (Alter and Oppenheimer, 2009; Boldt, de Gardelle, & Yeung, 2017).

How does confidence guide behavior?

If confidence constitutes a proxy for the probability of success in a task (Peirce and Jastrow, 1884; Pouget, Drugowitsch, & Kepcecs, 2016), it may act as a “common currency” signal for estimating and comparing the relative likelihood of success between different tasks (de Gardelle & Mamassian, 2014). Such a common currency would be particularly useful when deciding on which tasks or goals to pursue in the future, especially when external feedback is unavailable. For instance, when choosing a career, it would be advantageous to internally evaluate and compare our performance or skill in different potential jobs. The existence of a common currency for confidence is supported by studies showing that subjects are able to compare confidence across visual and auditory tasks with the same precision as when comparing two trials within the same task (de Gardelle & Mamassian, 2014; de Gardelle, Le Corre, & Mamassian, 2016).

The notion that confidence may be compared between task domains to facilitate flexible decision making is consistent with the existence of both domain-general and domain-specific signatures of confidence in neuroimaging data reviewed above. A midline network with hubs in pMFC and vmPFC may provide a nexus for monitoring arbitrary task-set shifts. In contrast, task-specific confidence representations in lateral aPFC may allow the hierarchical control of decision-making in situations in which subjects need to regularly switch between tasks or strategies on the basis of their reliability (Donoso, Collins, & Koechlin, 2014; Morales, Lau, & Fleming, 2018). However, we note that the ability to compare confidence between domains does not itself imply that the representation of confidence at the neural level is domain-general, and more investigation is needed into neural signals that support such cross-task comparisons.

Another intriguing possibility is that the existence of midline hubs for confidence formation leads to domain-general representations of confidence which in turn act as priors on confidence in other domains. In behavioral studies, a confidence “leak” has been identified between a color and a symbol discrimination task, where confidence in one task influences confidence in the other regardless of actual performance. Notably, the ability to resist such leakage was positively correlated with lateral aPFC grey matter volume across subjects (Rahnev, Koizumi, McCurdy, D’Esposito, & Lau, 2015). It is important, however, to distinguish the notion of confidence leak (the temporal autocorrelation of confidence ratings when several tasks are interleaved) from metacognitive bias (the overall tendency to rate confidence higher or lower irrespective of performance). A larger confidence leak across trials is not necessarily linked to a higher or a lower metacognitive bias, but confidence leak may represent a mediating factor in explaining why metacognitive bias often generalizes across tasks (Ais et al., 2016; Baird et al., 2013). In addition, if confidence leak is large, confidence may become more loosely coupled to current performance in tasks in which autocorrelations in stimuli are absent, which could in turn decrease metacognitive efficiency. However, it remains to be explored whether confidence leak extends across task domains beyond perception. It could be that some domains are more susceptible to leak than others.

Ultimately, to understand the structure of metacognition across task domains we should aim to understand what functions metacognition provides to the system (the computational level in Marr’s taxonomy). The formation of accurate beliefs about performance is useful for learning, cognitive control and for social interaction (such as when communicating confidence to others) (Bang et al., 2017; Donoso, Collins, & Koechlin, 2014). But when is it advantageous to share this computation across modalities or inputs? We can think of two possible reasons. First, as stated above, if tasks are composed of arbitrary state-action mappings then it may be more computationally efficient to infer performance in a global, task-independent frame of reference, for use in control of future behavior (Donoso, Collins, & Koechlin, 2014). Second, inferences about performance in one domain may be useful as priors for performance in other domains to the extent they share instrumental characteristics. For instance, if I infer that I am very skilled at skiing, I might infer that I would also be good at similar sports such as ice-skating, but, on this basis, it would not be wise to also think I will be able to remember my to-do list. In other words, it may be useful to generalize confidence level across tasks according to their distance in task space. Conversely, overgeneralization might be maladaptive, and the extent of metacognitive generalization may itself constitute a stable individual difference. For example, people with depression tend to generalize more strongly from poor performance in one domain to other domains, in turn reinforcing a lower level of self-esteem and poorer self-efficacy that cuts across various areas of life (Bandura, 1977; Elliott et al., 1996; Stephan et al., 2016).

Implications of domain-specific alterations in metacognition for clinical populations

The study of metacognition provides an experimental window into our subjective estimates of our internal states. The explanatory potential of metacognition for mechanisms of pathogenesis and maintenance of mental illness is therefore considerable, and metacognitive deficits might be usefully measured in the clinic to guide...
assessments of metacognition in domains other than memory (e.g., Chapman et al., 2018). Greater knowledge of the relationship between patterns of metacognitive deficits and individual neuropsychiatric profiles may eventually allow development of personalized therapeutic approaches and suggest pathways to train resilience to mental illness (Moros, Paulus, & Maia, 2016).

Conclusions and future directions

Here we have reviewed the recent literature comparing the neurocognitive architecture of metacognition across domains. We have distinguished between the constructs of metacognitive efficiency (one’s sensitivity to fluctuations in task performance) and metacognitive bias (one’s overall confidence level, irrespective of performance). We have considered the importance of taking into account variations in task performance when measuring metacognition to allow meaningful comparisons across domains. In particular, it is critical in future studies to match task and stimulus characteristics for across-domain comparisons (e.g., Y/N vs. 2AFC). Finally, the use of a hierarchical Bayesian framework allows uncertainty in metacognitive efficiency parameters to be taken into account when examining correlations across domains (Fleming, 2017).

Our review of neuroimaging studies indicates a more nuanced picture of domain-generality than studies of behavior or individual differences, which have tended to argue for either domain-specific or domain-general aspects of metacognition, but not both. Recent studies suggest that both domain-specific and domain-general signals co-exist in the human brain, and that there may exist a gradient in which some tasks (such as different types of perceptual judgment) are more likely to rely on shared circuitry for metacognitive evaluation than others. Finally, we have highlighted the utility of computational models in providing a framework for understanding how confidence is formed across different tasks, and why it might be useful to maintain confidence in a common currency when switching between tasks. We suggest that the formation of confidence in one domain may provide useful priors on confidence formation in other domains. Notably the extent of such generalization itself could represent an individual difference, with extreme over-generalization possibly contributing to pervasive low self-efficacy often seen in depression and anxiety disorders.

Acknowledgments. The authors thank Robert Jagiello for contributing research to this article. A. McW. is a National Institute of Health Research-funded academic clinical fellow. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust 201472/Z/18/Z.

Conflicts of Interest. None of the other authors declare any conflicts of interest.


---

Downloaded from [https://www.cambridge.org/core](https://www.cambridge.org/core). IP address: 54.70.40.11, on 20 Sep 2019 at 10:46:38, subject to the Cambridge Core terms of use, available at [https://www.cambridge.org/core/terms](https://www.cambridge.org/core/terms).


