July 3, 2014
Jennifer Nagel, University of Toronto
Forthcoming in *Philosophy and Phenomenological Research*

The Meanings of Metacognition

[Contribution to a book symposium on Joëlle Proust's *The Philosophy of Metacognition*]

The battle over metacognition begins with the word itself: the prefix indicates that the word denotes something about or above ordinary cognition, but 'about' and 'above' are flexible enough that researchers have been able to produce strikingly divergent theories of the domain they label 'metacognitive'.

Joëlle Proust's rich and complex book begins by raising the question of whether two leading groups of metacognition researchers aim to offer rival accounts of the same thing. She argues that from the perspective of one camp—the 'inclusivists'—the answer is 'yes', and from the perspective of the other, her own 'exclusivist' camp, the answer is 'no'. Inclusivists take metacognition to refer to 'knowledge and cognition about cognitive phenomena' (2). Metacognition, understood as cognition *about* cognition, involves 'the theoretical knowledge *that* one knows, understands, remembers, perceives and so on' (3), a capacity underwritten by a much more general conceptual ability to attribute mental states such as knowing and perceiving, whether to oneself or others. Proust's exclusivists, by contrast, see metacognition as something poised *above* ordinary cognition: this group has focused on the feelings generated by thinking, feelings such as certainty and tip-of-the-tongue states. These feelings seem to have some function connected to monitoring one's own mental activity, as opposed to being equally applicable to ourselves and others, and

Proust argues that they function in a way that does not depend on any prior conceptual ability to attribute mental states. In her view, 'metacognition' should be understood as 'referring *exclusively* to the capacity of self-evaluating one's own thinking'; she holds that 'metacognition is a natural kind: it has a set of functional features of its own, which are independent of those associated with the self-attribution of mental states' (4). Metacognition in her sense sometimes applies to self-attributed mental states, but it need not always do so: she contends that some metacognitive evaluations are made without any conceptualization of a mental state on the part of the agent, including, for example, evaluations made by animals relatively incapable of mental state attribution. Proust takes metacognition to be 'based in part on non-analytic, that is, procedural, knowledge (knowing how, rather than knowing that)' (4).

Whether or not we stipulate that the word 'metacognition' be restricted to the self-evaluative capacity that Proust singles out here—it is after all to some extent arbitrary how broadly we use this technical term—we can agree that she has selected an interesting class of feelings as her target of inquiry. There is something intriguing about the function performed by such sentiments as the feeling of knowing (FOK), the sense that one will be able to answer a question just posed, even when one's retrieval of the answer is temporarily blocked. Furthermore, it does seem right to say that feelings in this class have a special first-person relevance. Although researchers have identified some roughly equivalent third-person phenomena, such as 'the feeling of another person's knowing' (Jameson, Nelson et al. 1993), the use of the word 'feeling' is a stretch here, with these researchers themselves characterizing these third-person phenomena as judgments, rather than 'feelings' strictly speaking. However, one might agree with Proust's suggestion that

metacognitive feelings have a special self-evaluative function while crossing over to the far inclusivist side about the importance of mental state attribution, even to the point of arguing that metacognitive feelings can only serve their special self-evaluative function when they are appropriately coupled with self-attributed mental states. In what follows, I'll argue that Proust is right to emphasize the procedural knowledge afforded by metacognitive feelings, but I'll resist her suggestion that this knowledge is 'non-analytic' in character, in the sense of not involving conceptualization of mental states. It seems to me that the special 'knowing how' that is enabled by metacognitive feelings always depends directly on 'knowing that'.

To isolate the type of self-evaluative function that Proust regards as distinctive, it is useful to begin by describing a function that happens to contribute to broadly similar ends without counting as metacognitive in anyone's books. Multisensory integration provides a straightforward example. In gaining information about the objects in our environment, we simultaneously gain information about the reliability of our senses. As the human nervous system draws information from a variety of sensory modalities in making judgments about the environment, it assigns different weights to these channels of information as conditions change (for example, shifting to rely more on touch and hearing as darkness falls and the signal from vision is increasingly blurred). Strikingly, this process of integration is almost perfectly optimal, in the sense that the weight assigned to each sensory modality is an almost perfect reflection of the modality's relative current precision (e.g. Ernst and Banks 2002). Researchers examining sensory integration have suggested that 'the central nervous system encodes an estimate of measurement error along with every estimate of position, or other attributes' (Alais and Burr 2004, 261). While one might think that we would need

some kind of dedicated self-monitoring capacity to 'read' that estimate of error and ensure that the signal from each modality is weighted appropriately, it now seems that the optimality of sensory integration is achieved more directly, as a function of the way in which neural populations inherently encode probability distributions. As visual noise increases, for example, a wider spread of relevant neurons will fire in response to a visual stimulus, and the increased variance in the neuronal response from vision automatically dilutes the impact of this sensory channel's informational contribution to the combined output (Ma, Beck et al. 2006). Although neural activation patterns showing greater variance are sometimes said to represent greater 'uncertainty' about a stimulus, feelings of certainty or uncertainty play no functional role in multisensory integration.

Multisensory integration is not metacognitive: it automatically incorporates information about the perceiver as a byproduct of the way in which the perceiver responds to outer objects, and it requires no distinct higher-order cognitive function dedicated to evaluating the first-order cognitive functions of perception. It therefore fails to satisfy a core requirement of metacognition as Proust defines it, the requirement that 'an operative cognitive subsystem is evaluated or represented by another subsystem' (13). By contrast, in genuine metacognition, information about some cognitive subsystem takes on a life of its own in serving as input to another subsystem.

The metacognitive element to which Proust devotes the most attention—the one she describes as 'the overarching norm' (137)—has an important relationship with the aspect of sensation that we have just identified as non-metacognitive. Proust's central case of a metacognitive signal is fluency, or ease of processing. Sharper sensory signals (say,

black text on white paper) consistently cause a greater experience of fluency than their hazy counterparts (say, yellow text on an orange screen). However, at least in human beings, fluency itself is monitored, and takes on a life of its own; we consciously experience fluency, and these experiences have distinctive effects. One recent review describes fluency as being 'experienced the same way people experience emotional or bodily feelings like appetite or hunger' and as felt not only in connection with perception but across a spectrum of other types of mental activity: 'when people perceive, process, store, retrieve, and generate information, they experience the ease or difficulty of these cognitive operations' (Unkelbach and Greifeneder 2013, 3, 11). Another review describes fluency as a 'monolithic metacognitive cue', observing that 'every cognitive task can be described along a continuum from effortless to highly effortful, which produces a corresponding metacognitive experience that ranges from fluent to disfluent, respectively… fluency is a ubiquitous metacognitive cue that accompanies cognition across the full spectrum of cognitive processes' (Alter and Oppenheimer 2009, 220, 232).

In examining the metacognitive function of fluency, it is important to isolate the distinctive value of the metacognitive cue itself, as opposed to the underlying characteristics of the processing giving rise to it, and the content it accompanies. For example, when we experience shifts in our feelings of fluency as perceptual conditions change, while simultaneously undergoing corresponding shifts in our reliance on various sensory modalities, our feelings are epiphenomenal to those shifts in multisensory integration: they have a common cause, but the feelings play no part in the automatic regulation of the balance of power between sensory modalities.

Delicate empirical work has however shown that feelings of fluency do make a distinctive contribution in a wide variety of settings. In a classic experiment separating the feeling of fluency from the content it accompanies, Norbert Schwarz and colleagues asked experimental participants to recall either six or twelve examples of situations in which they 'behaved very assertively and felt at ease', having established in piloting that most would find it fairly easy to think of six and difficult to come up with twelve (Schwarz, Bless et al. 1991). Participants were later asked to evaluate how assertive they were. Despite having just generated twice as many examples of their own assertive behavior, the participants who had been asked to generate long lists of moments of assertiveness rated themselves as significantly *less* assertive than their short-list counterparts. Schwarz and colleagues argued that the metacognitive feeling of difficulty was taken as informative: if it is hard for me to recall examples of my assertiveness, I must not be an especially assertive person. A final experiment bolstered this interpretation. Participants completed the recall task while listening to music that they were told either hindered or aided the recall of the relevant type of behavior. Given the suggestion of crediting their experiences of difficulty in recall to the interference of the music, participants who had the difficult task of generating lists of twelve examples no longer evaluated themselves as less assertive: with the diagnostic value of their metacognitive experience discounted, its impact on judgment disappeared.

Subsequent work has extended Schwarz's results to a wide variety of other domains, across all of which the impact of fluency on judgment is mediated by some naïve or learned theory (Alter and Oppenheimer 2009). This work has helped to secure the status of fluency as a distinct and unitary feeling, rather than a family of analogous effects. If newly learned

responses to perceptual fluency are transferred over to retrieval fluency, this is a sign that fluency as such is a single factor guiding judgment.

The malleability of responses to fluency is remarkable. Proust describes fluency in somewhat fixed terms, as a positive norm: in her book, fluency serves as a 'correctness condition' for thought, 'a norm that gives a preferential status to cognitive actions that can be comparatively performed with the least cognitive effort' (125). It is true that fluency is typically experienced positively. Fluency increases liking; random stimuli are preferred when we have been subtly primed by prior exposure to process them more easily (Zajonc and Rajecki 1969). Fluency also conveys the 'illusion of truth'; obscure trivia statements come to seem more likely to be true when they are repeated or even just presented in a sharper color contrast. However, it is possible to reverse the perceived valence of fluency. Participants asked to evaluate the truth of trivia statements presented in high- and low-contrast colors on screen will by default rate the higher-contrast ones as more true; however, after a brief training period involving feedback indicating high-contrast statements are false, participants in a test condition (without further feedback) switch to associating greater fluency with falsity, even without reporting any awareness of the impact of color contrast on their judgments (Unkelbach 2007). While the training period manipulated color contrasts, it inculcated a new understanding of fluency itself, rather than anything specific to color contrast or even perceptual ease; the reversal effect carried over to a separate test period in which all statements were presented in black on white but some felt more fluent because they had been flashed on screen earlier.

If the meaning of fluency is learned, rather than fixed, this does nothing to diminish its value. The default truth effect would arise exactly because in non-manipulated environments we are generally likelier to have heard truths as opposed to falsehoods repeated; our general tendency to accept the more familiar answer as the right one is in these environments reinforced over time, and constitutes knowledge of how things ordinarily work (Reber and Unkelbach 2010). Seeing metacognitive cues as having learned meanings in some ways diminishes the distinctiveness of metacognition, however; the subjective signal of fluency comes to signify truth in the same way that the sound of the can opener comes to signify to the cat that it is about to be fed. The guidance provided by such a feeling is not hardwired into the nature of the feeling, but depends on acquired beliefs—learned from experience or generated by an experimenter—about the relationship between the feeling and the mental state of perceiving, imagining or recalling that it accompanies, or is seen to accompany. Presented with a new trivia statement that I've been subtly primed to find unexpectedly fluent, I interpret myself as remembering it, and if the fluency is high enough, I feel confident that this is a well-established truth I am remembering (Kelley and Lindsay 1993).

This way of describing the distinctive character of metacognition diverges from Proust's: she argues that mental state self-attribution is not needed for the guidance provided by metacognition. She takes metacognitive states to have a fundamentally nonconceptual representational structure, signaling the presence of 'knowledge affordances', understood roughly on the model of J.J. Gibson's affordances for action, applied in a 'feature-based representational system' (FBS) that presents possibilities for

cognitive action in response to characteristics of one's current mental task. Metacognitive feelings, in her view, are not 'about' self-attributed mental states; in her view, 'the reason why aboutness is not needed for feelings to guide decisions is that this guidance is procedural; the nonconceptual content of FBS thoughts expresses the specific knowledge affordance associated with a first-order cognitive performance, and thereby guides epistemic decisions' (144). Proust's argument is difficult to follow here. There is no obvious anathema between aboutness and procedural guidance: indeed, the clearest cases of a distinct contribution of metacognitive feelings do seem to be cases of the type identified by Schwarz, in which procedural guidance is made possible by judgments of aboutness.

According to Proust, noetic feelings—feelings such as fluency—'neither refer to all mental states nor to a particular one.' Her reasoning is as follows: 'If noetic feelings were nonconceptually metarepresenting distinctive mental states (such as remembering, versus perceiving, versus judging one's learning), then fluency would not be applied indiscriminately across mental states. For noetic feelings would be controlled by the nonconceptual representation that they concern a given mental state. If noetic feelings do not metarepresent distinctive states, however, fluency should regulate self-evaluation in a liberal way, without heed being paid to the mental state under scrutiny. This is what is found. Subjects easily misapply a norm of fluency to mental states that it cannot regulate, such as judgments of learning.' (144) Proust is certainly right to highlight the plasticity of feelings like fluency: the feeling of fluency can arise in connection with a wide variety of different mental states, and it does not directly report to us whether we are, for example, currently remembering a trivia fact or seeing it under special conditions of priming. We

can also misread the signal sent by fluency, as in her example of judgments of learning, where we naively underestimate the difficulty of later remembering what we are presently studying. However, even if the feeling itself does not have a dedicated internal meaning, it could still be true that fluency only ever serves as a guide for thinking when we take it to be about something, and more specifically, about a mental state of ours, a state seen (perhaps wrongly) as a state of knowledge, perception, recall or imagining.

The core of Proust's case against the self-attributive view lies in her discussion of metacognition in animals incapable of (much) mental state attribution. Do such animals get procedural guidance from metacognitive feelings such as certainty? After being trained to respond one way to a given category (say, dense patterns of dots) and another way to a contrast category (say, sparse patterns), animals can learn to 'opt out' or move to a new trial when they are 'uncertain' (in borderline cases). Proust ably summarizes recent experimental work establishing that borderline cases are not simply a third category reinforced by reward, deflecting one major type of challenge to the theory that animals have feelings of certainty.

However, other challenges await. In a recent article cautioning against assuming that metacognition is sharply distinctive from first-order cognition, Nate Kornell observes that judgments of certainty are not necessarily based on private access to characteristics of one's inner cognitive processes: publically accessible cues such as response time and wavering between options are often thought to serve as the basis for certainty judgments

in research on humans, and it is odd to assume that animals have a distinct and more direct way of evaluating their epistemic position.[1] Kornell writes:

> Being vigilant about the role public cues play in guiding ostensibly metacognitive judgments is a healthy kind of skepticism. Contemplating the framework presented here can produce a deeper level of healthy skepticism, however: Even in a best-case scenario, when metacognitive responses are made based on private, internal cognitive cues, how special is that, really? Is a judgment based on an internal cognitive cue more impressive than a judgment based on an internal noncognitive cue such as hunger? Is it more impressive than a judgment based on a complex, difficult-to-understand external cue? Metacognition may be like a Monet—it looks transcendent from afar, but if you look closely it can become almost ordinary. (Kornell 2014, 162)

Like Proust, Kornell sees animals as responding differently to more and less fluently processed signals. However, Kornell answers the question, 'Do animals monitor certainty?' with: 'Probably not' (Kornell 2014, 147). Animal performance can be explained without appeal to any conscious feelings, he argues, and without appeal to any capacity for self-evaluation or self-reflection, which he, following Janet Metcalfe, takes to be definitive of true metacognition. The 'procedural guidance' trained animals get from detecting their hesitation on borderline cases is on this view not really different in kind from the guidance

[1] Proust herself observes correctly that direct access would be more efficient than inferential access (p.99), but this does not quite settle the case in its favor, and as far as I know the question is still open, with several attractive theories on either side.

we get from perceptually detecting fuzzy as opposed to sharp signals; if animals behave differently when they are more or less certain, this is not necessarily because they are aware of feelings of certainty.

By contrast, human noetic feelings look special even close up. Because they are conscious, these feelings can enter into controlled decisions about how to act. In the tip-of-the-tongue condition, for example, we are aware of ourselves as searching for a word, and this awareness enables us to make a choice about whether to persist in the search or give up and shift attention elsewhere. Proust places a heavy emphasis on the importance of noetic feelings in triggering mental action in particular, suggesting that metacognition evolved because 'a mental agent needs to adjust her cognitive efforts and goals to her cognitive resources'. (4) It is clear that in humans the range of resources is large, and its use connected to metacognitive signals: we are capable of controlled, conscious thought of a type absent (or virtually absent) in other species, and feelings of disfluency are a key trigger for this kind of thought (Alter, Oppenheimer et al. 2007). For animals with a much more limited range of possibilities for cognitive action and communication, it is less clear what the distinct functional role of noetic feelings might be. Proust suggests that 'adequate opting out' is evidence that animals have feelings of confidence that are 'able to adjust to task and context in a flexible way', but it's not clear that nonhuman animals have yet demonstrated competence on tasks that demand the kind of flexibility that would call for anything like our kind of awareness of certainty. In emphasizing the procedural value of metacognition, Proust makes it harder to establish that there is a common form of metacognition in play for us and for animals who have a much more limited range of

procedures. There is much to admire in Proust's detailed discussions of a huge range of issues connected to metacognition—many more issues than I've been able to touch on here—but there's a tension near the heart of her position, and we may need substantially more evidence about the thinking of human and nonhuman animals before it can be resolved.

References:

Alais, D. and D. Burr (2004). 'The ventriloquist effect results from near-optimal bimodal integration.' Current Biology **14**(3): 257-262.

Alter, A. and D. Oppenheimer (2009). 'Uniting the tribes of fluency to form a metacognitive nation.' Personality and Social Psychology Review **13**(3): 219-235.

Alter, A., D. Oppenheimer, N. Epley and R. Eyre (2007). 'Overcoming intuition: Metacognitive difficulty activates analytic reasoning.' Journal of Experimental Psychology: General **136**(4): 569-576.

Ernst, M. O. and M. S. Banks (2002). 'Humans integrate visual and haptic information in a statistically optimal fashion.' Nature **415**(6870): 429-433.

Jameson, A., T. O. Nelson, R. J. Leonesio and L. Narens (1993). 'The Feeling of Another Person's Knowing.' Journal of Memory and Language **32**(3): 320-335.

Kelley, C. M. and D. S. Lindsay (1993). 'Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions.' Journal of Memory and Language **32**: 1-1.

Kornell, N. (2014). 'Where is the 'meta' in animal metacognition?' Journal of Comparative Psychology **128**(2): 143-149.

Kornell, N. (2014). 'Where to Draw the Line on Metacognition: A Taxonomy of Metacognitive Cues.' Journal of Comparative Psychology **128**(2): 160-162.

Ma, W. J., J. M. Beck, P. E. Latham and A. Pouget (2006). 'Bayesian inference with probabilistic population codes.' Nature neuroscience **9**(11): 1432-1438.

Reber, R. and C. Unkelbach (2010). 'The Epistemic Status of Processing Fluency as Source for Judgments of Truth.' Review of Philosophy and Psychology: 1-19.

Schwarz, N., H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka and A. Simons (1991). 'Ease of retrieval as information: Another look at the availability heuristic.' Journal of Personality and Social psychology **61**(2): 195.

Unkelbach, C. (2007). 'Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth.' Journal of experimental psychology. Learning, memory, and cognition **33**(1): 219-230.

Unkelbach, C. and R. Greifeneder (2013). A general model of fluency effects in judgment and decision making. The Experience of Thinking. C. Unkelbach and R. Greifeneder. New York, Psychology Press**:** 11-32.

Zajonc, R. B. and D. W. Rajecki (1969). 'Exposure and affect: A field experiment.' Psychonomic Science **17**(4): 216-217.