We doubt that Carruthers has possibility 1 in mind, as this would mean that one is confabulating even when one quite consciously uses interpretative processes to discern one's past mental states. If Carruthers has option 3 in mind, then we need to know much more about what distinguishes the proper subset. As a result, we proceed on the assumption that possibility 2 captures what Carruthers has in mind.

Our experience with identifying our own current mental states is characteristically quick, accurate, and confident. By contrast, when it comes to attributing mental states to others, our attributions seem much slower, more accident prone, and unsure. This subjective difference is thought to provide prima facie evidence that we have (non-interpretative) introspective access to our own mental states. Carruthers attempts to defeat this prima facie consideration by proclaiming that confabulated reports are subjectively indistinguishable from cases of alleged introspection. People confabulate attributions of their own propositional attitude events "while being under the impression that they are introspecting" (sect. 6, para. 1). Thus, we have no reason to think that canonical cases of "introspection" differ from confabulation in this respect (i.e., that we are interpreting in the latter case but not the former). Carruthers goes on to argue that since there is no other positive reason to believe in the reality of introspection for the attitudes, the best explanation is that all self-attribution (confabulation and alleged introspection) is subserved by the same kinds of processes: that is, interpretative ones.

Carruthers' argument depends on the claim that people confabulate attributions of propositional attitudes while being under the impression that they are introspecting. But we are given no evidence that this has been systematically investigated. Certainly no one has ever asked participants in these cases whether they think they are introspecting or interpreting. Without some more direct evidence, Carruthers is not warranted in claiming that when people confabulate they are often "under the impression that they are introspecting."

A closer look at the confabulation cases gives further reason to doubt the argument. The evidence on confabulation cited by Carruthers is all anecdotal, but even the anecdotes are illuminating if one looks at the behavior a bit more closely. For we find that across many different paradigms in which people confabulate, the confabulations are not reported with a sense of "obviousness and immediacy." Consider the following examples:

a. In a classic misattribution study, subjects took more shock because they thought a pill caused their symptoms. In a debriefing procedure subjects were asked, "I noticed you took more shock than average. Why do you suppose you did?" Nisbett and Wilson (1977) present one instance of confabulation and claim it as typical. The confabulation begins as follows: "Gee, I don't really know . . ." (p. 237).

b. In a dissonance reduction experiment involving shocks, Zimbardo reports that a typical confabulation would have been, "I guess maybe you turned the shock down" (Nisbett & Wilson 1977, p. 238).

c. Thalia Wheatley, one of the most inventive researchers using hypnotic suggestion (e.g., Wheatley & Haidt 2005), reports that when she has participants perform actions under hypnotic suggestion, she often asks them why they performed the action. Although they do often confabulate, their *initial* response to the question is typically "I don't know" (T. Wheatley, personal communication).

In each of these research paradigms, we find *typical* confabulations delivered with manifestly low confidence, rather than the sense of obviousness and immediacy that is supposed to be characteristic of introspective report.

Carruthers also draws on widely cited cases of confabulation involving split-brain patients. And, although Carruthers claims that split-brain patients confabulate with a sense of obviousness and immediacy, the situation is not so clear. In footage of split-brain patients, we find them showing little confidence when asked to explain behavior issuing from the right hemisphere. For instance, in a typical study with split-brain patient Joe, Joe is shown a saw to his right hemisphere and a hammer to his left. He is then told to draw what he saw with his left hand. Predictably, Joe draws a saw. Gazzaniga points to the drawing and says, "That's nice, what's that?" *Saw*. "What'd you see?" *I saw a hammer.* "What'd you draw that for?" *I dunno* (Hutton & Sameth 1988).

Carefully controlled studies are clearly needed. However, these anecdotes provide prima facie reason to think there are systematic differences in confidence levels between confabulation and apparent introspection, which in turn suggests a difference in underlying mechanism. The fact that confabulations are accompanied by low confidence does not, of course, provide conclusive evidence in favor of introspection. But it does suggest that given the present state of the evidence, the confabulation argument is toothless.

How we know our conscious minds: Introspective access to conscious thoughts

doi:10.1017/S0140525X09000636

Keith Frankish

Department of Philosophy, The Open University, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom.
k.frankish@open.ac.uk
http://www.open.ac.uk/Arts/philos/frankish.htm

Abstract: Carruthers considers and rejects a mixed position according to which we have interpretative access to unconscious thoughts, but introspective access to conscious ones. I argue that this is too hasty. Given a two-level view of the mind, we can, and should, accept the mixed position, and we can do so without positing additional introspective mechanisms beyond those Carruthers already recognizes.

In section 7 of the target article, Carruthers considers the proposal that we have two levels of mentality, conscious and unconscious, corresponding to the two reasoning systems posited by many psychologists, and that we have different forms of access to the attitudes at the two levels – merely interpretative access to those at the unconscious level, but introspective access to those at the conscious level. Prima facie, this mixed position is an attractive one, which does justice both to the evidence for psychological self-interpretation cited by Carruthers and to the everyday intuition that we can introspect our conscious thoughts. Carruthers rejects the option, however. Although conceding that we have introspective access to conscious *thinking*, he denies that we have such access to conscious *judgments* and *decisions*. I argue here that this conclusion is too hasty.

Carruthers' argument turns on the claim that judgments and decisions *terminate* reasoning processes and produce their characteristic effects directly, without further processing. Conscious thinking, on the other hand, involves rehearsing mental imagery, especially inner speech, and this has only an indirect influence on thought and action. The route may be metacognitive: A rehearsed assertion with content p may give rise to an (unconscious) metacognitive belief, to the effect that one believes that p or that one is committed to the truth of p, which, together with suitable desires, will lead one to think and act as if one believes that p. Or the rehearsed assertion may be processed as testimony, leading one to form the first-order belief that p, which will then guide behaviour in the normal way. On either route, Carruthers argues, the conscious event gives rise to the effects of a judgment only through the mediation of further cognitive processing, and so does not count as a judgment itself. Similar considerations apply to decisions, although here Carruthers mentions only the metacognitive route.

I am sympathetic to Carruthers' account of conscious thinking, and I agree that imagistic rehearsals influence thought and action through the mediation of unconscious cognitive processes. But this is not incompatible with the commonsense view that some conscious events are judgments and decisions. To see this, we need to take seriously the suggestion that the conscious mind is a distinct *level* of mentality. Carruthers has himself developed a version of this view, arguing that the conscious mind (the psychologists' System 2) is not a separate neural structure, but rather, a higher-level "virtual" one, realized in cycles of operation of a more basic unconscious system (System 1), which, among many other tasks, generates and processes the imagery involved in conscious thinking (Carruthers 2006; 2009; for a related version, see Frankish 1998; 2004; 2009). And from this perspective it is natural to regard appropriate utterances in inner speech as genuine judgments and decisions – at least when they achieve their effects via the metacognitive route. For these events will terminate reasoning processes *at the higher level* and *on the relevant topic*. The further processing occurs at the lower level and is devoted to a different topic. When I rehearse the sentence, "Polar bears are endangered" in assertoric mode, this terminates my reasoning about polar bears. The subsequent unconscious reasoning is about how to interpret and respond to this assertion, not about whether the conclusion it expresses is correct. These processes can be thought of as *implementing* the higher-level attitude, and their existence does not compromise the status of the conscious event as a judgment.

It is true that the lower-level processes may sometimes fail to generate the appropriate effects (for example, if the desire to execute one's commitments is overridden by a stronger desire), but this is irrelevant. On every view there are some implementing processes, at least at a neurological level, and these processes may go awry. And if we have a settled habit of interpreting appropriate utterance rehearsals as expressions of belief or commitment, and a settled desire to act consistently or to discharge our commitments, then the right effects will follow most of the time. Similar considerations apply to decisions.

The only peculiarity of the two-level view is that the processes that implement conscious judgments and decisions are cognitive ones. But why should that matter? Compare the way the judgments and decisions of a company are implemented. The edicts emerging from the boardroom require further processing in order to affect the activities of the organization, and this processing involves reasoning on the part of the staff involved. (Again, this will have a metarepresentational character, involving beliefs about what the directors have concluded.) But we still want to say that the judgments and decisions were made in the boardroom, rather than in the cubicles of the junior staff.

What about cases in which a rehearsed utterance generates its effects via the second route, being processed as testimony and generating a first-order belief? Here I think Carruthers is right. If further processing serves to evaluate the conclusion reached rather than simply to implement it, then this does disqualify the conscious event from judgment status. But note that in such cases, the agents themselves will not think of the conscious events as judgments. For if they did, they would naturally come to believe that they believed, or were committed to, the conclusions expressed, and the subsequent processing would follow the metacognitive route. Thus, there is no reason to regard such events as candidates for judgments in the first place. (We might think of them as hypotheses or self-suggestions.) Again, the same goes for decisions.

I conclude that Carruthers' case against a mixed position is not compelling. It is important to stress that the proposed mixed position does not involve positing additional introspective mechanisms. Carruthers allows that we have introspective access to conscious (System 2) thinking; I am simply claiming that some of the introspectable events can be legitimately classified as judgments and decisions. The proposal is merely a reconceptualization of the processes Carruthers describes. But it is a natural one, given a two-level view of the sort Carruthers endorses, and one that accords with intuition. For these reasons it should be preferred. Of course, it would be ad hoc if a two-level view were not independently motivated, but it is (see aforementioned citations).

Non-interpretative metacognition for true beliefs

doi:10.1017/S0140525X09000648

Ori Friedman and Adam R. Petrashek

Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.
friedman@uwaterloo.ca
http://www.psychology.uwaterloo.ca/people/faculty/friedman/
arpetras@uwaterloo.ca

Abstract: Mindreading often requires access to beliefs, so the mindreading system should be able to self-attribute beliefs, even without self-interpretation. This proposal is consistent with Carruthers' claim that mindreading and metacognition depend on the same cognitive system and the same information as one another; and it may be more consistent with this claim than is Carruthers' account of metacognition.

Mindreading often requires access to one's own beliefs.[1] Consider the following mental state attributions: Bill *believes* a first-aid kit contains bandages, though the kit actually contains feathers; Louise is an expert in British history, so she *knows* that the Battle of Hastings occurred in 1066; and Sally, age 2, *desires* candy when offered a choice between this and sushi as a snack. These mental state attributions do not depend on the interpretation of others' speech or behavior. Instead, they primarily depend on your beliefs (i.e., first-aid kits normally contain bandages; the Battle of Hastings occurred in 1066; children typically prefer candy over unfamiliar foods) in combination with other principles (e.g., experts in British history know a lot about British history).

The need to access beliefs is not restricted to just a few cases of mindreading. Instead, such access may be the rule in belief attribution: Most beliefs are true, and so one's own beliefs are indicative of what others believe. Because of this, people may have a default tendency to attribute their "true" beliefs to others (Fodor 1992; Leslie & Thaiss 1992; see Leslie et al. [2004] for a review of much evidence favoring an account making this claim). To operate according to this default tendency, the mindreading system requires access to beliefs.

The mindreading system's access to beliefs is problematic for Carruthers' account of metacognition, which denies such access (target article, sect. 2, para. 6).[2] For if the system accesses beliefs when attributing mental states to others, then it should also access them when attributing mental states to the self. For instance, if the mindreading system accesses the belief "the Battle of Hastings occurred in 1066" when attributing it to Louise the historian, then the system should also be able to attribute this belief to the self. The mindreading system's access to beliefs allows people to engage in non-interpretative metacognition.

This proposal does not necessarily imply non-interpretative access to other mental states, such as intentions, desires, and past (currently false) beliefs. Unlike currently held beliefs, these other mental states are typically uninformative about the world and about others' mental states. One's intention to drink coffee says little about the world except perhaps that people sometimes drink coffee; and it says little about other people because relatively few share this intention at any time, meaning that it will seldom be useful to quickly extend this intention to