(Forthcoming in *Behavioral and Brain Sciences*, commentary on Peter Carruthers' "How we know our own minds")

Confabulation, confidence, and introspection

Brian Fiala
Shaun Nichols

University of Arizona
Department of Philosophy
Social Science Bldg. Rm 213
PO Box 210027
Tucson, Arizona 85721-0027
U.S.A.

Abstract:

Carruthers' arguments depend on a tenuous interpretation of cases from the confabulation literature. Specifically, Carruthers maintains that cases of confabulation are "subjectively indistinguishable" from cases of alleged introspection. However, in typical cases of confabulation, the self-attributions are characterized by low confidence, in contrast to cases of alleged introspection.

What is confabulation? Carruthers' central argument hinges on this notion, so we need to get clear on what he has in mind. Carruthers doesn't present an explicit characterization, but the overall discussion suggests that the relevant confabulations are a class of first-person mental state attributions that are generated by an "interpretative" process, as opposed to an "introspective" process. By "interpretative," Carruthers means "any process that accesses information about the subject's current circumstances, or the subject's current or recent behavior, as well as any other information about the subject's current or recent mental life" (5). This characterization seems too broad since introspection itself is supposed to be a process that accesses information about the subject's current mental life. But Carruthers means to count as interpretative only those processes that do not employ any 'direct' access or any mechanism specifically dedicated to detecting one's current mental states.

On Carruthers' view, all attributions of propositional attitude events are, in fact, interpretative. So what is the relation between "confabulation" and "interpretation"? Here are several different possibilities:

(1) Confabulations include all self-attributions that result from interpretation.
(2) Confabulations include all *false* self-attributions that result from interpretation, and accurate interpretative self-ascriptions do not count as confabulatory.
(3) Confabulations include only a proper subset of false self-attributions resulting

from interpretation.

We doubt that Carruthers has (1) in mind since 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 (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 (noninterpretative) 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" (20). 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, i.e., interpretative.

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:

- 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 & Wilson present one instance of confabulation and claim it as typical. The confabulation begins as follows: "Gee, I don't really know…" (237).
- 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 238).
- 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" (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 while 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 that 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.

References:

Hutton, R. (Editor) & Sameth, J. (Director). (1988.) *The mind: Part 1: The search for the mind.* [VHS Tape]. Santa Barbara, CA : Annenberg/CPB Project.

Nisbett, R. & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.

Wheatley, T. & Haidt, J. (2005) Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, 16 780-784.