Forthcoming in *Philosophy and Phenomenological Research*

Contribution to Book Symposium on Joëlle Proust's *The Philosophy of Metacognition*

Penultimate Version

Unwitting Self-Awareness?

Peter Langland-Hassan
University of Cincinnati

I.

A central question in the philosophy of metacognition is whether the standard tests of metacognition are, in fact, tests of metacognition. For an animal to answer questions only when it knows the answer, and to "opt out" of answering otherwise, does it need to *metacognize*? It depends, obviously, on what you think metacognizing requires. If, like Josef Perner (2012) and Peter Carruthers (2011), you hold that metacognizing requires the use of mental state concepts in thoughts about your own mind, then—for reasons I will discuss—the answer may be no. But if, like Joëlle Proust (2013), you think that metacognition can occur in a non-propositional format, without the use of mental state concepts, and without one's representing one's own mental states, then the answer is not so clear. But neither, in that case, is it very clear what one *means* by "metacognition." Or so one might object.

In response, Proust points to metacognitive *know-how*—what she terms 'procedural metacognition.' On her view, humans, and a handful of other species, are able to monitor and evaluate their own cognitive states and dispositions, and to fruitfully use such self-evaluations to guide their behavior, without knowing *that* they have minds at all. This monitoring and evaluation (typically of confidence levels) is accomplished through one's sensitivity to nonconceptual, nonpropositional representations she calls "noetic feelings." While noetic feelings serve in the monitoring and control of one's own cognitive states, Proust holds that they are not *metarepresentational*; they do not require the organism to form representations of its own (first order) cognitive states. She contrasts procedural metacognition to "analytic" metacognition, which involves the use of mental state concepts in metarepresentational propositional thoughts about one's own mental states. (I will follow her in the using the terms 'procedural' and 'analytic' to mark this distinction.) Analytic and procedural metacognition are important yet distinct capacities, according to Proust. Her main aim is to defend the importance of the procedural variety against a tradition in philosophy and (parts of) psychology that will have doubts about its legitimacy (see, e.g., Carruthers & Ritchie (2012)).

As Proustian procedural metacognition does not bear any obvious relation to *thinking about thinking* as philosophers typically understand it, she needs to show that the processes at its heart have special import for understanding an organism's awareness of its own mind. As Proust recognizes, it is not enough to observe that animals succeed at the standard (nonverbal) tests of metacognition. First,

many philosophers will deny that success at those tasks requires metacognition in the first place. And, second, it may seem that the kind of cognitive control processes that Proust thinks explain the animals' behavior are rife in cognition generally, and have nothing in particular to do with an organism's awareness or understanding of mentality.

While Proust is alive to both worries, and addresses each in some depth, I found her replies lacking in important respects. Yet, otherwise, I am quite sympathetic with her approach, and found much to be excited by in her very substantial new contribution to the metacognition literature. My plan in this commentary, then, is to say why I think she does not adequately dispel the two worries, and to suggest another way forward—one that preserves the main thrust of her position. As my proposed "fix" is at odds with at least one important strain in her work—the idea that procedural and self-directed analytic metacognition are fundamentally distinct capacities—she will no doubt wish to take this friendly advice with a grain of salt.

II.

The standard metacognition studies have a common structure: the participant (animal or human) is presented with tasks of varying difficulty, with known rewards for answering correctly, and penalties for answering incorrectly. Often it is a perceptual discrimination task, where the participant must, for instance, select the longest of nine lines (Kornell, Son, & Terrace, 2007), or indicate whether an array of dots is qualifies as "dense" or "sparse," based on a previously established threshold (Smith, Beran, Couchman, & Coutinho, 2008). In the "opt-out" variant of such tasks, the participant is given the choice to opt out of answering the prompt. Opting out results in a lesser reward than answering the prompt correctly, but is preferable to the penalty (typically a time delay) received for answering incorrectly. Interestingly, some species (rhesus macaques, dolphins) have been shown to use the opt-out key adaptively, while others (capuchin monkeys, pigeons) are unable or unwilling. Proust aims to show how the adaptive use of the opt-out key is genuinely metacognitive, yet available to creatures who otherwise seem to lack concepts of mental states (as evidenced by their inability to pass "false belief" tasks).

There are at least two ways one can question the claim that adaptive performance in the metacognition tasks requires metacognition. The first is well known, and has been controlled for in recent experiments. This is the objection that animals may be conditioned to use the opt-out key not based on their own uncertainty, but on superficial features of the stimuli. So, for instance, on a task where touching the longest of nine lines results in the preferred reward, the animal could follow the rule: touch the longest line when it is much longer than the others, and touch the other key (which experimenters conceive of as the "opt out" key) whenever the two longest lines are very close to the same length. This would result in behavior that appeared "metacognitive." To rule out such alternatives, researchers (e.g., Kornell *et al.*, (2007)) now include a second phase, which incorporates novel types of stimuli that have not yet been associated with the opt-out key. The animal's ability to

immediately transfer use the opt-out key to the new stimuli, in adaptive way, shows that it was not simply associating it with superficial features of the prior stimulus type.

There is, however, a more subtle deflationary hypothesis that still poses a *prima facie* threat to viewing such tasks as measures of metacognition. As Perner (2012) notes, in lieu of interpreting use of the opt-out key as indicating the subject's awareness of its own uncertainty, one could equally well interpret it as an indication that subject has judged the trial to be *difficult*. Thinking that a trial is difficult does not obviously require any self-awareness, and would suffice for explaining the animal's ability to immediately transfer use of the opt-out key to new stimuli. True, judgments of difficulty depend on, and result from, subjective states and abilities of the animal. So there is *a sense in which* reports of difficulty are subjective. However, this is true of many ordinary reports we would never count as metacognitive, such "This room is hot" or "That person is attractive" (Perner, 2012, p. 99-101). I will call this "Perner's Challenge."

Proust does not, that I can see, offer a response to Perner's Challenge. She emphasizes that, in order to circumvent objections that have historically been raised about the metacognition studies, we should follow Hampton's (2009) criteria for appropriate metacognition task structure. Among those criteria is the stipulation that a subject's responses "must not be based on environmental cue association" (Proust, 2013, p. 83). Is the difficulty-level of a task an "environmental cue"? If it is, then existing studies do not satisfy Hampton's criteria. If it is not, then we need to understand *why* it is not, when the hotness of a room, or the attractiveness of a face, presumably *would* constitute "environmental cues."

In considering the related objection that procedural metacognition "boils down to primary task-monitoring," Proust emphasizes that "the information needed to *make a decision under uncertainty* is not the same as the information used in *assessing one's uncertainty*" (p. 104). According to Proust's "double accumulator model," two cognitive mechanisms called 'adaptive accumulator modules' (AAMs) underlie metacognition, and have processing elements relating to each of the two tasks: first-order choices between A and B are determined by the comparative rate at which evidence for either A or B is gathered in one accumulator (there being a specific threshold where one choice is made over the other); and a second accumulator has the function of adjusting the evidential or "confidence" threshold at which decisions are made, based on prior successes and failures when acting at that threshold (pp. 99-102). When put in this way, it is natural to conceive of the latter accumulator as metacognitive. However, we could alternatively describe the accumulators in an "outward looking" fashion. One can distinguish between *making a difficult decision* (i.e. where there is evidence for competing hypotheses), and *assessing the difficulty-level of the question* (where this amounts to assessing whether the level of difficulty is one where answering leads to reward). And one could conceive of Proust's second accumulator as aimed at assessing and calibrating the difficulty level at which answers should be given, rather than the confidence level. From this angle, the cognition in question appears entirely aimed at the first order task and its properties.

One strategy for tipping the balance in favor of a metacognitive reading is to argue that the balance of empirical evidence, combined with a teleosemantic approach to content ascription, warrants

the view that one cognitive state or process represents the content of another—even if the higher-order representation is *nonconceptual* in nature (see, e.g., Shea (2014)). For one might think that mentally representing the content of one's own mental state—via the use of mental state concepts or not—is the essence of metacognition. Yet this is not an approach Proust favors. She denies that procedural metacognition involves *meta*representation, nonconceptual or otherwise. According to Proust, noetic feelings "express dynamic properties in the cognitive vehicle" (p. 157). Her view is that, while noetic feelings represent states of a person's cognitive system (or "cognitive affordances"), they do not qualify as metarepresentations because they do not represent the *content* of those states; rather, they represent (neural) properties of the "cognitive vehicle." Yet, given that her dual accumulator model is a *computational* model—where later stages in processing monitor and evaluate earlier stages—it should be multiply realizable. It would then be an error to hold that later stages (i.e., the secondary accumulator) represent *neural* properties, instead of the content of the states they evaluate; for the neural realization of the model is presumably a contingent matter, while the computational values at which different evaluations are made are not. Be that as it may, what matters for present purposes is that, if noetic feelings are not metarepresentations, the rationale for calling them metacognitive will have to come from elsewhere.

That said, it may be just as well that Proust avoids the position that noetic feelings are nonconceptual metarepresentations, and metacognitive *for that reason*. For if nonconceptually representing the content of one's own mental state is assumed sufficient for metacognition, this will have the result that metacognition occurs wherever a cognitive process has the function of controlling, monitoring, or calibrating another cognitive process. Many agree that the most basic mechanisms governing action and perception involve subconscious prediction and comparison processes that fulfill these criteria (Wolpert, Miall, & Kawato, 1998). Suddenly fly-swatting and jump-roping become metacognitive events. At that point we seem to have lost sight of our original goal, which was to understand an organism's awareness and knowledge of its own mind.[1]

III.

Thus, the two worries noted at the outset persist: there remain ways of viewing performance on the metacognition tasks as hinging on the detection external environmental cues (i.e., task difficulty); and, in addition, there is a threat of overgeneralization if we hold that the tasks are metacognitive just because they require one to form (nonconceptual) representations of one's one own cognitive states. With these difficulties in mind, I want to sketch a different possibility for vindicating the metacognition experiments—one that I think coheres with much in Proust's account.

"First order" or not, the metacognition tasks can considered *highly relevant* to understanding metacognition if they tap an ability that is an essential component of "full blown," concept-involving, analytic, propositional, self-directed metacognition. Unlike judgments of hotness, or of attractiveness, judgments of task difficulty are intimately related to the conception we have of ourselves as fallible

[1] I say "seem to have" because I think this is, in fact, a far more delicate question than I can adequately address here.

cognitive agents, whose representations of reality may be inaccurate or incomplete. This can best be appreciated from the perspective of "outward looking" theories of introspection and self-knowledge.

Following Evans (1982), a number of philosophers have proposed that knowledge of one's own mind can be generated by "looking outward": when trying to decide if one believes that *p* (a metacognitive question), it is enough simply to consider whether p, where the question whether *p* is a "first order," world-directed question. In cases where one finds *p* to be the case*,* it is safe to infer that one believes that *p*. I will call this the "ascent-routine" approach, even if not all versions are equivalent. Alex Byrne defends a sophisticated version of this sort of view in a series of recent papers. As he emphasizes, one does not need to know that one has inferred that *p* in order to follow such a procedure; rather, one needs to learn to follow a rule that, from one's own perspective, can be understood as: "If *p,* believe that I believe that *p*" (Byrne, 2005). Bryne proposes similar "outward looking" procedures for desire (2012), thinking (2011), and a number of other mental states (see also Gordon (2007)). A virtue of the general approach is that it promises to explain the means by which we easily and securely generate true beliefs about our own current mental states, without having to posit cognitive mechanisms over and above the traditional (world-directed) senses, and a general ability to follow rules of inference. A sizeable gap in this literature, however, concerns the crucial ability we have of knowing when we do not know, one way or the other, whether *p*. From what fact about the external world can one infer that one has no opinion either way concerning, for example, whether it is raining in San Francisco?

It is here that the animal metacognition literature may hold some lessons: the inferential rule by which we can pronounce ourselves uncertain may proceed by way of determining that a particular question or present problem is *difficult*. That is, one can move, inferentially, from the degree of difficulty a question poses to one's own degree of confidence in the answer (so long as difficulty is being assessed via one's own engagement with the question, and not with respect to some more objective standard). In this way, the defender of the metacognition experiments can accept—even insist—that the tasks themselves are "first order," requiring only sensitivity to the trial's level of difficulty. However, they remain very *relevant to* understanding metacognition because the "environmental cue" they require one to discern—namely, task difficulty—can be featured in the antecedent of an epistemic rule one can learn to follow that generates propositional knowledge of one's own lack of knowledge (the rule being: if the question is very difficult, believe that you do not know the answer). Offering an account of the cognitive mechanisms that underlie our ability to find a question (subjectively) difficult—as Proust does in her appeal to the double accumulator model, and to "cognitive affordances"—can then be seen as explaining the mechanics of how the "outward looking" phase in an eventual metacognitive ascent routine (available to those in possession of the concept UNCERTAIN) is accomplished.

The key to maintaining the metacognitive relevance of the animal metacognition experiments, then, may lie in closing the gap between procedural and (self-directed) "analytic" metacognition, by revealing the former as an integral part of the latter. From this perspective, self-directed metacognition is seen as one capacity—involving both Proustian procedural elements, and analytic conceptual elements applied via "outward looking" ascent routines—while other-directed mindreading is an distinct ability, which enables one to understand others as having and acting from a different view of the world

than one's own. Unlike judgments of hotness or attractiveness, outward-looking judgments of task difficulty are a central component of the mature human ability to judge oneself to have an incomplete and possibly inaccurate representation of the world, and are relevant to metacognition *for that reason*.

This is not how Proust views the territory. She sees procedural metacognition as one (always self-directed) capacity, and analytic metacognition as an importantly distinct capacity used both with respect to oneself and others. "There is a phylogenetic difference between procedural and analytic metacognition," she tells us. "The first type relies on a variety of mechanisms for error detection and control; the second is a distinct adaptation, which enables agents to understand error as false belief" (p. 105). Proust is well aware of the ascent routine strategy, but dismisses it as offering a "very shallow, and indeed purely verbal" form of analytic self-knowledge; it has little to do with *genuine* self-directed analytic metacognition. Nor does she wish to rest the metacognitive relevance of procedural metacognition on its possible role in an ascent routine.

We can grant that there is a sense in which ascent routines are "shallow." The question is: does self-awareness get much deeper? What is it, really, that self-directed analytic metacognition makes possible, over and above the kinds of "procedural" capacities shown by animals in metacognition experiments? If it is a distinct *adaptation*, it must bring with it an ability to *do* something other than engage in procedural metacognition and supposedly shallow ascent routines. Proust does not offer many details here, other than that it enables "new executive abilities" such as the ability to "refrain from acting impulsively," and to "reject what does not cohere with [one's] values" (p. 52). But it is easy enough to conceive of those capacities in first-order, non-metarepresentational terms. Belief revision can be modulated by changing levels of evidence for competing commitments; and impulsive action may be avoided by developing appropriately strong competing desires.

It would help Proust defend the depth of the procedural/analytic distinction, and resist the kind of collapse I am suggesting, if she could better clarify the special capacities self-directed analytic metacognition makes possible. (Though, in resisting the collapse, I think she leaves herself open to the question of why procedural metacognition should count as metacognitive in the first place). One way to do so would be by appeal to a non-verbal task that specifically assesses one's capacity for (so-called) analytic metacognition. The task should be nonverbal in order to clarify what, if anything, self-directed analytic metacognition enables other than the (possibly "shallow") self-ascription of mental states via ascent-routines.

Interestingly, there do not currently seem to be any such experiments. There is no first-person equivalent of the classic (non-verbal) false belief task. Is this because enjoying the fruits of self-directed analytic metacognition requires language? That is not a view Proust defends. If it were true, we would like to know *why*. Or is it just that applying mental state concepts to ourselves does not do that much for us, over and above facilitating the kinds of capacities Proust identifies as merely procedural? Answering these questions will tell us how best to apply the many insights contained in Proust's rigorous and groundbreaking work.[2]

[2] Thanks to Christopher Gauker for helpful feedback on this commentary.

References

Byrne, A. (2005). Introspection. *Philosophical Topics, 33*(1), 79-104.

Byrne, A. (2011). Knowing that I am Thinking. In A. Hatzimoysis (Ed.), *Self-Knowledge* (pp. 105-124). Oxford: Oxford University Press.

Byrne, A. (2012). Knowing What I Want. In J. Liu & J. Perry (Eds.), *Consciousness and the Self* (pp. 165-183). Oxford: Oxford University Press.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Carruthers, P., & Ritchie, J. B. (2012). The emergence of metacognition: affect and uncertainty in animals. In M. J. Beran, J. Brandl, J. Perner & J. Proust (Eds.), *Foundations of Metacognition* (pp. 77-89). Oxford: Oxford University PRess.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Gordon, R. M. (2007). Ascent routines for propositional attitudes. *Synthese, 159*, 151-165.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comp Cogn Behav Rev, 4*, 17-28.

Kornell, N., Son, L., & Terrace, H. (2007). Transfer of metacognitive skills and hint-seeking in monkeys. *Psychological Science, 18*, 64-71.

Perner, J. (2012). MiniMeta: in search of minimal criteria for metacognition. In M. J. Beran, J. Brandl, J. Perner & J. Proust (Eds.), *Foundations of Metacognition* (pp. 94-116). Oxford: Oxford University Press.

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

Shea, N. (2014). Reward Prediction Error Signals are Meta-Representational. *Nous, 48*(2), 314-341.

Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review, 15*(4), 679-691.

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal Models in the cerebellum. *TRENDS in Cognitive Science, 2*, 338-347.