Philosophical Transactions of the Royal Society B (Biological Sciences). Metacognition: computation, neurobiology and function. Organized and edited by Stephen M. Flemming, Raymond J. Dolan and Christopher D. Frith.

Metacognition and consciousness

Stuart WG Derbyshire, Reader in Psychology, University of Birmingham, B15 2TT, UK
Email: s.w.derbyshire@bham.ac.uk

Not so long ago, discussions of consciousness were considered, at best, an entertaining distraction. In the International Dictionary of Psychology, for example, Stuart Sutherland (1995) declared of consciousness: "Nothing worth reading has been written about it." Fifteen years later and consciousness is all the rage with new journals, research institutes and conferences dedicated to uncovering and detailing the mechanisms of conscious thought. A recent issue of the prestigious Philosophical Transactions of the Royal Society B continues the trend.

The issue comprises an introduction followed by twelve articles that variously address the mechanisms and nature of metacognition, the relationship of metacognition to conscious thought, the possibility of metacognition and conscious thought in animals and the neural correlates or causes of metacognition and conscious thought. In brief, there appears a general consensus across the authors that metacognition involves cognition about cognition (and not mere computation or association) but some dispute over whether metacognition is, by itself, a form of consciousness or gives rise to consciousness through emergence. Those that write about animals also address a very similar dispute. In general, there is consensus that at least some animals are capable of metacognition but there is much less certainty whether that means animals have conscious thoughts. Finally, those authors that address the brain generally concur that metacognition involves various regions of the prefrontal cortex but whether that knowledge explains or advances the understanding of metacognition is debated.

Fleming, Dolan and Frith (2012) explain in the introduction that metacognition essentially means cognition about cognition - cognitions that therefore go beyond the immediate content of any single cognition. An obvious example of metacognition is reflection, where a person considers their thinking, for example, as a means to monitor how well they are understanding or performing a task. Examples based on reflection tend to equate metacognition with consciousness itself and with introspection. But as Overgaard and Sandberg (2012) observe, metacognition and introspection are differently defined. Any cognitive state that is about another cognitive state is metacognitive and it does not necessarily have to be conscious. Introspection, in contrast, is directed towards specific conscious states and is necessarily a conscious process; unconscious introspection does not make obvious sense.

A good example of a metacognition that may not be conscious is uncertainty. Smith, Couchman and Beran (2012) describe a series of animal studies that involve deliberately difficult trials intended to create uncertainty or 'doubts' about performance. When the animal is given a response option that allows them to decline a difficult trial (and, typically, perform a simpler task for a smaller reward) the uncertainty or doubt can be observed behaviourally. Forfeiting a potential higher reward by opting to perform the simpler task expresses the animal's uncertainty about their task performance. In the first study of this kind, a dolphin was trained to respond to a high tone (2100 Hz) versus a low tone. The range of the low tone included tones very close to the high tone and as the low tone approached the dolphin's psychophysical limit of discrimination (around 2086 Hz) the dolphin began to decline the trials and display other evidence of uncertainty such as slowing, wavering and hesitating. Humans perform similarly at this task and subjectively report that their uncertainty reflects a conscious, metacognitive, state of uncertainty.

The authors spend some time going through various mechanistic and behavioural interpretations for these kinds of findings. It has been suggested, for example, that the discrimination follows a computational pattern that can be described using signal detection models and thus the animal's behaviour is computational rather than metacognitive (or even cognitive). Smith *et al.* rightly point out that just because a behaviour can be explained computationally does not mean that the behaviour is empty of psychological content. It is the model that is psychologically empty because it is purely mathematical but that does not imply, much less demonstrate, a low-level information - processing description of the behaviour: it implies no information-processing description. Similarly, a behavioural focus on reward-maximisation is also psychologically empty. Reward-maximisation may be the animal's meta-cognitive goal just as humans may consciously declare their pursuit of reward-maximisation via a particular strategy (such as opting out of uncertain trials to take an easier, more definite, reward). Again, reward-maximisation is a

description of the behaviour that does not imply, much less demonstrate, a low-level information-processing description of the behaviour: it implies no information-processing description.

On the whole, Smith *et al.* provide a strong case that at least some animals have metacognitive abilities that involve evaluations of ongoing cognitive processing to induce changes in behaviour. The more difficult issue is whether that amounts to a reflective, conscious, state of uncertainty. On this, 'difficult problem of scientific inference' (p. 1298) Smith *et al.* appear to be on the fence. I cannot sit with them because I see no means by which metacognitive processes might be reflected upon by animals. Formulating uncertainty involves concepts - 'right', 'wrong', 'different', 'difficult' - that are rooted in general beliefs about actual and possible states of the world that no dolphin or other animal could have. In short, animals lack a symbolic system that might enable isolation of one thought from another thought, one sensation from another sensation, a thought from a sensation and so on. Consequently, there appears to be no means by which metacognition could bubble up into conscious reflection for animals (Tallis, 2004).

Whatever might be happening in the animal mind, it is clear that conscious reflection does bubble up in the human mind. How that happens remains one of the hardest questions posed by contemporary philosophy and neuroscience (Tallis,2004; 2005; Fodor, 2007). It is captured well by Timmerman *et al.* (2012):

"[For consciousness] knowledge in the system has to become knowledge for the system. First-order systems - those that merely transform, however appropriately, inputs into outputs - can never know *that* they know: they simply lack the appropriate machinery... Sensitivity does not require consciousness in any sense. A thermostat can be appropriately characterized as being sensitive to temperature... Sensitivity can involve highly sophisticated knowledge, and even learned knowledge, but such knowledge is always first-order knowledge, it is always knowledge that is necessarily embedded in the very same causal chain through which processing occurs... *Awareness*, on the other hand, always seems to minimally entail the ability of knowing *that* one knows." (p. 1413, emphasis in the original).

The difficulty is that consciousness is always about something whereas the material world and computations are not about anything: they just are. Consequently, trying to transition from parts of the brain and computations to conscious awareness automatically imports something that is not a part of the brain or computations and so it starts to feel assertive or magical.

Despite being aware of the problems of basing solutions to the issue of consciousness on computational processes, the solution proposed by Timmerman *et al.* involves at least some assertion and magical thinking. Their proposal of three interwoven loops linking the brain with itself, the brain with the world and the brain with other agents producing consciousness is an attempt to pull consciousness out of the brain as an emergent property of interactions with the outside world. But that only displaces the hard question of how we become conscious from the computations of the brain to the computations of the brain in association with the world and other agents. I am not sure that gets us much beyond assertion and magic.

Neural theories of consciousness do not fare any better. It is almost certainly the case that the PFC is necessary for consciousness, as two papers in the current volume argue (Fleming and Dolan, 2012; Rosenthal, 2012), but being necessary is not the same as being causal. The proposed causal mechanisms, involving input from non-conscious monitoring loops or the rendering of implicit knowledge explicit, remain extremely sketchy. Perhaps more promising is the suggestion, pursued by Bahrami *et al.* (2012), Metcalfe *et al.* (2012) and Timmerman *et al.* (2012), that there is some form of common representational code that binds our self to our bodies and our bodies to other selves and that also grounds and defines our conscious awareness. Patients with schizophrenia, for example, appear unable to focus on internal cues that clue the rest of us into whether we are controlling our actions or not. Thus, being in control, having intentionality and volition may require bodily feedback that steps beyond neural loops. And it seems likely that the community of minds, which allows language and culture to flourish, certainly has a profound impact on conscious thinking. Understanding this may take us away from the isolated brain and some of the problems of associating consciousness with a brain rather than explaining consciousness as a sociodevelopmental process.

Jerry Fodor (2007) recently suggested that, at the moment, we can't even imagine a solution to the hard problem of how consciousness comes about. The authors of this themed issue of Philosophical Transactions have certainly risen to that challenge by imagining several solutions. Maybe none will stick but, at the very least, Sutherland can no longer claim that nothing worth reading has been written about consciousness.

References:

Bahrami B, Olsen K, Bang, D, Roepstorff A, Rees G, Frith C. What failure in collective decision-making tells us about metacognition. Phil Trans R Soc B (2012); 367: 1350-1365.

Fleming SM, Dolan RJ, Frith CD. Metacognition: computation, biology and function. Phil Trans R Soc B (2012); 367: 1280-1286.

Fleming SM, Dolan RJ. The neural basis of metacognitive ability. Phil Trans R Soc B (2012); 367: 1338-1349.

Fodor J. Headaches have themselves. London Review of Books (2007); 29: 9-10.

Metcalfe J, Van Snellenberg JX, DeRosse P, Balsam P, Malhotra AK. Judgements of agency in schizophrenia: an impairment in autonoetic metacognition. Phil Trans R Soc B (2012); 367: 1391-1400.

Overgaard M, Sandberg K. Kinds of access: different methods for report reveal different kinds of metacognitive access. Phil Trans R Soc B (2012); 367: 1287-1296.

Rosenthal D. Higher-order awareness, misrepresentation and function. Phil Trans R Soc B (2012); 367: 1424-1438.

Smith JD, Couchman JJ, Beran MJ. The highs and lows of theoretical interpretation in animal-metacognition research. Phil Trans R Soc B (2012); 367: 1297-1309.

Sutherland S. Consciousness. In *MacMillan Dictionary of Psychology*. Second edition. London: The MacMillan Press, 1995, 95.

Tallis R. *I Am: A Philosophical Inquiry Into First-Person Being,* Edinburgh University Press, Edinburgh, 2004.

Tallis R. *The Knowing Animal: A Philosophical Inquiry into Knowledge and Truth,* Edinburgh University Press, Edinburgh, 2005.

Timmermans B, Schilbach L, Pasquali A, Cleeremans A. Higher order thoughts in action: consciousness as an unconscious re-description process. Phil Trans R Soc B (2012); 367: 1412-1423.