themselves intact. The answer is that there are no such channels. Rather, the attended outputs of perception are globally broadcast to all conceptual systems, including the metarepresentational faculty inter alia. See section 2 for some discussion and references.

2. All of these authors endorse broadly "theory-theory" accounts of mindreading. A very different kind of "mindreading is prior" account is defended by Gordon (1986; 1996), who develops a form of simulation theory that denies any need for introspection. But this account makes both mindreading and metacognition dependent upon the acquisition of natural language. Likewise, Dennett (1991) is a sort of theory-theorist who denies introspection for attitudes, but he, too, appears to make our knowledge of our own mental states dependent upon their expression in language. Discussion of these issues would take us too far afield. For present purposes I assume, as seems plausible, that basic capacities for both mindreading and metacognition are independent of our capacity for natural language.

3. Note that for this reason Nichols and Stich's (2003) introduction of a separate perception-monitoring mechanism is wholly unnecessary. Since the mindreading system would need to have access to the agent's own perceptual states in order to do its work, there is simply no need for a distinct system to monitor and self-attribute those states.

4. In allowing that perceptual *judgments* are introspectable, I don't mean to imply that perceptually based *beliefs* are likewise introspectable. On the contrary, once formed and stored, the only way that those beliefs can be consciously accessed is via their expression in visual imagery (in the form of an episodic memory, perhaps) or in inner speech. But such events, although introspectable, will need to be interpreted to extract the information that they are, indeed, expressive of belief (as opposed, for example, to supposition or mere idle fantasy). See section 2.1 for further discussion.

5. An alternative account to the one sketched here is outlined by Wilson (2002), who suggests that the introspective assumption may make it easier for subjects to engage in various kinds of adaptive self-deception, helping them build and maintain a positive self-image. In fact, *both* accounts might be true.

6. We also know that in other domains – such as physics – the unconscious theories that guide behavior often make false, but simplifying, assumptions. See, for example, McCloskey (1983).

7. This isn't quite accurate. For, to the extent that apes, for example, do have limited mindreading abilities (e.g., in respect of perception and goal-directed action), to that extent one might expect to find metacognitive processes also. At any rate, this is what a "mindreading is prior" account would predict.

8. *Sometimes* a System 2 utterance *does* express an underlying System 1 judgment with the same content, no doubt. But in such a case it is all the clearer that the utterance in question isn't *itself* a judgment. Nor does the expressibility of judgments in speech provide any reason for believing in introspection, as we saw in section 2.1.

9. Similar claims are made by Bayne and Pacherie (2007). They argue against an interpretative account of self-awareness of the sort defended here, preferring what they call a "comparator-based" account. But I think they mis-characterize the models of normal action-monitoring that they discuss. Properly understood, those models lend no support for the claim that metacognition is damaged in schizophrenia. See the paragraphs that follow.

10. The claim that we have introspective access to our own motor intentions seems also to underlie the idea that "mirror neurons" might play an important role in the development of mindreading (Gallese & Goldman 1998). For what would be the use, for purposes of social understanding, of an activation of one's own motor system in response to an observation of the action of another, unless one could acquire metacognitive access to the motor plan in question? (For a variety of criticisms of this account of the mirror neuron system, see Csibra [2007] and Southgate et al. [2008].)

11. Russell and Hill (2001), however, were unable to replicate these results. This is probably because their population of autistic children, although of lower average age, had higher average verbal IQs, suggesting that their autism was much less severe. Since most researchers think that intention-reading is among the easiest of mindreading tasks, one might predict that only very young or more severely disabled individuals with autism would be likely to fail at it.

Open Peer Commentary

What puts the "meta" in metacognition?

doi:10.1017/S0140525X09000557

Michael L. Anderson[a,b] and Don Perlis[b,c]

[a]*Department of Psychology, Franklin & Marshall College, Lancaster, PA 17604;* [b]*Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742;* [c]*Department of Computer Science, University of Maryland, College Park, MD 20742.*

michael.anderson@fandm.edu **http://www.agcognition.org**
perlis@cs.umd.edu **http://www.activelogic.org**

Abstract: This commentary suggests an alternate definition for metacognition, as well as an alternate basis for the "aboutness" relation in representation. These together open the way for an understanding of mindreading that is significantly different from the one advocated by Carruthers.

Carruthers suggests that cognitive scientists are confused about the meaning of "metacognition," citing our work as an illustrative example. In fact, we follow a standard definition of the term, adopted from Nelson and Narens (1990). (This particular formulation appears in Anderson & Oates [2007], but the definition is in widespread use. See, e.g., Dunlosky 2004; Dunlosky & Bjork 2008; Dunlosky & Metcalfe 2009; Metcalfe 1993; Metcalfe & Shimamura 1994.) The definition runs as follows:

> Imagine two components *X* and *Y* (where *X* and *Y* could be the same), related in such a way that state information flows from *Y* to *X*, and control information flows from *X* to *Y*. Component *X* is in a monitoring and control relationship with *Y*, and when *Y* is a cognitive component, we call this relationship metacognitive monitoring and control.

This offers an information-theoretic characterization of metacognition that is neutral regarding the form that information takes, or the processing it undergoes. Thus, it is quite incorrect to say that cognitive scientists use the term "in two quite distinct ways, often without noticing the difference" (target article, sect. 5.1, para. 2). We use the term consistently in a way that leaves open the various ways in which such a relationship could be implemented. We are not confused about the difference between systems that involve "metarepresentations of [its] own first-order cognitive processes as such" (sect. 5.1, para. 2) and those that don't; rather, this distinction is not relevant to the definition of metacognition.

In fact, *some* of the processes in the systems we implement are indeed metacognitive in Carruthers' more restricted sense. To take just one example, mentioned by Carruthers: If an active logic system notices the presence of both P and $\neg P$ in its knowledge base (KB), it will assert $Contra(P, \neg P, t)$.

That is a statement *about* – a metarepresentation of – the state of the KB at time *t* (i.e., that it contained that contradiction). Our systems can reason about this fact with that metarepresentation, and consequently take various control steps, the simplest of which is to refrain from using these premises in further deduction (Anderson & Perlis 2005a). But other processes in active logic systems, and other of our metacognitive systems, effect such monitoring and control without explicit metarepresentations of this sort (see, e.g., Anderson et al. 2006).

Of course, Carruthers is free to define his terms and circumscribe his interests as best serves his argument, and if this were merely a terminological dispute, we would not be submitting a commentary. But there is a more substantive point in the background, which potentially affects Carruthers' overall proposal. Carruthers writes: "Generally the term is used, as it has been throughout this article, to mean cognition *about* one's own cognition. Metacognition, in this sense, is inherently higher-order, involving metarepresentations of one's own first-order cognitive processes as such" (sect. 5.1, para. 2, emphasis in original). The implication seems to be that for something to be *about* another requires a higher-order metarepresentation. But we would like to suggest that this associates *higher-order*-ness with *meta*-ness and *about*ness (if we can be forgiven the neologisms) in a way that is not necessary.

First, it is not clear that aboutness requires higher-order-ness. Surely a representation or a process can be about another without being at a different level, or in a different representational language. Indeed, can't a process (or representation) be about itself? (See, e.g., Perlis 1985; 1988; 1997; 2000; Perlis & Subrahmanian 1994.) It is a common bias, perhaps stemming from Tarski, that there must be a hierarchy of meta-languages, each standing back from the one it refers to. But Tarski adopted that approach to avoid technical difficulties in formal logic; it is not necessary a priori.

Second, it is not clear that meta-ness requires higher-order-ness. In related writings, we have suggested that representation requires only the following: tokens, whatever their form/content, that can be used to guide actions with respect to certain targets (Anderson & Perlis 2005b; Anderson & Rosenberg 2008). On these accounts, the information being used and manipulated during cognition is representational just in case it is used to guide behavior with respect to targets in various circumstances. Likewise, a metacognitive monitoring and control process represents a cognitive process, just in case it allows the metacognitive component to guide actions with respect to the cognitive process. Such monitoring and control is indeed (we maintain) cognition *about* cognition – is thus *meta*cognition – without having to be/utilize higher-order representations of cognition as such.

As should be clear from the preceding, we have a somewhat different understanding of what the representational aboutness relation requires. This most definitely applies to self-representation as well (Anderson & Perlis 2005b), although it is perhaps worth noting that the account of self-awareness we develop in the cited paper is – despite differences in the fundamental criteria for aboutness – nevertheless compatible with the "mindreading is prior" framework that Carruthers advocates.

So why might all of this matter to Carruthers? Because of Carruthers' understanding of what aboutness requires, he is driven to adopt a higher-order, meta-representational account of what having certain thoughts about another's thoughts ("mindreading") requires. In contrast, the less restrictive option offered by us opens the door for a broader range of theories of what our responsiveness to the mental states of others requires. This would include, for instance, Shaun Gallagher's interesting, and interestingly different, interaction-based account of understanding self and others (Gallagher 2004; 2005). It would have been useful and instructive to see how this rather broader portrayal of the competing possibilities might have affected Carruthers' argument, discussion, and conclusions.

Is feeling pain just mindreading? Our mind-brain constructs realistic knowledge of ourselves

doi:10.1017/S0140525X09000569

Bernard J. Baars

The Neurosciences Institute, San Diego, CA 92121.

baarsbj@gmail.com

http://bernardbaars.pbwiki.com

Abstract: Carruthers claims that "our knowledge of our own attitudes results from turning our mindreading capacities upon ourselves" (target article, Abstract). This may be true in many cases. But like other constructivist claims, it fails to explain occasions when constructed knowledge is *accurate*, like a well-supported scientific theory. People can know their surrounding world and to some extent themselves. Accurate self-knowledge is firmly established for both somatosensory and social pain.

Brain imaging studies show that social pain (like social rejection, embarrassment, and guilt) activates brain regions characteristic of painful *bodily* experiences. The brain regions that are activated by both evoked social and physical pain include the anterior cingulate cortex, the right prefrontal lobe, the insula, amygdala, and somatosensory cortex. Even deep brain structures, such as the brainstem periaqueductal gray (PAG), are known to be evoked by mother–infant separation, marked by intense and repeated distress cries. These functions are highly conserved among mammals and, perhaps, birds (Eisenberger & Lieberman 2004; Nelson & Panksepp 1998).

This evidence contradicts Carruthers' hypothesis that we learn about ourselves by turning our social mindreading capacities upon ourselves. No doubt we do learn about ourselves based upon what we have learned about others. After all, we constantly transfer knowledge between different domains of reference. However, it is simply not the case that *all* of our introspective self-knowledge is of this kind. Children acquire "theory of mind" abilities in about the fourth year of life. But long before that time we can observe, pain and pleasure perception, the distress of abandonment, anticipatory fear and joy, and a wide spectrum of social and imaginary emotional experiences.

Carruthers could maintain that such emotional experiences are not true cases of "metacognition" and "introspection." It is possible to define such terms in very limited ways, but there is no doubt that emotional feelings express propositional attitudes: They are *about* something, namely the well-being of the self. Thus, hunger, thirst, air-hunger, social distress, fear of rejection by the mother, peer envy, and numerous other infant emotions are by no means simple "reflexes." They are socially contingent, though not explicitly deliberated, reactions to real-world events that are critical to the infant's survival. This crucial self-related information has extraordinary breadth of conservation among mammals, suggesting an evolutionary history of some 200 million years (Baars 2005).

Pain is not the only kind of introspective experience humans have with minimal social input, but it is perhaps the most compelling. Metacognitive self-report ("introspection") has been used for two centuries in psychophysics. It is a well-established methodology that converges extremely well with other empirical evidence, such as brain recording methods (Baars & Gage 2007).

Science is a constructive enterprise, but it is tightly constrained by evidence. That is why, like other human activities such as farming and tax accounting, it is not merely constructed, but also bound by considerations of accuracy and predictability.