Précis of *The Philosophy of Metacognition*

JOËLLE PROUST
Institut Jean-Nicod

This book deals with metacognition, that is: with the epistemic self-evaluation that allows mental agents to *predict* whether they will be able to remember a proper name, to discriminate an object, to solve a given problem etc., and to *retrospectively evaluate* whether their cognitive outputs (what they seem to remember, to discriminate, to demonstrate) are likely to be valid. There has been for several years an important controversy among psychologists and philosophers about the best definition of the concept of metacognition, based on a number of assumptions concerning mental architecture, consciousness, the continuity or discontinuity between humans and nonhumans, and the respective roles, in self-evaluation, of beliefs and feelings. On a classical view, dating back to pioneering articles by James Flavell and by Nelson and Narens, "metacognition" refers to knowledge about one's own knowledge, or thinking about one's own thinking, which involves metarepresentation of one's own epistemic states. This is the "attributivist" conception of metacognition. More recently, the term "metacognition" has been extended from recursive (metarepresentational) to experience-based evaluations: on this latter approach, self-directed mindreading is no longer constitutively involved in a phenomenon which is still called "metacognition". Self-directed mindreading is rather seen as contributing to a more restricted area called "analytic" metacognition. Metacognition "at large", then, is defined as epistemic self-evaluation, whether based on affects or on concepts (and metarepresentations).[1] This is the "evaluativist" conception of metacognition.

Some theorists, nevertheless, have stuck to the initial attributivist definition of metacognition as knowledge about one's knowledge, where metarepresentations are constitutively involved. However, their views about the relation between metarepresentations and self-evaluations vary to a considerable extent. Some deny that feelings can play any role in epistemic self-evaluation and are skeptical about nonhuman metacognition (attribution does it all).[2] Others recognize that feelings can guide uncertainty-based decision-making, and that evidence for this ability in animals is impressive, while also denying that this function is metacognitive.[3] Still others claim that nonhumans use metarepresentations, rather than mere feelings, when evaluating what they perceive or remember, even though they are unable to read others' minds.[4] The first chapters of the book aim at clarifying both the meaning and the scope of metacognition. To do so, a neutral definition is proposed in chapter 1, that does not pre-empt the respective roles of evaluation and of representation. Chapters 2 and 3 aim to spell out four opposing theoretical claims constituting respectively an "evaluativist" and an "attributivist" view of metacognition. These claims are responses to four questions: 1) does appraisal originate uniquely for the self, or can it also be applied to others? 2) what kind of information does epistemic appraisal rely on? 3) Does appraisal require an ability to represent the attitudes being appraised? 4) Does appraisal need to be part of an agentive context? The attributivist view takes appraisal to apply similarly to self and others. It responds positively to

[1] This wider use of the term is instantiated in work by Koriat, Bjork, Strack, Reber, and by most experimental and social psychologists studying metacognition. See for example Koriat & Levy-Sadot, (1999).

[2] Cf. Perner (2012).

[3] Carruthers & Ritchie (2012), p. 82.

[4] Smith et al. (2003).

question 3, negatively to question 4, and considers that the information being relied upon is contained in a metarepresentation of the contents and attitude involved in the task at hand. On the evaluativist view that the book defends, metacognition originates in the self, relies on affective as well as doxastic types of information, does not require metarepresentation of the attitudes being appraised, and is an essential ingredient of mental agency. This set of claims – evaluativism - is only implicit in the experimental study of metacognition, which has mainly concentrated on metamemory. From a philosophical viewpoint, however, it offers a new foundation for a general theory of mental agency.

On this account mental actions are what one does in order to acquire information that one does not currently have available. One can try to gain a better view of an object by moving one's body, acquire a new set of beliefs by learning from others, or one can try to reason about what one already believes to be true. Metacognitive episodes are the constitutive normative steps in these actions. Any efficiently selected and conducted new cognitive action has to meet two sets of structural preconditions. Agents must be able to assess whether the task considered is within their reach and solvable within a certain time span. They must also be in a position, once the action is performed, to reliably evaluate its success or failure.

Granting that we have this capacity to predict how valid our attempt will be (or has been) at retrieving a memory, discriminating an object, etc., there must exist a source of information and associated processes underlying this ability. The attributivist account sketched above belongs to the "representation-as-contemplation" model: In order to control our memory, our perception, etc., we need to represent *that* we are currently attempting to remember, to perceive, etc. But in addition we also need to judge the likelihood of our success by metarepresenting the content of our task: How many times, in the past, have I succeeded at remembering proper names? How does the mathematical content of this problem look to me?

To this second type of question, however, the metarepresentational story has no answer. A particular metarepresentation allows one to shift viewpoints about its first-order embedded content, i.e. allows a decoupling between that content and the world as the attributor knows it to be. But it does not generate novel knowledge relative to the embedded content.[5] Even if metarepresentations allow thinkers to infer the Bayesian probabilities attaching to success in this *type* of task (e.g. if they have always received straight As in math courses), these probabilities cannot comparatively tell them how they will do in this *particular* case. As shown in Chapter 4, an attributive approach fails to provide a natural account for the fact that metacognitive appraisal is activity-dependent. Furthermore, it cannot explain why metacognition exists in species unable to metarepresent their mental states.

Chapter 5 reviews the evidence about metacognitive abilities in non-human primates. The initial methodological problems inherent in the possibility of animal's using behavioural or conditional learning cues to form uncertainty assessments have been progressively identified and, on my view, overcome, in more recent paradigms. Present studies satisfy the operational definition offered by Robert Hampton for metacognition as tested on nonverbal creatures. Thanks to neuroscientific evidence, the mechanisms involved in assessments of uncertainty are now better understood (in studies involving rodents and rhesus monkeys). These studies favour an accumulator model of confidence assessments as a main source of information for epistemic decisions (a model that had already been used to account for experimental evidence in human metaperception and metamemory). In other words, metacognitive prediction and retrospective evaluation depend on the mental activity triggered by a task: this information can only be generated if the agent *engages* in the task.

Chapter 6 has a more speculative aim: to determine the representational format of metacognition in macaques, given that such metacognition does not rely on the representation

[5] See Dokic (2012).

of the subject’s own attitudes. Strawson's notion of a feature-placing system is used as a springboard for hypothesizing that procedural metacognition relies on a similar, fluency-based non-conceptual system, whose function is to detect epistemic affordances. A feeling of fluency, when present, is also used by humans as a cue for likely truth. But fluency does not coincide with truth: and there is no specific experience of truth. Chapters 7 and 8 aim to provide a general conceptual framework within which procedural and analytic forms of metacognition coexist and compete for guidance. Identifying the full scope of the normative conditions of human epistemic agency is no less important than identifying some of the primary mechanisms that have been selected, over phylogeny, to control mental actions via a norm of fluency. Analytic metacognition evaluates cognitive actions when they involve a sensitivity to higher-order epistemic norms – norms that need to be metarepresented in order to guide epistemic decisions. Determining, for example, whether one has exhaustive or only partial knowledge of a complex event is a precondition for many human decisions. Assessing one's conformity to these norms cannot be done in an exclusively affective way.

This said, epistemic norms are not the only constraints that apply to mental agency, because we generally perform mental actions (like trying to remember one's forgotten shopping list) in order to reach further distal goals (like bringing home new items as planned). Two kinds of motives, then, must be present for a mental action to develop. The first kind is instrumental: a mental action is performed because of some basic instrumental need, such as remembering the name of a composer, shopping for food, or solving an equation. The second is epistemic: given one's distal goal, there is a specific epistemic norm relevant to that goal. Fluency is involved in assessing whether the name of the composer can be retrieved. If the goal is to solve a math problem, higher-order norms will be tracked, such as exhaustiveness, or higher-order coherence. These two types of motive are reflected in two different phases of mental or cognitive actions. The first phase determines which epistemic norm to select as a function of one's distal goal. Should I remember my shopping list accurately (only true positives admitted, with misses tolerated) or exhaustively (all true positives retrieved, with false positives tolerated)? This instrumental choice is sensitive to a norm of rationality: The agent tries to use the means that he believes necessary for doing what he intends to do. The second motive is associated with a phase of epistemic evaluation - of the feasibility of the action, and then of its final adequacy. Confidence evaluations will now be governed by the normative requirements associated with the selected mental action, such as accuracy, exhaustiveness, consensus, plausibility or coherence. While the instrumental conditions of success can be overruled without compromising the action (say by taking a longer route than necessary), the epistemic conditions are non-negotiable: once one's epistemic goal is fixed, the associated constraints apply in a strict way.

This difference between conditions of felicity and constitutive conditions deserves emphasis. Some agents may intend to change the world by using means of action that are actually entirely inappropriate. They may also be wrong in believing that they will benefit from the change they intend to bring about. Cumulating these two infelicities, however, does not prevent the corresponding bodily intention to act from being formed and carried out. Mental actions differ strikingly on this account. Instrumental norms also apply to goal selection in their case: agents may be instrumentally wrong in selecting a given informational goal. They can violate a self-imposed norm, for example in confusing one name for another when trying to remember a name and fail to recognize their mistake. Even then, however, the agent is still committed to the goal of correctly remembering the name that fits that description. If he does not, he cannot be said to have tried to remember. An agent cannot be wrong, then, about the norm(s) that constitutes his/her own action, without having failed to act. Agents who are not sensitive to this requirement did not act mentally. This dis-analogy, defended in detail in Chapter 7, pleads for the distinctive functional character of metacognition with respect to ordinary action control. Chapter 8 uses the distinction between two sets of norms to propose a two-tiered theory of acceptance, inspired by Goldsmith & Koriat. Once an epistemic decision is reached about the subjective certainty to attach to a given judgment, an *epistemic acceptance* is formed. Now, however, instrumental reasons can step in and modify the threshold of confidence sufficient for triggering an overt action, as a function of its stakes; this is a *strategic* acceptance. Even when quite certain of your knowledge, you may not want to use it when it concerns a matter of life or death.

Chapter 9 explores the implications of an evaluative perspective on metacognition for epistemology. According to epistemic internalism, justification of an epistemic agent's beliefs depends on factors that are accessible to, or knowable by, the agent. Epistemic externalism, in contrast, is the view that justification depends on the objective reliability of the subject's cognitive systems, which believers may not be in a position to evaluate. Although epistemic feelings seem to pave the way to internalism, these feelings themselves depend, for their reliability, on objective environmental conditions (which influence calibration of assessments). A thought experiment, inspired by an actual experiment that we performed concerning the metacognition of change blindness, suggests that epistemic externalism fares better with metacognitive calibration processes than internalism does.

Chapter 10 discusses Peacocke's view that awareness of having acted mentally is what entitles a thinker to make true judgements about her own mental actions. Epistemic feelings, rather, are proposed to be a nonconceptual source of information that can (defeasibly) entitle an agent to acquire true beliefs about her own mental agency. Chapter 11 explores the relations between human metacognition and self-reidentification, through the notion of a semi-hierarchical control system. Chapter 12 attempts to show how such a system can be used to account for schizophrenic delusions. Patients usually are normally conscious of moving and of thinking -- they have a preserved sense of "ownership" for their ordinary or mental actions. They may deny, however, that they are the author of certain of their actions or thoughts -- they have an impaired sense of agency. Failures in the predictive and executive power of the models of action and on the metacognitive processes involved are hypothesized to influence the phenomenology of agency, and to generate the impression of thoughts being inserted. Chapter 13 focuses on "conversational metacognition", i.e., the set of abilities that allow an embodied speaker to make available to others and to receive from them specific markers concerning his or her own “conversational adequacy”. Because of its rather quick tempo, conversation is a case where procedural metacognition rules the game. Fluency and informativeness jointly determine what is relevant in a message. Speaker and hearer monitor each other's uncertainty through facial gestures, and to some extent, control, when necessary, what they display. A concluding chapter proposes that the solutions offered in the book are compatible with a two-system view of metacognition.

References

Carruthers, P. & Ritchie, J.B. 2012. The emergence of metacognition: affect and uncertainty in animals. In: Beran, M.J., Brandl, J. Perner, J., Proust, J. (eds), *The Foundations of Metacognition,* Oxford: Oxford University Press, 76-93.

Dokic, J. 2012. Seeds of self-Knowledge: noetic feelings and metacognition. In: Beran, M.J., Brandl, J. Perner, J., Proust, J. (eds), *The Foundations of Metacognition,* Oxford: Oxford University Press, 302-321.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34 (1979): 906-911.

Goldsmith, M., & Koriat, A. 2008. The strategic regulation of memory accuracy and informativeness. In A. Benjamin & B. H. Ross (Eds.), *The Psychology of Learning and*

Motivation. London: Academic Press, 48: 1–60.

Hampton, R.R. 2009. Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition and Behaviour Reviews*, 4, 17-28.

Koriat, A. & Levy-Sadot, R. 1999. Processes underlying Metacognitive Judgments: Information-based and experience-based monitoring of one's own Knowledge. In: S. Chaïken & Y. Trope, (Eds.), *Dual-Process Theories in Social Psychology*. London: The Guilford Press, 483-502

Nelson, T.O . & Narens , L. 1990. Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (pp.125-173). New York: Academic Press.

Peacocke, C. 2008. Mental Action and Self-Awareness (II): Epistemology, in L. O'Brien and M. Soteriou (eds.) *Mental Action* Oxford, Oxford University Press, 192-214.

Perner, J. 2012. Minimeta: in search of minimal criteria for metacognition, in: Beran, M.J., Brandl, J., Perner, J., Proust, J. (eds.), *The Foundations of Metacognition,* Oxford: Oxford University Press, 94-116.

Smith, J. D., Shields, W. E., & Washburn, D. A. 2003. The comparative psychology of uncertainty monitoring and metacognition. *Behavioural and Brain Sciences*, 26, 317-373.

Strawson, P. F. 1959. *Individuals.* London: Methuen.