Stephen M. Fleming[1]

Metacognitive Psychophysics in Humans, Animals, and AI

A Research Agenda for Mapping Introspective Systems

Abstract: *Kammerer and Frankish (this issue) propose an exciting new research programme on the computational form of introspective systems. Pursuing this goal requires measures that can isolate introspective capacity from response biases and first-order processes. I suggest that metacognitive psychophysics is well placed to meet this challenge, allowing the mapping of introspective architectures in humans, animals, and artificial systems.*

1. Introduction

The capacity for a mind to interrogate itself via introspection is both beguiling and elusive. Introspection is commonplace — we wonder to ourselves whether we have a headache coming on, or muse about why we always react in a certain way when receiving an email from a particular colleague. But it is also elusive and slippery — the mechanisms of introspection themselves are opaque to introspection and hard to pin down using the tools of cognitive science. These

Correspondence:
Email: stephen.fleming@ucl.ac.uk

[1] Department of Experimental Psychology and Wellcome Centre for Human Neuroimaging, University College London, 26 Bedford Way, London, WC1H 0AP, UK.

Journal of Consciousness Studies, **30**, No. 9–10, 2023, pp. 113–28
DOI: 10.53765/20512201.30.9.113

challenges notwithstanding, we are now in an era of the mind sciences when a rigorous approach to reverse engineering introspection is flourishing. After false starts in the nineteenth century, where introspection was used as a (notoriously misleading) tool for the study of the mind, rather than the target of explanation itself, psychologists and philosophers are now increasingly adopting a healthily sceptical stance towards introspection. We no longer take it at face value, and instead attempt to use behavioural data to creep up on its inner workings.

Kammerer and Frankish (this issue; henceforth, K&F) advance a project that clears new ground in this research programme. Instead of restricting themselves to the features of human introspection, they ask 'what forms could introspective systems take?'. This question is allied with a computational functionalist stance — introspection may not be restricted to biological brains, and is instead conceived as a feature of information-processing systems. It also motivates an engineering-like approach to empirical research, in order to determine the types of processes that could enable a cognitive system to learn about its own mental states. This is an exciting new research direction, and one that has the potential to shed light on the introspective processes in non-human animals and AI systems.

My goals in the current commentary are threefold. First, I will argue that a psychophysical approach to the study of metacognition is well-placed to address many of the challenges raised by K&F. Doing so requires marshalling a defence of the psychophysical approach in general, which is often challenged as being too narrow and artificial — for instance relying on constrained paradigms with confidence ratings, rather than free-form narrative data. I will argue that, far from being a drawback, this is often an advantage, as it allows researchers to both tease out metacognitive processes from first-order processes, and develop computational models of the latent processes underpinning introspective judgments across different domains. These endeavours meet K&F's challenge of developing a 'minimal mind' approach to introspection.

Through this lens, I will briefly review some of the nascent findings in this literature that begin to address some of the open questions raised by the target article — including the link between introspection and theory of mind, the domain-generality of introspection, the neural architecture of introspection, and the influence of meditation on introspection. A powerful side-benefit of these research efforts is that there is now a robust ecosystem of computational models of metacognition,

which are being actively tested in both humans and animals. By formalizing metacognition within computational models, we can identify a range of possible introspective mechanisms. One salient open question concerns the format and structure of metacognitive representations — and, as we will see, neuroscientific investigations of confidence judgments are well placed to answer these questions. In a final section I explore how these aspects of metacognition are being built into AI systems under the umbrella of 'introspective robotics'.

2. Definitions: Metacognition and Introspection

Let us start with some definitions. K&F define introspection as 'a process by which a cognitive system represents its own current mental states, in a manner that allows the information to be used for online behavioural control'. This is closely overlapping with the definition of metacognition typically given in psychology — where metacognition is the set of capacities through which a cognitive subsystem is evaluated or represented in the service of self-regulation and/or for communication to others (Nelson and Narens, 1990; Proust, 2013).

The term 'metacognition', however, carries unhelpful baggage when our goal is to pursue research on introspection. A colloquial definition of metacognition is often given as 'cognition about cognition' or 'thinking about thinking'. This is unfortunately restrictive and limits the targets of introspection to ongoing thoughts, rather than encompassing percepts, memories, and mental states in general. The colloquial definition is also misleading as empirical research on metacognition is increasingly focused on understanding how human subjects can reflect on (introspect) specific first-order judgments. For instance, the rapidly developing field of perceptual metacognition focuses on how subjects self-evaluate first-order perceptual judgments (see Rahnev, 2021, for a recent review). It is this broader notion of metacognition that I have in mind in this article, and one that I believe is closely allied to K&F's definition of an introspective system. I will use the terms 'metacognition' and 'introspection' interchangeably in what follows.

3. Overcoming Challenges in the Empirical Study of Introspection

As K&F point out, a key challenge in developing an empirical science of introspection is to find a way to isolate inference on the

mechanisms of introspection from changes in first-order processes or response biases.

First-order confounds are particularly pernicious in consciousness science (Lau, 2022). Imagine trying to identify patterns of neural activity covarying with introspection of a perceptual experience of the colour red. Naïvely we might set up this experiment by showing someone a patch of red on a screen, and ask them to introspect about the colour experience they are having. The obvious problem here is that a change in my introspective judgment is correlated with both a change in both a) sensory input and b) a first-order perceptual experience of red in the world. Any change in neural activity we observe could be a result of first-order processing, introspection, or both.

Conversely, imagine a patient who has brain damage or dementia, and appears to be unable to introspect that they have a memory problem (in clinical parlance, they would lack insight). This may be an isolated deficit in introspection. But it could also be a secondary consequence of the memory problem itself: without sufficient signal from memory systems, introspective systems are unable to form a reliable belief about performance capacity.

Another challenge in developing an empirical science of introspection is to rule out the possibility of response biases, either within- or between-individuals. For instance, if asked to reflect on how painful an experience is, it is possible that two individuals could have similar first-order pain experiences, and introspect them in a similar manner, but differ considerably when reporting their experience on a scale or describing its qualities to others.

For an empirical science of introspection to get off the ground, we need to find ways of dealing with these two confounds of first-order processes and response biases. In the next section, I will describe how signal detection theoretic models of metacognition meet this challenge.

3.1. Psychophysical assays of introspection

Signal detection theory (SDT) is a mainstay of cognitive psychology. At its core, SDT is a framework for distinguishing between 'sensitivity' — how well a system can discriminate between two or more states of the world — and 'response bias' — the often idiosyncratic criterion adopted for making a report. Consider a task in which you are asked to detect a faint light in a dark room. If the intensity of the

light is reduced sufficiently, sometimes you will make errors — saying you saw the light when it was off, and saying the light was off when it was in fact on. All else being equal, your sensitivity to perceiving the light should increase with the brightness of the light. However, whether an observer says that they saw the light on any given trial is a joint outcome of their capacity to discriminate the light, and the criterion they adopt for reporting — are they conservative, only reporting they saw something when they have strong evidence, or are they liberal, responding 'yes' with only limited evidence? SDT provides a formal, mathematical framework for separating sensitivity from response biases in behavioural data.

This application of SDT is first-order — it seeks to characterize an observer's sensitivity in distinguishing (or remembering) different aspects of the world. But it is also possible to extend the logic of SDT to introspective judgments. The idea here is that we can determine people's bias and sensitivity with respect to first-order mental states. This extension of SDT, known as 'type 2' SDT, was first proposed by Clarke, Birdsall and Tanner (1959), and was then resurrected by Galvin and colleagues in a landmark 2003 paper (Galvin *et al.*, 2003). In type 1 SDT, the target that is being detected is a property of the outside world: the light is either on or off. In type 2 SDT, the target that is being judged is one's own first-order judgment of the world. Often the type 2 (metacognitive) judgment takes the form of a confidence rating — confidence in a first-order response being correct.

How do introspective bias and sensitivity manifest in this context? Introspective bias denotes how confident we are overall about a particular type 1 judgment, and may be subject to idiosyncratic criterion effects — some people might be more or less confident about their perceptual judgments than others due to general factors such as being dogmatic or anxious (Rouault *et al.*, 2018b; Schulz *et al.*, 2020). Introspective (or metacognitive) sensitivity, on the other hand, refers to whether the metacognitive judgment systematically tracks — is sensitive to — the first-order state. Introspective sensitivity can be assessed by measuring the separation between the distributions of confidence in correct and incorrect trials: a healthy separation between these distributions is consistent with good metacognition, whereas greater overlap belies a lack of introspective access to fluctuations in performance, and therefore poor metacognition. The overlap between confidence distributions on correct and incorrect trials can be quantified as the area under the type 2 ROC, or via model-based estimation of meta-d′ (Maniscalco and Lau, 2012). Good metacognitive

sensitivity is consistent with introspection being sensitive to the same or similar evidence that drove first-order behaviour, whereas poor metacognitive sensitivity reveals a decoupling between introspective judgments and first-order processes.

We can now see how adopting a type 2 SDT framework allows controlling for response bias in the empirical study of introspection. A critical point is that, just like perceptual (first-order) sensitivity, introspective sensitivity cannot be derived from a single trial or report. If I say 'yes I saw the light', it could be that I have good perceptual sensitivity to a faint light, or poor perceptual sensitivity together with a very liberal criterion for saying I saw something. Similarly, if I say 'I am confident that my judgment that the patch is red is correct', it might be that I have good metacognitive sensitivity about my fine-grained colour discrimination, or poor metacognitive sensitivity together with a tendency towards overconfidence. In the case of perception research, it is only by applying psychophysical analysis to data collected over many trials that we can identify the latent perceptual sensitivity of the system. Similarly, in introspection and metacognition research, it is only by applying type 2 SDT models to data collected over many trials that we can securely identify the *introspective* sensitivity of the system.

One powerful side-benefit of quantifying introspective sensitivity within SDT is that it becomes straightforward to control for contributions of first-order processes. In particular, d′ (first-order performance) and meta-d′ (metacognitive sensitivity) are in the same units. The ratio meta-d′/d′ (also known as metacognitive efficiency) is therefore a useful summary statistic that tells us about the introspective capacity of the system, controlling for first-order sensitivity. Under an ideal observer model, d′ and meta-d′ change together. This is because stronger sensory evidence both increases the detectability of the signal (d′), and also makes correct and incorrect detections themselves easier to detect (meta-d′). To pump the intuition for this relationship, consider a subject participating in a visual psychophysics experiment with their eyes closed. They will perform at chance level in the task, and also be unable to introspect about whether they were right or wrong on any given trial — as they have no basis on which to make this judgment. Now imagine they open their eyes: their first-order performance increases, but so does their metacognitive sensitivity, as now any errors they make are more easily detectable. The upshot of this expected relationship between d′ and meta-d′ is that it is hard to pin a change in introspective judgment to introspection alone. However, if

we find a case in which meta-d′ changes without an accompanying change in d′, then we can be more confident that we have identified a selective change in introspective processing. We will encounter cases such as this below.

3.2. In praise of artifice

There is a legitimate concern that the measures and frameworks developed for the psychophysical study of metacognition elide naturalistic introspective content by focusing on contrived experimental paradigms. Indeed, K&F worry that 'we shall need to focus on more complex introspective tasks if we are to detect potential variation in the *content* of introspective representations'. Elsewhere I have written about the need for ensuring that the laboratory study of metacognition appropriately encompasses all the facets of the construct — including personal-level self-knowledge and beliefs (Katyal and Fleming, 2023). For probing the mechanisms of introspective systems, however, a psychophysical approach holds considerable promise. To make this case, it is useful to compare the scientific progress that has been made in understanding first-order perceptual systems such as vision.

Vision science since the first half of the twentieth century has made great strides in understanding the components of the primate visual system. The approach taken here has been to conduct psychophysics using minimal stimuli — such as points of light, moving bars, Gabor patches, sinusoidal gratings, and so on. These experiments typically require many hundreds or even thousands of trials, in order to construct psychometric functions for different properties of the stimulus — for instance, relating detection sensitivity to variation in contrast, or using adaptation designs to ask how adapting to one feature of a stimulus (such as spatial frequency) affects subsequent judgments about other stimuli. From these minimal experimental designs, a great deal has been learned about the architecture of primate vision. For instance, we now know that there are parallel 'channels' for processing different spatial frequencies and orientations within an image (Graham, 1989). In turn, classical neurophysiology has revealed single-neuron substrates for tuning curves over the different component features identified by psychophysical experiments. These findings have been built upon to understand more broadly how viewpoint-invariant object recognition is achieved by a neural hierarchy organized along the ventral visual stream (DiCarlo, Zoccolan and Rust, 2012).

While this research endeavour is far from complete, together these findings have led to a 'standard model' in vision science — in which incoming information is processed through a set of retinotopically organized spatio-temporal filters tuned to different orientations and spatial frequencies, before subsequent layers with larger receptive fields enable linear decoding of object-level features. This model is, in turn, remarkably successful at predicting neural and behavioural responses to more naturalistic images (e.g. Freeman and Simoncelli, 2011; Mante, Bonin and Carandini, 2008). Notably, however, the experiments that were done to obtain this mechanistic picture were, frankly, long, boring, and very unlike the richness of regular human vision (Rust and Movshon, 2005).

In comparison, a psychophysics of metacognition remains in its infancy. But the isolation of metacognitive mechanisms in well-controlled psychophysical paradigms holds similar promise for developing a detailed understanding of introspective computation (Peters, 2022). Importantly, the appropriate starting point for these experiments is not the richness of everyday introspective content — just as perceptual psychophysics does not start with the richness of everyday vision. Instead, once a core mechanism has been identified, it can be validated in naturalistic contexts, and (ideally) extended beyond the paradigm or perceptual/cognitive domain in which it was originally discovered.

4. Emerging Findings from a Psychophysics of Introspection

Having described and defended a psychophysical approach to introspection, I will now briefly summarize some of the emerging findings in this literature that speak to some of the open questions raised by K&F.

A first advantage of adopting a metacognitive psychophysics approach is that it does not presuppose a relationship between theory of mind (ToM) and introspection, as the judgments being made are all self-directed. This enables unbiased tests of hypotheses relating introspection to ToM, with recent studies finding initial evidence of a link (Nicholson *et al.*, 2021). For instance, Nicholson *et al.* found that performance-controlled metrics of metacognitive efficiency were impaired in autism spectrum disorder when confidence in performance was measured using explicit confidence ratings, but not when confidence was measured using implicit gambles of the type often used in

animal metacognition research (*ibid.*). Moreover, in neurotypical subjects, a concurrent task involving mentalizing interferes with explicit metacognitive efficiency, but a similarly demanding non-social concurrent task does not. While there is more work to be done to understand the computational overlap between metacognition and ToM, these findings support a model in which explicit, conscious introspection (isolated from variation in the strength of first-order processes) covaries with ToM, whereas implicit metacognition does not (Fleming, 2021; Carruthers and Williams, 2022).

Second, the structure of covariation in metacognitive efficiency across different tasks can be estimated, allowing questions about the number of distinct introspective systems to be tackled in both human and non-human animals. In humans, there are now several such studies which point towards a shared domain-general capacity (Rouault *et al.*, 2018a; Mazancieux *et al.*, 2020), with occasional outlying domains or islands of introspective ability, such as for pain (Beck *et al.*, 2019).

Third, precise, quantitative measures of metacognition allow tests of how introspective capacity varies across individuals and cultures (Heyes *et al.*, 2020; van der Plas *et al.*, 2022), how such capacities are affected by meditation (Baird *et al.*, 2014) and introspective training (Carpenter *et al.*, 2019; Rouy *et al.*, 2022), and how introspection may be impaired by brain damage or disease (Pannu and Kaszniak, 2005). This research programme has uncovered stable individual differences in introspective capacity linked to the structure and function of the primate frontopolar cortex, with lesions to this area and/or networks interconnected with it leading to systematic impairments in introspection, without changes in first-order performance (Fleming *et al.*, 2014; Miyamoto *et al.*, 2018).

Finally, and perhaps most importantly for the purposes of the current commentary, a psychophysics of metacognition provides a comprehensive test bed for detailed, quantitative models of introspective computation. There are several competing models of how confidence is formed, and how different potential sources of noise and distortion affect introspective judgments. Reviewing each distinct modelling approach is beyond the scope of this article. As a brief example, Boundy-Singer, Ziemba and Goris (2023) proposed that a key source of noise in confidence estimates is due to an imprecise higher-order representation of the precision of first-order evidence. They obtained initial support for their model in a carefully controlled visual psychophysics experiment, and then validated it on data derived

from the Confidence Database — a large, open-source database of millions of introspective judgments provided across a number of distinct experiments (Rahnev *et al.*, 2020). The ambition of this research effort is to arrive at a detailed picture of the computational underpinnings of human introspection and metacognition, just as the field of vision science has characterized the computational underpinnings of human vision. As noted above, progress in identifying the building blocks of introspection from constrained laboratory tasks will shed light on the general principles governing naturalistic introspection. To date, this research programme is yet to make contact with the broader theoretical frameworks put forward to account for introspection in philosophy (e.g. K&F's Figure 1). An attractive research direction here is to ask how different potential models of introspection articulated at the psychological level map onto the computational components of metacognitive judgments.

The research endeavour outlined in the previous section has largely been pursued in humans. However, it also holds promise for achieving K&F's vision of a 'minimal mind' approach to understanding introspective systems beyond humans. So far this has gained prominence in the neuroscientific investigation of animal metacognition. As an example, work carried out in Adam Kepecs' lab has begun to carefully dissect the neural architecture underpinning perceptual metacognitive judgments in rodents (Kepecs and Mainen, 2012). To achieve this, rats and mice learn to signal their confidence by waiting for a reward that is conditional on the correctness of a first-order perceptual judgment (in these studies, the first-order judgment is often an olfactory or auditory discrimination). As the reward is occasionally withheld, even on correct trials, it is adaptive for the animal to give up and move onto a new trial rather than waiting indefinitely for a reward that might never arrive. Under an ideal observer model, the amount of time that should be invested waiting is proportional to confidence in the initial choice. Rodent behaviour shows this signature, and confidence (as derived from waiting time) shows classical signatures of metacognition — being higher after correct than incorrect decisions, and scaling with evidence strength (Sanders, Hangya and Kepecs, 2016). Notably, distinct neural signatures of confidence have been observed in the rodent orbitofrontal cortex (OFC) using single-unit recordings, and lesions to this brain area disrupt confidence but not first-order performance (Lak *et al.*, 2014).

One limitation is that the models and experimental paradigms being tested in the animal literature are often inherited (with clever back-

translation into non-verbal forms) from the human literature, rather than developed *de novo* for the study of potentially distinct forms of introspection. There is also a restricted range of species under regular test in experimental neuroscience (primarily rodents and non-human primates), rendering investigation of introspective architectures currently limited. In contrast, K&F's proposal for investigating radical computational architectures for introspection provides an exciting opportunity to reverse this usual direction of travel and think afresh about how introspection and metacognition may work in non-human systems. This is already happening in artificial systems, which I turn to now.

5. Comparing the Format of Introspective Computation in Humans, Animals, and AI

The engineering mindset of AI research means that architectures are typically built from the ground up with function in mind, rather than with the goal of mimicking models abstracted from the human case. Nevertheless, the importance of metacognition and 'introspective robotics' has increasingly been emphasized. To operate AI devices in novel environments and in collaboration with others, appropriate sensitivity to first-order uncertainty ('knowing what you do not know') is essential to avoid the pitfalls of overconfidence, especially when faced with input data that fall outside of the training distribution. For instance, in one study, autonomous drones were trained to navigate around a cluttered environment. A second neural network was trained to detect the likelihood of crashes during test flights, allowing the fully trained architecture to adaptively bail out of navigation decisions that it predicted may end in failure (Daftry *et al.*, 2016). However, with notable exceptions (Webb *et al.*, 2023), these architectures have typically not been investigated with the kind of metacognitive psychophysics that has been deployed to characterize introspective computation in humans and non-human animals.

In the architectures described above, the format of minimal introspective systems is often very lean — they track the overall performance of the system, but do not re-represent its internal states in a finer-grained sense, or have any knowledge of the source of the performance decrement (Pasquali, Timmermans and Cleeremans, 2010; Webb *et al.*, 2023). Interestingly, similar coarse-grained neural codes for confidence have been observed in a recent study of the neural architecture supporting metacognition in rats. Masset and

colleagues found that neural populations tracking confidence (waiting time) in the rodent OFC showed similar signatures in both an olfactory and auditory first-order task — indicative of a generic neural code for confidence that could be used to guide behaviour, but which does not re-represent the content or modality of the first-order state (Masset *et al.*, 2020).

In humans, at least, it seems clear that introspection enables much finer-grained assessments about the presence and reliability of different mental states. In the human brain, both domain-general and domain-specific confidence signals are observed in the prefrontal cortex (Morales, Lau and Fleming, 2018). Such a picture is suggestive of the human system combining both a more generic, coarse-grained performance-tracker with a more granular capacity for introspection of particular domains or mental states. This may be achieved by estimates of uncertainty in first-order representations being 'tagged' with their source. Indeed, there is recent evidence that neural signatures of confidence in human brain imaging track the precision of stimulus-specific neural codes in the early visual cortex — consistent with a mechanism that reads out the reliability of first-order mental states, allowing finer-grained introspection (Geurts *et al.*, 2022). Alternatively, it may be that introspection is achieved by combining relatively low-dimensional signals of vividness or intensity with first-order content, without requiring higher-order re-representation of such content (Teng, 2022). New paradigms that can measure introspective judgments for a variety of first-order states in combination with brain imaging are well-placed to make progress on this issue.

A distinct open question concerns the functional role performed by introspective systems of different levels of granularity. If the main function of introspective systems in AI is to allow performance prediction and error monitoring, then a low-dimensional metacognitive signal may suffice. But if there is a functional need to communicate the internal states of the system to others, then a more granular introspective system may be required. In biological agents, it is plausible that the evolutionary pressures associated with developing this more granular form of mental-state introspection in the service of social communication only emerged in more complex systems (Fleming, 2021). But with AI, this coupling between first-order complexity and the richness of introspection may be violated, leading to a new space of intermediate cases in which it is useful for AI to be able to introspect about some aspects of its functioning and reliability but not others. For instance, when arriving at a road junction, it may be

adaptive for a self-driving car to communicate confidence in its representation of other drivers' intentions to grease the wheels of social coordination, just as human drivers intuitively do with glances, looks, and hesitant moves forward. But it may have no similar need to represent its confidence in its (first-order) representation of battery status.

6. Conclusions

The 'minimal mind' approach to introspective systems advocated by K&F meshes neatly with emerging work seeking to identify general computational principles supporting metacognitive function. This research programme takes inspiration from the success of psychophysics in vision science, and aspires to identify the building blocks of introspective capacity in humans and non-human animals. A science of metacognition is now poised to expand its scope to include artificial systems, allowing the investigation of hitherto undiscovered forms of introspective computation.

Acknowledgments

SMF is a CIFAR Fellow in the Brain, Mind & Consciousness Program, and funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme Trust.

References

Baird, B., Mrazek, M.D., Phillips, D.T. & Schooler, J.W. (2014) Domain-specific enhancement of metacognitive ability following meditation training, *Journal of Experimental Psychology. General*, **143** (5), pp. 1972–1979.

Beck, B., Pena-Vivas, V., Fleming, S. & Haggard, P. (2019) Metacognition across sensory modalities: Vision, warmth, and nociceptive pain, *Cognition*, **186**, pp. 32–41.

Boundy-Singer, Z.M., Ziemba, C.M. & Goris, R.L.T. (2023) Confidence reflects a noisy decision reliability estimate, *Nature Human Behaviour*, **7** (1), pp. 142–154.

Carpenter, J., Sherman, M.T., Kievit, R.A., Seth, A.K., Lau, H. & Fleming, S.M. (2019) Domain-general enhancements of metacognitive ability through adaptive training, *Journal of Experimental Psychology: General*, **148** (1), pp. 51–64.

Carruthers, P. & Williams, D.M. (2022) Model-free metacognition, *Cognition*, **225**, 105117.

Clarke, F., Birdsall, T. & Tanner Jr., W. (1959) Two types of roc curves and definitions of parameters, *The Journal of the Acoustical Society of America*, **31** (5), pp. 629–630.

Daftry, S., Zeng, S., Bagnell, J.A. & Hebert, M. (2016) Introspective perception: Learning to predict failures in vision systems, in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 1743–1750.

DiCarlo, J.J., Zoccolan, D. & Rust, N.C. (2012) How does the brain solve visual object recognition?, *Neuron*, **73** (3), pp. 415–434.

Fleming, S.M. (2021) *Know Thyself: The Science of Self-Awareness*, New York: Basic Books.

Fleming, S.M., Ryu, J., Golfinos, J.G. & Blackmon, K.E. (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions, *Brain*, **137** (Pt 10), pp. 2811–2822.

Freeman, J. & Simoncelli, E.P. (2011) Metamers of the ventral stream, *Nature Neuroscience*, **14** (9), pp. 1195–1201.

Galvin, S.J., Podd, J.V., Drga, V. & Whitmore, J. (2003) Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions, *Psychonomic Bulletin & Review*, **10**, pp. 843–876.

Geurts, L.S., Cooke, J.R.H., van Bergen, R.S. & Jehee, J.F.M. (2022) Subjective confidence reflects representation of Bayesian probability in cortex, *Nature Human Behaviour*, **6** (2), pp. 294–305.

Graham, N.V.S. (1989) *Visual Pattern Analyzers*, Oxford: Oxford University Press.

Heyes, C., Bang, D., Shea, N., Frith, C.D. & Fleming, S.M. (2020) Knowing ourselves together: The cultural origins of metacognition, *Trends in Cognitive Sciences*, **24** (5), pp. 349–362.

Katyal, S. & Fleming, S. (2023) Construct validity in metacognition research: Balancing the tightrope between rigor of measurement and breadth of construct, *PsyArxiv*. doi: 10.31234/osf.io/etjqh

Kepecs, A. & Mainen, Z.F. (2012) A computational framework for the study of confidence in humans and animals, *Philosophical Transactions of the Royal Society B*, **367** (1594), pp. 1322–1337.

Lak, A., Costa, G.M., Romberg, E., Koulakov, A.A., Mainen, Z.F. & Kepecs, A. (2014) Orbitofrontal cortex is required for optimal waiting based on decision confidence, *Neuron*, **84** (1), pp. 190–201.

Lau, H. (2022) *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*, New York: Oxford University Press.

Maniscalco, B. & Lau, H.C. (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings, *Consciousness and Cognition*, **21** (1), pp. 422–430.

Mante, V., Bonin, V. & Carandini, M. (2008) Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli, *Neuron*, **58** (4), pp. 625–638.

Masset, P., Ott, T., Lak, A., Hirokawa, J. & Kepecs, A. (2020) Behavior- and modality-general representation of confidence in orbitofrontal cortex, *Cell*, **182** (1), pp. 112–126.

Mazancieux, A., Fleming, S.M., Souchay, C. & Moulin, C.J.A. (2020) Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks, *Journal of Experimental Psychology: General*, **149** (9), pp. 1788–1799.

Miyamoto, K., Setsuie, R., Osada, T. & Miyashita, Y. (2018) Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates, *Neuron*, **97** (4), pp. 980–989.

Morales, J., Lau, H. & Fleming, S.M. (2018) Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex, *The Journal of Neuroscience*, **38** (14), pp. 3534–3546.

Nelson, T.O. & Narens, L. (1990) Metamemory: A theoretical framework and new findings, *The Psychology of Learning and Motivation: Advances in Research and Theory*, **26**, pp. 125–173.

Nicholson, T., Williams, D.M., Lind, S.E., Grainger, C. & Carruthers, P. (2021) Linking metacognition and mindreading: Evidence from autism and dual-task investigations, *Journal of Experimental Psychology: General*, **150** (2), pp. 206–220.

Pannu, J. & Kaszniak, A. (2005) Metamemory experiments in neurological populations: A review, *Neuropsychology Review*, **15** (3), pp. 105–130.

Pasquali, A., Timmermans, B. & Cleeremans, A. (2010) Know thyself: Metacognitive networks and measures of consciousness, *Cognition*, **117** (2), pp. 182–190.

Peters, M.A. (2022) Towards characterizing the canonical computations generating phenomenal experience, *Neuroscience & Biobehavioral Reviews*, **142**, 104903.

Proust, J. (2013) *The Philosophy of Metacognition: Mental Agency and Self-Awareness*, New York: Oxford University Press.

Rahnev, D. (2021) Visual metacognition: Measures, models, and neural correlates, *The American Psychologist*, **76** (9), pp. 1445–1453.

Rahnev, D., Desender, K., Lee, A.L.F., Adler, W.T., Aguilar-Lleyda, D., Akdogan, B., Arbuzova, P., Atlas, L.Y., Balcı, F., Bang, J.W., Begue, I., Birney, D.P., Brady, T.F., Calder-Travis, J., Chetverikov, A., Clark, T.K., Davranche, K., Denison, R.N., Dildine, T.C., Double, K.S., Duyan, Y.A., Faivre, N., Fallow, K., Filevich, E., Gajdos, T., Gallagher, R.M., de Gardelle, V., Gherman, S., Haddara, N., Hainguerlot, M., Hsu, T.-Y., Hu, X., Iturrate, I., Jaquiery, M., Kantner, J., Koculak, M., Konishi, M., Koß, C., Kvam, P.D., Kwok, S.C., Lebreton, M., Lempert, K.M., Ming Lo, C., Luo, L., Maniscalco, B., Martin, A., Massoni, S., Matthews, J., Mazancieux, A., Merfeld, D.M., O'Hora, D., Palser, E.R., Paulewicz, B., Pereira, M., Peters, C., Philiastides, M.G., Pfuhl, G., Prieto, F., Rausch, M., Recht, S., Reyes, G., Rouault, M., Sackur, J., Sadeghi, S., Samaha, J., Seow, T.X.F., Shekhar, M., Sherman, M.T., Siedlecka, M., Skora, Z., Song, C., Soto, D., Sun, S., van Boxtel, J.J.A., Wang, S., Weidemann, C.T., Weindel, G., Wierzchon, M., Xu, X., Ye, Q., Yeon, J., Zou, F. & Zylberberg, A. (2020) The Confidence Database, *Nature Human Behaviour*, **4** (3), pp. 317–325.

Rouault, M., McWilliams, A., Allen, M.G. & Fleming, S.M. (2018a) Human metacognition across domains: Insights from individual differences and neuroimaging, *Personality Neuroscience*, **1**.

Rouault, M., Seow, T., Gillan, C.M. & Fleming, S.M. (2018b) Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance, *Biological Psychiatry*, **84** (6), pp. 443–451.

Rouy, M., de Gardelle, V., Reyes, G., Sackur, J., Vergnaud, J.C., Filevich, E. & Faivre, N. (2022) Metacognitive improvement: Disentangling adaptive training from experimental confounds, *Journal of Experimental Psychology: General*, **151**, pp. 2083–2091.

Rust, N.C. & Movshon, J.A. (2005) In praise of artifice, *Nature Neuroscience*, **8** (12), pp. 1647–1650.

Sanders, J.I., Hangya, B. & Kepecs, A. (2016) Signatures of a statistical computation in the human sense of confidence, *Neuron*, **90** (3), pp. 499–506.

Schulz, L., Rollwage, M., Dolan, R.J. & Fleming, S.M. (2020) Dogmatism manifests in lowered information search under uncertainty, *Proceedings of the National Academy of Sciences*, **117** (49), pp. 31527–31534.

Teng, L. (2022) A metacognitive account of phenomenal force, *Mind & Language*, advance online publication. doi: 10.1111/mila.12442

van der Plas, E., Zhang, S., Dong, K., Bang, D., Li, J., Wright, N.D. & Fleming, S.M. (2022) Identifying cultural differences in metacognition, *Journal of Experimental Psychology: General*, **151** (12), pp. 3268–3280.

Webb, T.W., Miyoshi, K., So, T.Y., Rajananda, S. & Lau, H. (2023) Natural statistics support a rational account of confidence biases, *Nature Communications*, **14** (1), art. 3992.