Synthese (2007) 159:271–295
DOI 10.1007/s11229-007-9208-3

Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition?

Joëlle Proust

Published online: 28 September 2007

Abstract Metacognition is often defined as thinking about thinking. It is exemplified in all the activities through which one tries to predict and evaluate one's own mental dispositions, states and properties for their cognitive adequacy. This article discusses the view that metacognition has metarepresentational structure. Properties such as causal contiguity, epistemic transparency and procedural reflexivity are present in metacognition but missing in metarepresentation, while open-ended recursivity and inferential promiscuity only occur in metarepresentation. It is concluded that, although metarepresentations can redescribe metacognitive contents, metacognition and metarepresentation are functionally distinct.

Keywords Metacognition · Metarepresentation · Theory of mind · Reflexivity · Epistemic feelings · Self-prediction · Causal-contiguity · Inferential promiscuity

A good part of our mental life is devoted to evaluating our mental performance, and predicting how well (or badly) we can do, have done, or are doing in a new job, a new task, or a new social situation. This is the domain of metacognition: thinking about one's own thinking. It is exemplified in all cognitive activities in which one is trying to appreciate, in retrospect, a cognitive achievement (did I do this task right?, haven't I forgotten something?), to remember the source of the information one is using, or to predict whether one will be able to attain some cognitive goal (learn new material, retrieve a proper name within seconds, or make efficient plans in a new context).

J. Proust (✉)
Institut Jean-Nicod (CNRS-EHESS, ENS),
29 rue d'Ulm, 75005 Paris, France
e-mail: jproust@ehess.fr

Most researchers in theory of mind have taken for granted that metacognition, in this sense, necessarily involves the second-order representation, i.e., the metarepresentation, of first-order cognitive contents. In other terms, modularists as well as theory-theorists have assumed that the capacity of three-year-old children to attribute knowledge to themselves, as well as the capacity to distinguish a real from an apparent property of the world (reviewed in Flavell 2004) develop as a consequence of the children's mindreading skills, and more specifically, are made possible by their having metarepresentational ability.

When trying to explain why children attribute true or false beliefs to others (solve a false belief task—or FBT) at around 4 and a half, most theoreticians have emphasized that children only succeed when they are able to metarepresent the corresponding mental states. Radical simulation theorists (Gordon 1996) have rejected this assumption. Their claim is that children succeed at FBTs when they are able to simulate the perspective of another child on a given situation, when it is different from the one they know themselves to hold. Radical simulation theorists, however, have not produced a fully adequate explanation of children's understanding of false belief.[1] Non-radical or "hybrid" simulation theorists, on the other hand, have recognized that *decoupling* simulations and reasoning *acros*s them—rather than reasoning *within* a simulation—was a key ingredient in a mindreading capacity.[2] Such a decoupling has been shown to be semantically equivalent to forming a metarepresentation of a first-level simulated content. For reasons that will not be discussed here, exploiting a simulation from without amounts to forming a metarepresentation in the broad sense of the term (decoupling the interpreted content from a larger set of contexts where it can receive different truth evaluations).[3]

On the basis of this large consensus, this article will take it for granted that metarepresentation is necessary to ensure the kind of decoupling involved in mindreading. Furthermore, it may be a precondition for certain forms of metacognitive effort, in particular those that are regulated by *explicit* self reference. An interesting and debated question, however, is whether metarepresentation has to occur for metacognition to be possible at all. This question is interesting, for if it can be shown that metacognition can be dissociated from metarepresentational capacity, one might infer that different evolutionary pressures are at play in metacognition and in mindreading. There is, as will shall see, some evidence that metacognition may be present in animals without a theory of mind. These results are still controversial, however, because the animal data on which the case for non-metarepresentational metacognition is based raise methodological problems that may interfere with interpretation of the results.

There is another dimension to the question that has theoretical import: metacognition is by definition a form of self-predication; metarepresentation, however, is not essentially reflexive. As a semantic skill, it can be used in self- or other-attribution, even

[1] For critical reviews of radical simulation theory, see Dokic and Proust (2003) and Nichols and Stich (2003).

[2] See Perner (1996), Nichols and Stich (2003), Proust (2003b). The contrast between exercising simulation and exploiting it allows one to contrast performance in 3-year-olds and 4-year-olds (Recanati 2000). We will come back to this contrast in Sect. 1.

[3] See Recanati (2000) chapters 4–7 for a detailed defense of this claim.

if it can be argued that having access to one's own beliefs and desires is more immediate than inferring beliefs and desires in others. We will review below the conceptual and empirical arguments that can be adduced to show that there are non-metarepresentational forms of metacognition.

The central section of this article will be devoted to an examination of the main defining features of metacognition. We will further examine how a decision about the respective structures of metacognition and metarepresentation impacts on issues such as epistemic transparency and the ability to pursue recursively into higher orders. We will begin our discussion, however, by observing that an utterance constructed as a metarepresentation can admit various cognitive interpretations. The distinction between deep and shallow types of processing will turn out to be quite important in discussing the relations between metacognition and metarepresentation.

1 Metarepresenting: shallow and deep forms

By a "metarepresentation", is meant a representation whose content includes: (1) a first-order representation, such as "it is raining" and (2) the representation of an epistemic or a conative attitude directed at that content, such as, "I believe (I perceive, I regret) that it is raining". A metarepresentation is thus structured as follows:

(1) PA [self] {TC}

If metacognition is defined as "thinking about thinking", then it seems (at least prima facie) to necessarily require representing not only the thought content TC together with the propositional attitude PA_1 (the belief or desire) that attaches to it and the representation of oneself as having that attitude, but also a second-order propositional attitude PA_2 that bears on PA_1. When evaluating one's current belief state, or learning state, or when predicting whether one will remember a name, one must represent oneself as believing that P or as learning material M. That is,

(2) PA_2PA_1 [self] {TC}

It is on the basis of this kind of analysis that Tim Shallice has claimed that in order to monitor an intention, one must metarepresent (PA_2) that one intends (PA_1) to do A. On this view, what explains why patients with schizophrenia have trouble monitoring their actions is that they have a disturbed mindreading capacity.[4]

A similar theory has been defended to interpret 3-year-olds' failures in judging the modal source of acquired knowledge. Flavell considers that such a task requires that a child understand the relations of the isolated event [of visually perceiving that X is in the box] with inputs, other mental states, and behavior. In short, the child becomes successful in such tasks when he can use (that is: learn) a functional theory of mental states. This theory allows the child to understand that representation tokens acquired in specific places, at specific times, and from specific viewpoints have differential roles in modulating behavior. Dienes and Perner,[5] more recently, defend a higher-order

[4] For a detailed discussion of this view, see Proust (2006a).

[5] Dienes and Perner (2002): see the last section below for a critical discussion.

theory of consciousness designed to account for the metarepresentational structure of metacognition.

Let us observe that, on a metarepresentational view of metacognition, not every metarepresentation is metacognitive: one can metarepresent a first-order representation in the absence of a mental/predictive, self-directed goal. Take, for example, the following assertion:

(3) This photo fails to convey the luminosity of Leonardo's Last Supper.

This sentence expresses a metarepresentation: the full-view photograph represents the fresco, and the fresco represents a religious scene. There are several ways to understand (3) according to the *background capacities* that may or not be invoked in processing (3). One may or may not actively search one's own episodic memory of having seen Leonardo's Last Supper; one may or may not recognize that this task involves visual perception (a pigeon, say, could be trained to recognize the similarity of two pictures without using the concept of vision or of visual experience, and without identifying the scene to which both the fresco and the photo refer). In other terms, one may miss one or several of the associated metacognitive thoughts that are necessary for a deep understanding of (3). If asked to assent to (3) as soon as it is heard, (assuming that one is currently perceiving both), for example, one might simply compare the colors. Provided with more time, one might remember one's initial experience of the fresco, notice finer details of the faces and come up with a deeper interpretation of what (3) means.

This case suggests that a distinction should be made between several ways of understanding (3)—a distinction that generalizes to all kinds of thoughts. A first shallow way to understand (3)—let us call it 3A—is one that restricts itself to the *perceptual* aspects described in (3) while ignoring the representational facts that are communicated in (3). When comparing the photo and the fresco (assuming again that our subject is currently perceiving both), a viewer does not need to explicitly think that the photo is supposed to represent the fresco, which itself *represents the Last Supper*—an event that has relevance for Christians, etc. Even someone who does not know that a full-view photo is an optical duplicate of the full-view appearance of an independent object can compare the perceptual appearances of a painting and a photo and form a judgment of similarity between two presentations. In that sense, (3) can be understood in a non-metarepresentational way (although, admittedly, an important portion of the information contained in (3) is lost). This does not, however, prevent our observer from making a competent judgment on whether there is the same luminosity in both pictures.

A second type of *shallow understanding* of (3)—3B—is one that restricts itself to the *semantic status* of (3). This time, the hearer captures the metarepresentational content of (3), but cannot evaluate its truth for herself (never having seen the original fresco). To understand (3) in this second shallow way, one need only understand the fact that a given photo is representing a given referent—namely a painted wall representing a religious scene. In this case, although one knows the general truth conditions of (3) in the abstract, one is prevented from evaluating the truth of (3) for lack of perceptual and memory access to the fresco. In sum, a correct semantic metarepresentation offers no guarantee that a rich mental meaning is constructed. Even if one thinks that

mindreading involves metarepresenting thought contents, it must involve additional ingredients. Otherwise, mental understanding remains shallow in the sense indicated. As an assertion, (3) does not involve mental concepts. As a communicated thought content, reached after the relevant inferences are made, however, it does. Mental concepts contribute to widening the scope of the knowledge relevant to interpreting its meaning. But shallow types of interpretation do suffice in many cases: a photographer working on the print does not need go beyond a perceptual analysis to think (3).

Verbal representation is therefore no sure guide to the actual content of a thought. It does not necessarily reflect the kind of information-processing that is actually used by a subject when forming that thought. Information-processing, however, determines representational form and content. As is well known to cognitive scientists, a cognitive subject may have a given capacity, but fail to exercise it in some circumstances. Shallow metarepresentation, therefore, occurs not only when one does not have the appropriate knowledge; it may simply be a manifestation of one's rational propension to simplify a task, and to make the cost appropriate to the stake.

The next question, then, is how to characterize "deep processing" in general terms. My proposal in this article is that processing is "deep" when it deploys information in a way that offers a dynamic prediction and/or justification of its output (3C), rather than associating perceptual (3A) or linguistic cues (3B) with a predetermined answer, as is done in shallow processing.

The 3C type of understanding can be gained when one represents one's own previous encounter with the work, and engages in a comparison of one's visual percept of the photo with one's memory of the fresco. This last part involves, aside from clear-cut mindreading considerations as to the painter's intentions, various cognitive operations such as: focusing attention on a context or event so as to relate one's memory to one's perception in the relevant way, which in turn involves: (a) recognizing the identity of reference between fresco and photo; (b) controlling one's memory to retrieve one specific episode with its relevant properties; finally (c) comparing present and past visual experiences.

In other words, a "rich" understanding, in both the Last Supper case, and the mental case of belief/desire self-ascription, can only be attained if a subject can evaluate or predict, for herself, in a first-person engaged mode of thinking, the truth value of a verbally expressed or mentally represented proposition. First-person engagement, as we shall see below, refers to a way of thinking that exercises a subject's reflexive metacognitive capacities. This proposal accounts for an interesting asymmetry between metacognition and metarepresentation that we will discuss in the last section: metarepresentation has a semantic structure that allows for shallow processing, metacognition does not; metarepresentation however has an inferential capacity that metacognition lacks. In other words, metarepresentation and metacognition can complement each other, but have different specializations.

The reader may object at this point that the distinction between shallow and deep metarepresentational thinking may well apply to the relations between a photograph and its object; but not to *epistemic or conative self-attributions*. If I attribute to myself or to another subject the second-order propositional attitude of having the first-order propositional attitude of, say, believing that it is raining, then the objection goes, I must necessarily form an engaged, reflexive judgment to that effect. I must form the belief

to be able to form the metabelief, and be committed to the truth of the first operation to infer the truth of the second.

2 Metarepresenting mental content

This objection can be phrased in either first-person or third-person attributional terms. There seems to be no room for having *shallow access* to a conceptual relationship between the fact that I presently believe (or know) that P and the fact that I presently believe that I presently believe (or know that I know) that P. There seems to be no room, similarly, for having shallow access to what someone else believes, and yet metarepresenting him as believing it.

A full answer to this objection, however, will only be possible in the last section of this article, when the asymmetry between shallow and deep forms of processing has been examined in all its relevant aspects. As a starter, let us examine two facets of the contrast between first-person and third-person mental attribution: (a) Epistemic transparency applies only in the first-person case, not in third-person mental attribution; (b) recursive metarepresentation is much more easily performed in third person than in first person cases. Explaining this double contrast will help us strengthen and develop the point made in Sect. 1: metacognition is present in first-person self-attributions, but not necessarily in third-person attributions. Shallow processing is never possible in metacognition, while in metarepresentation, contrary to the objection raised above, it is commonplace. We will see in the final section how to derive from these facts the conclusion that where deep processing is needed, recursion is restricted.

(a) Epistemic transparency—the principle allowing one to infer knowing that one knows that P from knowing that P—is valid in modal and epistemic logic. Epistemic transparency about knowledge may fail in certain cases even in second order—we will remain agnostic on this debated issue (Williamson 2000; Dokic and Egré in press). But from the third order on, it seems clear that recursive semantic ascent is *not* within the reach of ordinary thinking. Even if one has a clear picture of what it means to know that one knows that P, it is less clear that one has different ways of appreciating

(4) I know that I know that I know that P,

from the way of appreciating that one knows that

(5) I know that I know that I know that I know that P.

This fact is quite puzzling,[6] in particular when viewed in light of the fact that we do have (in general) epistemic transparency at the first recursive level.

(b) recursive metarepresentation in first and third person attributions.

How can one explain that we face such severe limits on recursion in self-knowledge, whereas we have no problem using and understanding recursion with embedded higher-orders attributions when several *different* attributors are hierarchically involved? It is not very difficult, for example, to understand the following sentence:

[6] Other metacognitive predicates seem also to be subject to the same limitation: "believe", "be aware", "feel", "doubt", etc.

(6) "I know that Anna knows that her father knows that her mother knows that her grandmother knows that she is invited for lunch on Sunday".

What kind of account can we offer for this difference? Our previous distinction provides the beginning of an answer. Our limitation does not reside in an inability to understand the *formal possibility of recursion*. Indeed logical systems capture in exact terms the properties of recursion. A written formula for a recursive proposition is easy to learn and to reproduce. Nor does the problem reside in an inability to form self-attributive judgments, for we succeed at the second-order level. *It resides rather in our limited ability to engage first-personally in a self-directed n-order recursive thought for n greater than one*. Although we understand fourth- or fifth-order attributions in the third-person case, we fail to simulate self–self–self simulations as being *distinct* cognitive operations.[7] While the third-person case can be accounted for by the recursive properties of metarepresentation, the first-person limitations can be explained by the properties of metacognition.[8] To complete this point, we need to turn to these properties.

3 What is metacognitive content?

There is a form of engagement that explains "rich" understanding in both the Last Supper case, and in the mental case of belief/desire self-ascription. Constraints associated with such an engagement should account for the limitations that apply to recursive self-attribution. We will now attempt to list the various properties that jointly define metacognitive engagement. Our goal is to discuss these various properties to get a clearer picture of the relationship between metacognition and metarepresentation.

[7] Richard Carter has objected here that a sentence including different tenses works better: "I know that I knew (yesterday) that I knew (the day before) etc." On his view, this suggests that what makes it hard to find any context or interpretation in which "I know that I know that I know that P" is true, is that the three "knows" have present reference. Such contexts however can be constructed: "I know that I know that I have always known/that I will always know that P" has a clear interpretation. Carter finally observes that the limitations also hold for the corresponding third-person thoughts: "John knows that he knows that he knows that P" (where the two "he's" refer to the same John). Carter's objections can both be addressed in a similar way. In both cases (adjunction of tensed contexts, and third-person attribution), metacognition and metarepresentation overlap. When representing my epistemic relation to P at other times, and comparing them, I am not scrutinizing my present cognitive adequacy (as in "Do I presently know P?" or in "May I predict whether I will remember P"?). I am not therefore performing a metacognitive action; I am rather using the mental concept of "knowing" in its fully general scope across various metacognitive episodes, which requires using a metarepresentation of various metacognitive states. Conversely, it is as difficult to understand "John knows that he knows that he knows that P" as the first-person parallel sentence, because in both cases, the only interpretation that would make sense is the "engaged" one, i.e., an interpretation gained through metacognition. As simulationists have argued, self-simulation is a necessary ingredient in many kinds of other-attributions (those that qualify as "engaged"). It is an interesting consequence of the present view that overlaps between metacognition and metarepresentation can be apriori predicted/explained on the basis of the engaged/shallow distinction.

[8] Compare with the preceding footnote: each "know" has a subject with a different reference, unlike the sentence above with "John". This allows creating metarepresentational contexts from a succession of engaged metacognitive meanings.

1. Metacognitive engagements *are predictive or retrodictive*.
2. Prediction and retrodiction are part of *a self-directed evaluative* process.
3. They have a *normative and motivational* function: evaluation in turn produces revision, adjustments, based on prediction or retrodiction.
4. This evaluative process is not explicable *in first-order terms.*
5. This evaluative process is not explicable in *second-order terms.*

3.1 They are predictive or retrodictive

Prediction is the essence of cognition as a mental function.[9] Cognitive systems excel at extracting in an implicit way the world regularities on which their actions depend. Not all cognitive systems, however, are able to extract regularities about their own *epistemic* or *motivational* properties. The capacity to predict whether one will be able to retain an item of knowledge in memory until the following day, or week, is not available to younger children, nor to most non-human animals. Predicting how one will cope with physical danger, or social stress, or whether one has the skills required for a certain project to be executed smoothly are also relatively late acquisitions in human ontogeny.

Metacognition is also exercised retrodictively: one may judge after the facts that one has made a bad decision, one may regret a purchase, or, on the contrary, feel justified in having made it. One may notice that one has reacted inappropriately in a social context. Retrodiction thus engages episodic memory, with a prospective aim of revising one's behavior or motivations. Actually, prediction and retrodiction are closely associated, as self-prediction relies on feedback from past operations, and self-retrodiction has the function of reorienting one's future actions.

Should prediction and retrodiction be explained in metarepresentational terms? As we saw in the introduction, most theorists have been prompt to infer from the notion of metacognition that one needs to metarepresent cognition in order to control it. Early work on metacognition (understood at the time in the more limited sense of "metamemory") accordingly proposed that prediction of future ability requires two-level cognitive processes. A metalevel "contains a dynamic model (e.g., a mental simulation) of the object level." (Conant and Ashby 1970; Nelson and Narens 1992). The metalevel is seen as a control level: it promotes commands for the initiation, continuation, and termination of epistemic actions (such as controlled remembering). The object level, on the other hand, informs the control level by sending it feedback from its commands, i.e., by "monitoring" them.

This general schema is helpful in contrasting a command level, which has a world-to-mind direction of fit (because it aims at producing an outcome in the world, such as naming an individual person), with an observation level, which has a mind-to-world direction of fit (because it registers the responses that are actually produced, which may be a simple feeling of knowing the name, a tip of the tongue impression). I will try to show, however, that a metarepresentational account for the relation between command and monitoring is not adequate.

[9] See Proust (in press a).

Let us start with a non-metacognitive case such as evaluating a possible bodily action. In order to predict whether you can jump over a ditch, for example, you have to simulate, on the basis of your implicit knowledge of your motor ability and the perceptual cues available, whether the jump will be easy or problematic. In such cases, you simply simulate your jumping, i.e., you imagine yourself jumping over the ditch in front of you. This simulation normally allows you to predict in a very reliable way what you are able to do "in reality"[10]. Pace Nelson and Narens, simulating does not involve representing *the fact that you jump*: simulating is just running a dynamic motor representation off-line, and obtaining internal predictive feedback on this basis. In conceptual terms: the function of simulation is not to represent yourself as doing something; it is rather to prepare to do something, that is, to do it in pretend mode.

Nor does receiving internal feedback as a consequence of your simulation require any metarepresentation of your past jumpings. Receiving feedback has a quasi-observational nature, as it results from the perceptual consequences of the simulated command (this is why it is often called "introspection"). In what sense is this kind of observation "internal"? Feedback is called "internal" because it is not *presently* based on external cues; it originates, however, in a memory of standard visual and proprioceptive percepts, as it is constituted by bodily (vestibular, visual, etc.) reafferences[11] stored in prior jumpings. You merely use your implicit non-conceptual, dynamic knowledge; you don't need to declare to yourself that you have it. Thus it is simply false that a metarepresentational redescription of the object-level is a precondition for control to occur.[12]

What is true for the prediction of bodily action also holds for mental action (i.e., metacognitive action). When you need to retrieve a memory, you simulate initiating the action of remembering, and compare the feedback you obtain with that you expect to find. If the cues match your expectations, (the norm, as computed from prior episodes of successful remembering), then you consider that it is worth spending time and energy trying to remember. If the cues fail to match, you abandon the goal.[13]

In sum: a metarepresentational conception of the relationship between control and monitoring fails to capture the dynamic simulatory nature of the command-level. Control does not report or describe what is the case at the object level. It rather causes new events at that level, that will in turn drive new commands (of a different sort). The reason why theorists have considered that control was describing an object level is that, as we shall see, both levels share content (see the discussion of representational promiscuity across levels in the next section).

We will analyze in Sect. 3.5 below various important connections between command and monitoring that a metarepresentational view of their relations fails to capture.

[10] See Decety et al (1997).

[11] I.e., perceptual inputs collected as a consequence of an action.

[12] Cf. the evolution of Frith's view on the perturbation of controlled action in schizophrenia; for a review, see Proust (2006a).

[13] For a model of metamemory bond on experimental evidence, see Koriat (1993).

3.2 Prediction and retrodiction are part of *a self-directed evaluative* process

As we saw in Sect. 3.1, control and monitoring are the two relata of a closed causal loop. This property can be rephrased by saying that there is *causal contiguity* between the control and monitoring levels (which means that control directly affects monitoring, which in turn directly affects further control). One crucial aspect of this dynamic reafference is that it retains the causal structure of on-line control-monitoring. What is retrieved in memory is retrieved as a result of the present intention to retrieve it. The metacognitive prediction ("that I know the name of Peter's daughter") is a response to a prior command with a closely related content ("Do I know the name of Peter's daughter?"). Causal contiguity is the solution that Hugh Mellor[14] has developed for explaining what distinguishes what he calls "subjective beliefs" (such as "I face food now") from "objective facts" (such as "X faces food at t"), which are fully explicit truth-conditional expressions of states of affairs. What makes a belief subjective is that causal contiguity guarantees that an individual's belief will make his desires cause him to act to satisfy them. "Causal contiguity is how a subjective belief refers to whomever has it, and when". (Mellor 1991, p. 24).

Causal contiguity, as a general causal structure, belongs to every adaptive control system where feedback comparison is made to causally respond to command in a systematic way. In our case, however, the control system is cognitive, which means that representations—beliefs and desires—are used to control behavior. As philosophers of action have shown, using various terminologies, causal contiguity relations between the cognitive subsystems engaged in action are associated with an intriguing conceptual relation. If you intend to act, your intention constitutes the condition of satisfaction of the completed action.[15] This property of an intention to be embedded in the content of an action, is what I will call *representational promiscuity.*[16] In every cognitive control structure, there is such representational promiscuity. Monitoring *instructs* the command, while the command *directs* or organizes upcoming monitoring (by offering predictions about what is to be the case if the action develops normally). There is representational promiscuity because control and monitoring share the basic informational pattern that drives the causal loop to the expected goal.

As noted by Mellor (1991), an important consequence of causal contiguity is that self-reference does not need a self concept to be instantiated. More generally, there can be token-reflexive dimensions of thinking with no corresponding concept; the modes of presentation that are expressed by words like "here", "now", or "I" can be exhibited by organisms that do not master the corresponding concepts, but are equipped with the corresponding causally contiguous control-informational device. In other words, the very structure of the causal and informational exchanges between control and monitoring builds reflexivity into the operating mode, and into its output. Such "procedural" or architecture-bound reflexivity is more than a precursor of explicit semantic reflexivity (as expressed by indexical words). It is the locus where all forms of reflexivity

[14] See Mellor (1991, ch. 2).

[15] See Searle (1983), and Proust (2003a) for an analysis of Searle's view.

[16] On a closely related use of this concept, see Carey and Xu (2001).

are generated. Where there is reflexivity, *there must be an adaptive control system engaged in monitoring its commands.*

Every token of metacognition can provide illustrations for this procedural reflexivity: a student who is rote-learning a poem, for example, must appreciate whether *she now* knows the poem; an aging speaker must appreciate whether s*he* will be able to remember a proper name in a minute (*from now*), etc. This dependence of operation on self-evaluation and various task-relative (temporal and spatial) indexicals does not *need* to be achieved in explicit semantic terms as long as self-reference is the indefeasible background default of the operating system.

To qualify as procedurally reflexive, a further condition must hold, and it deserves to be spelled out even though it results from the conjunction of causal contiguity and representational promiscuity. A procedurally reflexive system must not be a mere combination of distinct devices that happen to produce a joint effect via a causal loop; it must have the intrinsic *function of affecting itself*, and do so in a causal-representational way (through informational means). *In other words, there is an evaluative dimension in procedural reflexivity* that should appear as a definitory feature of metacognition. When producing an evaluation of its own operating mode, the system reflects its intrinsic reflexivity: being closed under a norm, the cognitive system generates commands to change itself in order to adjust to a changing environment. As we saw in Sect. 3.1, the decision to initiate/revise a command is produced by an information-based (simulatory) estimate of the probability of success/failure at a task. Success or failure are outcomes that *motivate* the agent to metacognize (i.e., predict and evaluate) its own states.

In contrast to (metarepresentational) mindreading, metacognition is *always evaluative, rather than merely predictive*. You can predict, for example, that Jane intends to eat bananas rather than pears, but this prediction does not involve your own preferences (even though it involves an evaluation by Jane of what is best for her). *Predicting* future states generally implies anticipating trajectories in internal (mental) or external (physical or behavioral) dynamic events or event sequences. *Evaluating* future states involves in addition appreciating the efficiency of a given course of action, which means comparing internal resources with objective demands for the task. A judgment of learning, or an evaluation of one's emotional level, for example, involve norms of adequacy: the goal of such judgments is to find an efficient or reliable way of coping with a set of requirements.

Note that ordinary bodily action also requires flexibility and adjustment to local conditions, and these corrections are part of what "intending" or "willing to do P" mean. In metacognition, however, flexibility and adjustment concern the *informational* resources presently available. A subject cannot change much of the set of epistemic or motivational resources available to her at any point in time, but she may acquire control over pragmatic aspects of her thinking that crucially affect its outcome. She may plan ahead of time to gather resources for a demanding task. The time she spends on it (looking, considering, reasoning), the acuity with which she perceives an object, the chosen distance from and orientation to the perceptual input, the duration of the mental effort given to collecting information, all these correlate with significant differences in the validity of the output. In sum: an agent cannot change the content of her

attitudes, but she can improve the *informational quality* and the *cognitive adequacy* of her mental processes in a controlled way.[17]

3.3 Monitoring information quality and cognitive adequacy have a normative and motivational function: evaluation produces revisions and adjustments

By "informational quality", is meant the optimal signal-to-noise ratio of a sensorimotor or classificatory operation. The longer one attends to an object, the less one is distracted, the better one will do at recognizing or categorizing it. By "cognitive adequacy" is meant the correct evaluation of the resources needed in a reasoning task, given its importance. Correct choice of resources obviously depends on the value or utility of available outcomes. It presupposes preferences, a trade-off mechanism to integrate values,[18] and a correct appreciation of the constraints that apply to each action in the repertoire.

Although the function of metacognition is to approximate a norm, cognitive adequacy, it is far from evident that the mind/brain has mastered multi-purpose and reliable heuristics applicable to a majority of situations. Classical work on metacognition, for example, has studied on which basis subjects appreciate how well they have learned a list of names. It has been shown that people often confuse ease of processing (associated with block learning) with efficient memory encoding (associated with temporally distributed learning). Confusion between these two types of evaluation inevitably leads the subjects to an incorrect evaluation of how well they have learned a given piece of material—block learning subjects overestimate their learning rate, while distributed learning subjects underestimate it. What this case shows is that metacognitive self-attribution (i.e., monitoring) can be negatively affected by a mistake in detecting the kind of epistemic feeling relevant for a task. The feeling that material is *easy to understand* is confused with the feeling that it is *easy to retrieve from memory*. Confusing the feelings, however, leads to two types of cognitively inadequate decision: one chooses to stop learning too early (in the block practice condition); or, if one has learnt in a distributed way, one prefers to return to block practice on a later occasion, (although it is in fact much less efficient).

3.4 This evaluative process is not explicable in first-order terms

Critics of metacognition argue that all that is needed for "metacognition" is a capacity to predict the Bayesian likelihood with which a certain desired outcome will occur. A dedicated neural reward system encodes *disparities* between an animal's expectations and its experience of success. In other terms, it encodes errors in the prediction of future reward.[19] Such outcome evaluation, the criticism goes, can proceed on a first-order cognitive basis: actions will be selected if they maximize rewards in

[17] See Proust (2006b, 257 ff, in press b).

[18] See Sugrue et al. (2005).

[19] See Schultz (1998).

response to specific sensory stimulations. A common currency for reward seems to be implemented in the dopaminergic neurons.[20] Metacognition thus boils down to decision-making under uncertainty. On this view, called "an actor-critic model",[21] there is no substantial cognitive difference between judging one's chances at succeeding in a given bodily action (such as hitting the bullseye) and evaluating whether one has learned a list of names. In both cases, one just stores one's past success rate for tasks of the same type. One therefore does not need to appeal to inner resources: one judges one's capacity as anyone else would, by the number of hits versus failures. To know what your future disposition is, just look at your prior performance: look at the world, not at the self.[22] Self-report performed on this basis will be shallow, masquerading as second-order evaluation.

To respond to this objection, we can use two types of arguments. One type is conceptual.

(7) Knowing (believing) that a reward of probability p is associated with stimulus S is not equivalent to

(8) Knowing (believing) with probability p of accuracy that a reward is associated with stimulus S.

In addition to the changing world, a distinctive source of uncertainty may be that generated in the knower. Pragmatically, the two kinds of uncertainty are similar, because the patterns of response look identical when the crucial variables fail to be controlled. Although it may be difficult to disentangle them when studying non-human behavior, everyone will appreciate that the following two sentences refer to very different states of affairs:

(9) I am so confused that I don't know any more whether Peter will come.

(10) The circulation is so difficult that I don't know any more whether Peter will come.

What makes the distinction difficult to establish on the basis of behavior alone is that most experimental paradigms use noisy percepts as stimuli in collecting decisions made by monkeys. These stimuli are indeed ambiguous between a cognitive and a non-cognitive interpretation. A second difficulty is that the distinction as we have presented it uses explicit mental concepts. In order to circumvent this ambiguity, however, experimental paradigms have to be framed in comparative psychology in such a way that objective probabilistic interpretations can be excluded.

Our second argument is that precisely this type of distinction seems to explain why some kinds of animals (primates, marine mammals) perform well at certain "metacognitive" tasks, while others, which are good first-order predictors, such as pigeons and rats, fail them. Let us examine the tasks involved in more detail.

Bottlenosed dolphins (*Tursiops truncatus*) have been offered a task in which they have to press one pedal for low-pitched auditory stimuli, and another pedal for high-pitched ones. The animal subject was offered a third pedal, which allowed it to decline

[20] See Sugrue et al. (2005).

[21] Dayan and Abbott (2001), Sugrue et al. (2005).

[22] Gareth Evans'routine might also be discussed in this context. See Sect. 4, where it is discussed.

trials when it was uncertain.(Smith et al. 1995) Animals used the third pedal as human subjects do. It is now recognized by the authors that the "uncertain" response might have been reinforced to cope with objective uncertainty. In that case, it would not qualify as metacognitive. Rhesus monkeys have also been tested in their ability to perform, or to bail out from, a visual density discrimination task (Smith et al. 2003). The animals had to discriminate whether a display contained 2,950 pixels. They responded in positive cases by moving a cursor to a box, in negative cases by moving it to an S shape, and to a star to decline the trial. Monkeys have again been found to have a pattern of response similar to humans.

The uncertain response is known to have a specific profile, both in humans and in monkeys: it is quite variable across individuals. For this reason, some researchers consider that it is prompted by "extraserial attitudes and decisional temperaments" (Smith 2005). A potential problem for this experiment, however, is that the animals have to respond to an occurrent situation. It is particularly difficult, in a simple perceptual paradigm, to tease apart objectively produced and subjectively appreciated uncertainty.[23]

A research project responding to this worry has been conducted using a memory monitoring paradigm: a serial probe recognition task. In one version of this task by Hampton (2001), monkeys are presented with an icon, that they will subsequently need to recognize among distractors after a varying delay, in order to obtain a reward. A first crucial constraint is that they decide to take the test (after the delay) *before* seeing the test stimuli, or rather chose an alternative, easier task. Taking this decision before seeing the test stimuli, they have to rely on their memory of the icon. As Hampton emphasizes,[24] this task elicits a prospective judgment of memory in monkeys. To solve it, monkeys must rely on the equivalent of our human feeling of knowing. They cannot base their response on the perceived familiarity of the test stimuli, for they have not yet seen them: they cannot consult the world to decide what to do.[25] Exposed to the same material, pigeons show that these two conditions are critical.[26] They can only perform the task when they perceive the test stimuli, not on the basis of their memory of the icon. These experimental results indicate that animals without a mindreading capacity, such as monkeys or dolphins, can still succeed in using metamemory. This in turn suggests that mental metarepresentation might not be the initial input that makes metacognition possible.

Let us summarize this point. Metacognition, when it is present, draws on a kind of information that is not delivered by the problem situation, but by the subject's own procedural self-knowledge. For that reason, metacognition can deal with novel decisions, while well-practiced routines remain within the scope of cognition (where external cues can be used as predictors). These two features seem to provide a divide

[23] Smith and colleagues addressed this difficulty since then in Smith et al., 2006. See also Kornell et al., 2007.

[24] Hampton (2005).

[25] That absence of memory was causing the response was experimentally controlled by introducing trials where no icon was presented. The decision to decline is also present in these cases, suggesting that the animals did not learn to predict result from such cues as noises, grooming, or motivational changes.

[26] Inman and Shettleworth (1999).

between the capacities exemplified in primates and marine mammals, on the one hand, and in pigeons and rats,[27] on the other.

3.5 This evaluative process is not *explicable* in second-order terms

The intuition that metacognition involves second-order self-knowledge is compelling, and researchers in the field themselves have found it attractive (see introduction and section 3.1). In summary: while cognition is typically engaged with the external world, metacognition is engaged with one's own informational and motivational states. Its goals are typically endogenous. Given that metacognition is about representational capacity, we can, at least, conclude that a crucial aspect of metacognitive reasoning seems to require metarepresentations.

The evidence discussed in Sect. 3.4, however, suggests that the intuition in question should be resisted. Marine mammals as well as monkeys seem to have good metacognitive capacities (metaperception and metamemory). On the other hand, they typically fail false belief tasks. They do not seem able to metarepresent (conspecifics' or their own) mental states *as mental states*. We are thus confronted with a dilemma: either (1) we reject the view that metacognition is a distinct capacity from mindreading, with the burden of showing either—against present evidence—that there is no metacognition in non-human animals, or—also against present evidence—that they have a theory of mind. Neither of these avenues seems promising. Or (2) we accept that these animals have metacognition, but we show that metacognition does not require a mentalistic metarepresentational capacity. We are now in a position to argue that such is the case: cognitive adequacy can be a goal for a cognitive system unable to metarepresent its own states as representations.

We have already shown, in Sect. 3.1, that metarepresentation is not involved in performing the kind of simulation that a cognitive system needs in order to evaluate its own capacities when engaged in a given task. In Sect. 3.2, we examined the causal-contiguity structure of a metacognitive loop, its representational promiscuity, and its functional role in generating procedural reflexivity. Assembling these various considerations, we can conclude that (a) metacognition is not inherently metarepresentational, and (b) that metarepresentation is not inherently metacognitive. Let us elaborate both points.

(a) Metacognition is not metarepresentational in the sense that there is no "report" relation between command and monitoring, but functional complementarity of a basic kind (Sect. 3.1). In each metacognitive intervention, a command token inquires whether typical conditions for a desirable/undesirable outcome (learning, remembering, versus forgetting, confusion, etc.) are "now" present; the corresponding monitoring token uses present reafferences to offer a context-based answer (feeling of learning, feeling of knowing, cue-based inferences, etc).

One might insist that the command metarepresents the condition that should prevail in the cognitive engine. For example, a typical metacognitive outcome could be rephrased in terms of the structure of (1):

[27] For a recent study on rats, suggesting meta-cognition ability, see Foote and Crystal (2007).

(11) I [self] believe (PA_2) that I remember (PA_1) that P {TC}

But this analysis fails to capture two properties of metacognition. First, in a standard metarepresentation such as (1) or (11), the first-order content is *independent* of its second-order metarepresentation. P can be thought without being the object of higher-order propositional attitudes. Second, there is also truth-functional independence. It may be false that P, while true that I believe it. Someone else can recognize this fact, and attribute to me the false belief that P.

In a genuine metacognitive interpretation of (11), however, these two properties do not apply. Let us take a metacognitive outcome such as

(12) I can remember how to come back home.

Here, the monitoring (or reafferent) content is gained as a consequence of a metacognitive query ("will I remember how to come back home?"): it does not have any autonomy relative to the command level. The command must be issued for perceptual input to be anticipated and compared with observed cues for the purpose of evaluation. Monitoring is not a passive information pick-up that would have occurred no matter what in the absence of a previously "trained" command-level.The same idea can be articulated by contrasting the representational promiscuity[28] that occurs in metacognition and the redeployment of content that characterizes mental metarepresentations. In (1) and (11), a thought is represented along with the *concept* of the propositional attitude in which it is embedded. In contrast, in (12), remembering does not have to be conceptually represented; it only has to be *exercised* as a trying; that is, simulated as a consequence of the need to come back home at some future time.

Second, there is no truth-functional independence between control and monitoring contents. If the monitoring level delivers a non-matching result (e.g. I don't remember P), the control level cannot use it correctly as a matching result. A metacognitive control loop aims at establishing a coherent and reliable picture of the presently available cognitive capacities. In contrast, metarepresentations generally are truth-functionally independent from their embedded representations: the attribution of a belief to John may be true even though what John believes is false.

These differences in content can now be further explored and characterized using the contrast discussed in Sect. 1 above. In metacognition, a thinker engages in a simulation in a motivated, first-person way, on the basis of his/her prior procedural knowledge of the conditions of success of a given mental task. Metacognition can never go "shallow"; it cannot predict or evaluate without simulating, which means spending significant time and resources on running a dynamic model for the task. Given however that metacognition does not necessarily involve mental concepts, it does not enjoy "inferential promiscuity", i.e., the capacity to combine inferences and generalize across domains.[29] The kind of prediction or evaluation that metacognition produces is domain-specific, and can only become inferentially rich if metarepresentation is used to redescribe its outcome through mental concepts. For example, if as a young child, I do not remember how to get back home, I will need to find alternative ways to acquire the information. If this occurs when I am an adult, I can obtain

[28] On this concept, see Sect. 3.2.

[29] On inferential promiscuity, see Stich (1978).

inferential knowledge based on my mental concepts, I will be able to infer that my memory loss is alarming, that I need see a doctor, train my memory etc. These inferences and subsequent decisions are only accessible to agents able to (conceptually) metarepresent *that they have a bad spatial memory.*

In summary: metacognition is neither first-order, nor second-order. We might call this initial, emergent metacognitive level "level 1.5".

(b) Metarepresentation is not inherently metacognitive.

When metarepresentation redescribes metacognition, it automatically receives a deep reading. But it can also apply to other kinds of content in a way that does not involve a deep reading. It might be because autistic children perform well on a task of picture/source relation, but fail in a task of mental metacognition, that they perform well on the Zaitchik photograph task (which is metarepresentational), while failing on a false belief task.[30] The difficult question that we raised in Sect. 2 is whether metarepresentation can be applied in a shallow way to mental contents. It is now time to address this question.

As we saw in Sect. 1.1, metarepresentational shallowness may occur without jeopardizing inferential capacity. This suggests a possible line of response to our question. As long as you don't need to evaluate the truth of an embedded content to correctly apply a mental concept to it, you don't need, a fortiori, to test and evaluate your own judging or learning capacity in order to form that metarepresentation. There are many inferences to be drawn from the attribution of belief to someone, even when you don't know what the actual state of affairs is that is represented by the believer. For example, you may report to yourself Anna's belief by saying.

(13) Anna believes that the train leaves at 6

Just because you heard her say "the train leaves at 6", without taking it to be true (or caring whether it is true or not) that the train leaves at 6.[31] We store and report to others in this shallow way many ordinary beliefs, with their informational source, without submitting them to critical examination. This in turn means that there are concepts and representations that we can metarepresent without having the ability or the need to apply to objects and states of affairs. This way of dealing with embedded contents might be particularly useful in forming metarepresentations when one has failed to fully understand the embedded representation. As Sperber observes, we can "think about" a thought without "thinking with it". In such cases, we have an intuitive belief whose content is metarepresentational without having an intuitive belief at the object level. As a result, as Sperber shows, it is impossible to "disquote" the reported content, for it cannot be used as a regular belief. Sperber's concept of a non-intuitive, or purely reflective metarepresentation, therefore, seems to qualify as "shallow" metarepresentational processing.

[30] See Zaitchik (1990), Leslie and Thaiss (1992).

[31] In Sperber (1997), a distinction is offered between intuitive and reflective beliefs which is useful in the present context. But as will appear in this section, it is interesting to see how shallow a reflective belief (i.e., a metarepresentation) can get. The notion of a validating context implicit in any metarepresentation according to Sperber might in many cases derive from metacognition rather than from the metarepresentational structure of the belief.

It might be objected, however, that when a subject stores the informational source from which she acquired a reflective belief, as seems necessary in order to utter sentences like (13), she certainly needs to use her own metacognitive competence. For to store the original conveyor of the information that *p*, the objection goes, one must be able to engage in a simulation of the way one acquired that information. Therefore deep first-person engagement is needed in most reflective, metarepresentational beliefs.

Although this kind of analysis is quite popular among mindreading theorists, it may not square well with recognition of the contrast stressed in Sect. 3. The point of contrasting "having a capacity" and "exercising it" was to show that, although one can use one's mentalizing abilities in many circumstances, one can also approach a question in non-mental terms, by simplifying the problem. There are deep forms of engagement available to an agent who has access to metarepresentational thinking, because normally such an agent independently possesses metacognitive capacities. But she does not need to exercise them, in particular when under time pressure, or in routine situations. For example, simple association of a sentence with an agent [Anna, *p*] can replace a full-blown, engaged simulation of the corresponding belief activity (with its network of normative constraints). A subject who already has these mental concepts in her repertoire may replace them by their shallow, non-mental counterparts.

Such counterparts, furthermore, have been recognized by several authors. Among psychologists, Josef Perner has observed that although human adults can use a metarepresentational approach to fully understand the mental import of a conversation, or of a social context, they don't need to. In most cases a "situational" approach to the problem at hand will help one recover the information that needs to be shared and jointly processed. Processing a social context through "situation theory" amounts to using the world as the model from which to predict behavior, instead of using another subject's representation of the world. We resort to representational theory "only when we need to": "remaining a situation theorist whenever possible can save many unnecessary complications … it saves the gory details of reshuffling mental representations."[32] The reasons Perner offers for staying with situation theory processing of contexts are the same we offered in our discussion of having/exercising a capacity to exploit deep levels of processing: simplicity (and reliability given the task at hand).

Among philosophers, Gareth Evans described a shallow understanding of belief, through a procedure that has come to be called "the ascent routine".

> In making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward – upon the world. (…) I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states. (…) But it seems pretty clear that mastery of this procedure cannot constitute a full understanding of the content of the judgment 'I believe that *p*'. (Evans 1982, p. 225)

[32] Perner (1991, p. 251 & 252).

Evans's point (also made by Gordon 1996), is that non-mentalizers can make *shallow linguistic* use of a metarepresentation such as "I desire that p", or "I want that p". Looking at the world, seeing what is the case, and reporting it as a belief may be done in conversing with someone without deploying the concept of belief: the procedure of embedding a content within an epistemic verb may be socially learned before one has realized what "belief" actually means.

It is interesting to observe that, although Evans's point was made to account for a *shallow form of metarepresentation*, the phenomenon he describes may also be interpreted *as a deep metacognitive (non-metarepresentational) self-engagement*. As suggested in the preceding section, the verbal expression "I believe that *p*" should be analyzed differently if it is an explicit answer to a linguistically expressed utterance (as suggested by "answering" a question, in which case it is metarepresentational) or to a mental token of metacognition (the question "answered" is raised by the thinker herself, in which case it is metacognitive and not necessarily metarepresentational). A non-mentalizer might thus entertain in thought a *metacognitive equivalent* of "I believe that p"—in a way that requires neither a metarepresentation nor the mastery of the concept of belief. This is the case, for example, when the subject has reasons to be subjectively uncertain of a situation P (where P is not objectively uncertain: for example her memory for a given event is weak). What forming this metacognitive self-attribution does require, however, is that the subject engage in appropriate self-simulation (comparing the feedback obtained to a stored norm).

The expression "self-simulation" may be somewhat misleading, if one takes it to involve "turning inward". In self-simulating herself, a subject also "turns to the world"—but in a "deeper" way, through dynamic prediction of reafferences and evaluation.[33]

If these observations are correct, there are several ways of opposing a "precursor" to a "genuine" "I believe". The metarepresentational type of precursor "I believe" is what Evans had in mind. In that case, an utterance of "I believe" merely emphatically expresses P. In contrast, a metacognitive precursor "I believe" is not verbally expressed, but rather procedurally exercised by giving attention to the cognitive adequacy of (what is in fact) the belief that *p*.

The full-fledged deep metarepresentational form of "I believe" requires exercising *both* the capacity to metacognize that I believe and the ability to master the concept of belief in a general way, i.e., to apply it in third-person attributions.

[33] "Turning to the world", let us note, does not mean only "turning to peripersonal space". Bodily states belong to the world, as do the contents of memories which were generated by "external" states of affairs. It is arguable that the epistemic feelings generated in the monitoring part of the metacognitive loop have a "mind to world" direction of fit: the thinker who enjoys these feelings extracts information on her mental dispositions from her bodily states. The fact that, most of the time, we perceive the external world through vision should not obfuscate the fact that proprioception is always an important mode of perceiving the world, our body being a part of it. Most metacognitive reafferences are proprioceptive rather than visual. See however, the fascinating results discussed in Levin (ed.) (2004).

4 Recursivity, transparency and reflexivity

The preceding section allows us to dissipate the puzzle about recursivity and transparency presented in Sect. 2. On the one hand, transparency holds in metacognition, not in mindreading: When I know or believe something, I immediately know that I know or believe it, but I do not immediately know what you know or believe. Your mental contents are not transparent to me: I need to infer them from your behavior. This is intuitive, but it should now also be theoretically obvious: the only system that can be used as a direct model of a knowledge state is the system that instantiates the corresponding internal model of the world. It is in virtue of this model structure of the mind that each mind is only equipped to directly simulate its own belief states.

In order to extend simulation of one's own mental states to the mental states of others, as one does in mindreading, one will have to modify the model of the world to be simulated, that is, construct an ad hoc simulation in which some crucial beliefs are missing or have different truth values. In this process the notion of evaluation will be lost in part: one cannot compare in an engaged way (metacognitively informed) the reafferences imputed to another subject with the norm *that* subject acquired via his previous experiences. The modified dynamic model does not coincide with the self-simulation dynamic model either, but only overlaps with it. This is why self-simulation can only partly account for mindreading, and needs to be supplemented with third-person kinds of generalizations. In mindreading, the evaluative component thereby becomes less prominent than the descriptive–predictive component.[34]

Now the link between transparency and recursivity becomes clearer. The reason why recursivity is open to modeling multiple others' embedded mental states rather than a single individual's is that this operation relies on metarepresentation, which relies on the syntactical phenomena of natural languages. It is indeed a universal formal property of human languages that they admit embedded clauses. If our hypothesis above is correct, metarepresentation is *not as such* cognitively demanding. It is implicitly mastered through language use. Embedded clauses are understood quite early in life, although children typically make mistakes in processing correctly the information delivered by such sentences when the clauses conflict in meaning or reference.[35] An experienced speaker, however, can sort out the variations in references and modes induced by example (6), repeated here:

(6) "I know that Anna knows that her father knows that her mother knows that her grandmother knows that she is invited for lunch on Sunday".

Some features of (6) help us in parsing the various attributions of mental contents: first, a "mental" word followed by the syntactical marker (that) introducing an embedded clause signals a new attribution; second, the different proper names work as social cues on which to anchor a new attributional verb and a new content (spatial cues used in gesturing can also be used to index perspective). Third, the temporal succession of the clauses is a cue made linguistically available and rehearsable. Iteration

[34] An exception might be constituted by "collaborative" mindreading (occurring in joint action), where each agent needs to predict and evaluate jointly (in an engaged way) the dynamics of knowledge acquisition and revision available to the group. This interesting problem cannot be examined here.

[35] See de Villiers (2000) and Tager-Flusberg (2000).

is thus inscribed in the expressive vehicle as well as in the organization of information. Understanding recursion, however, does not require engagement. What makes metarepresentation difficult is not its recursive structure, but rather the necessity it may involve of representing simultaneously conflicting views of the same situation (decoupling them).

Now why is recursive embedding so limited in the case of metacognition? The reason, again, has to do with the function and structure of metacognition. The function of metacognition is to predict, retrodict and evaluate one's own mental dispositions, states and properties. Such a function is practical and content-oriented. It has no interest in varied perspectives on the same content, for the only perspective of interest is one's own. Simulating one's simulation of P coincides with simulating P, because both simulations run the same dynamic self-model. If as I argued above, metacognition is predictive and evaluative, it does not need to develop hierarchies of higher-order self-simulations, except in the sense of running simulations for contexts of various scopes or interests (with behavioral, contextual, episodic, or social identity goals).[36] Neither is a recursive hierarchy of self-simulation available to metacognition, if, as argued above, there is no automatic informational access to linguistically based recursive mechanisms in non-linguistic metacognitive processes. The conclusion to be drawn, therefore, is that transparency only apparently involves recursion; it is only when translated into verbal terms, for the benefit of self-justification to others, that one uses recursion to report one's metacognitive states. Those states however are generated independently of the syntax of the verbal report.

4.1 Can self-engagement be subpersonal? A defense of procedural reflexivity as semantically relevant

An important objection to the present approach to metacognition (in particular as articulated in Sects. 3.4 and 3.5) is that only a conscious agent can *give an instruction*, or *issue a command*, as well as *register what is the case* in response to a former *question*. It would seem, however, that metacognition occurs essentially at a subpersonal level, which would well account—the objection goes—for the fact that it does not involve metarepresentation or mental concepts. A further observation can be leveled against the present analysis. It was suggested above, (particularly in the discussion concerning (12)) that the verbal expression of a metacognitive evaluation *does not* correspond to its functional role, as such verbalization inevitably superimposes mental concepts on practical procedures. Is not a metacognitive command (such as one that could be verbally expressed by "is this proper name available?") a causal process, rather than an intentional one, given that a thinker may have no conscious access to it and does not interpret it in rich mental terms? To get to the core of the objection: is it not a category confusion to claim that reflexivity for mental content is derived from the properties of the mind as an adaptive control system engaged in monitoring its commands (see Sect. 3.2)?

[36] On this question, see Proust (2003b).

To respond in full to these important objections would require more space than is available here. I will however summarize the various methodological and theoretical considerations that can be adduced in defending level 1.5 of metacognitive processing. Consciousness is often identified with person-level information processing, taken to constitute an adequate grounding for self-engagement. The present approach suggests a different picture of self-engagement, as well as of the emergence of a self-representation. *The present proposal is that metacognition is an essentially reflexive mental function allowing an explicit form of self-representation to eventually emerge*.[37] There are however forms of self-engagement that do not require the conscious mode, as we already know from neuroscientific evidence. Insofar as metacognition is *the* crucial capacity for building up self-identity, its important property is not consciousness, understood as the capacity to verbally report one's mental states or to attribute them to oneself, but reflexivity, the capacity to evaluate and revise one's cognitive states (whether epistemic or conative). Therefore, our definition of metacognition, to be non-circular, should rely on normative forms of self-guidance, rather than on a full-fledged representation of oneself.

Several researchers have recently defended a metarepresentational view of metacognition in combination with a higher-order-theory of consciousness (Rosenthal 2000a, b; Dienes and Perner 2002). The view is that conscious seeing, remembering, acting, etc., require some form of metarepresentation. Metarepresentation, when it is applied reflexively to one's first-level thought in an immediate way (not mediated by inference), makes the first level representation conscious. The second-order thought, in turn, is made conscious when it is immediately represented by a third-order thought. In this view, metacognition, being cognition about one's cognition, is seen as a representational mechanism for producing conscious thoughts. A twist in the theory is that a rich conceptual articulation of the relation of the experience to the event that caused it is not necessary for consciousness (although it is for mindreading). Self-referentiality that is inherent to metacognition does not need to be made explicit, a view closely similar to the view defended here.[38]

Metacognition, on the view of these higher-order theorists, has two functions. The first is to make mental states conscious by making them the content of higher-order thoughts.[39] Metacognitions make first-order contents conscious by providing the non-inferential metarepresentation that makes a mental state conscious. They are not themselves conscious, however, unless a third-order thought metarepresents them. The second function is inherent to the semantic structure of a metarepresentation: it is that of applying [implicit or explicit] concepts to first-order contents, which in turn provides the inferential structure needed for reasoning about one's own states as well as those of others. It is important to note that, on this view, the concepts used are used tacitly unless they are metarepresented in appropriate third-order thoughts.

[37] For a full account, see Proust (2003b, 2006b).

[38] "Children and animals, by using such concept of seeing as they do have, can have conscious visual experiences (but only when they use this concept in an assertoric way that does not appear mediated to them)" (Dienes and Perner, Sect. 5). See also Perner and Dienes, 2003.

[39] Dienes and Perner write "Fully explicit knowledge, in our sense, is thus a case of metacognition; what implicit representations lack is metacognition about them" (Sect. 5).

This higher-order approach to metacognition seems to offer a promising route to integrating a theory of consciousness with a theory of metacognition. But it has serious drawbacks.

First, it fails to characterize some aspects of metacognition that we have attempted to account for in this article, namely the relationship between recursivity and metarepresentation, as opposed to the relationship between metacognition and self-engagement. The open-endedness of having recursively conscious access to $n-1$ contents through our n-level reports does not square with the restricted capacity we have recognized in such self-directed recursive escalation. It is also open to the traditional objections against higher-order theories of consciousness, as discussed in Gennaro (2004). On the view defended here, metacognition is in part unconscious, and therefore does not presuppose a general theory of consciousness.[40] Rather, a theory of consciousness needs to account for the fact that metacognition only becomes conscious through its reafferences: epistemic feelings provide conscious feedback on the operation of an (often unconscious) prior command. Feedback however may not need to be conscious to play a causal role in guiding behavior.[41]

Second, it fails to account for the existing evidence of a phylogenetic stage at which animals enjoy some metacognitive capacities (and thereby "reason" about their mental states), while failing to metarepresent them.

But a more interesting and deeper problem for these theories is that they do not question the fact that representing oneself as oneself (i.e., as the same over time) is a precondition for metarepresenting one's own states. They thus accept a static representational view of the self supposed to be there independently of the cognitive history of the system. In a metacognitive view of self-construction, a self results from the process through which a system is able to affect itself in order to maintain stable values across time. Levels of reflexivity—from procedural to explicit—need to be distinguished and constructed in developmentally plausible terms; these levels cannot merely result from the ability to escalate a metarepresentational ladder. Philippe Rochat,[42] following Ulrich Neisser, has listed five different senses of selfhood successively realized throughout ontogeny. These successive reflexive states, in the view developed here, all presuppose at bottom a form of adaptive control, i.e., a structure that exemplifies the properties of causal contiguity and representational promiscuity. Without the *cognitive engagement* in thought that metacognition makes possible, it is not clear how a self might represent itself "from the inside" and develop (more or less) coherent preferences over time.

Acknowledgements I am deeply indebted to Radu Bogdan and to Richard Carter for their help with linguistic form and philosophical content for an earlier version of this article. I am also grateful for discussions with Jérôme Dokic, Paul Egré, Josef Perner, François Recanati, and the participants in the APIC seminar at the Institut Jean-Nicod.

[40] For empirical evidence, see Cary and Reder (2002).

[41] See Cary and Reder (2002).

[42] See Rochat (2003)

References

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: Beyond object files and object tracking. *Cognition, 80*, 179–213.

Cary, M., & Reder, L. M. (2002). Metacognition in strategy selection: Giving consciousness too much credit. In M. Izaute, P. Chambres, & P. J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 63–78). New York: Kluwer.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1*, 89–97.

Dayan, P., & Abbott, I. F. (2001). *Computational and mathematical modeling of neural systems*. Cambridge: MIT Press.

de Villiers, J. G. (2000). Language and theory of mind: what are the developmental relationships? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds* (pp. 83–123). Oxford: Oxford University Press.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). Brain activity during observation of action. Influence of action content and subject's strategy. *Brain, 120*, 1763–1777.

Dienes, Z., & Perner, J. (2002). The metacognitive implications of the implicit-explicit distinction. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 171–190). Dordrecht, Netherlands: Kluwer Academic Publishers.

Dokic, J., & Egré, P. (in press) Margin for error and the transparency of knowledge.

Dokic, J., & Proust, J. (Eds.) (2003). *Simulation and knowledge of action*. Amsterdam: John Benjamins.

Evans, G. (1982). *The varieties of reference*. Oxford: Clarendon Press.

Flavell, J. H. (2004). Development of knowledge about vision, in Levin, D.T. (ed.), *Thinking and seeing. Visual metacognition in adults and children* (pp. 13–36). Cambridge, MIT Press.

Foote, A. L., & Crystal, J. D. (2007). Metacognition in the rat. *Current Biology, 17*, 551–555.

Gennaro, R. (Ed.) (2004). *Higher-order theories of consciousness. An anthology*. Amsterdam: John Benjamins.

Gordon, R. M. (1996). 'Radical' simulationism. In P. Carruthers, & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 11–21). Cambridge: Cambridge University Press.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences USA, 98*, 5359–5362.

Hampton, R. R. (2005). Can rhesus monkeys discriminate between remembering and forgetting? In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition. Origins of self-reflective consciousness* (pp. 272–295). New York: Oxford University Press.

Inman, A., & Shettleworth, S. J. (1999). Detecting metamemory in nonverbal subjects: A test with pigeons. *Journal of Experimental Psychology: Animal Behavioral Processes, 25*, 389–395.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science, 18*(1), 64–71.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition, 43*, 225–251.

Levin, D. T. (2004). *Thinking and seeing. Visual metacognition in adults and children*. Cambridge: MIT Press.

Mellor, H. (1991). I and now. In *Matters of metaphysics* (pp. 17–29). Cambridge: Cambridge University Press.

Nelson, T. O., & Narens, L. (1992). Metamemory: A theoretical framework and new findings. In T. O. Nelson (Ed.), *Metacognition, core readings* (pp. 117–130).

Nichols, S., & Stich, S. (2003). *Mindreading*. NewYork: Oxford University Press.

Perner, J. (1991). *Understanding the representational mind*. Cambridge: MIT Press.

Perner, J. (1996). Arguments for a simulation-theory mix. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories mind* (pp. 90–104). Cambridge: Cambridge University Press.

Perner, J., & Dienes, Z. (2003). Developmental aspects of consciousness: How much theory of mind to you need to be consciously aware? *Consciousness and Cognition, 12*(1), 63–82.

Proust, J. (2003a). Action. In B. Smith (Ed.), *John Searle* (pp. 102–127). Cambridge, Mass.: Cambridge University Press.

Proust, J. (2003b). Thinking of oneself as the same. *Consciousness and Cognition, 12*(4), 495–509.
Proust, J. (2006a). Agency in schizophrenics from a control theory viewpoint. In W. Prinz & N. Sebanz (Eds.), *Disorders of volition* (pp. 87–118). Cambridge: MIT Press.
Proust, J. (2006b). Rationality and metacognition in non-human animals. In S. Hurley & M. Nudds (Eds.), *Rational animals*? Oxford: Oxford University Press.
Proust, J. (in press a). What is a mental function? In A. Brenner & J. Gayon (Eds.), *French philosophy of science, Boston Studies in the Philosophy of Science*.
Proust, J. (in press b). Is there a sense of agency for thought? In L. O'Brien (Ed.), *Mental action*. Oxford: Oxford University Press.
Recanati, F. (2000). *Oratio Obliqua, Oratio Recta*. Oxford: Blackwell.
Rochat, Ph. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and cognition, 12*(4), 717–731.
Rosenthal, D. (2000a). Consciousness, content, and metacognitive jujdgments. *Consciousness and Cognition, 9*, 203–214.
Rosenthal, D. (2000b). Metacognition and higher-order thoughts. *Consciousness and Cognition, 9*, 231–242.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*, 1–27.
Searle, J. R. (1983). *Intentionality. An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Smith, J. D. (2005). Studies of uncertainty monitoring and metacognition in animals and humans. In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition. Origins of self-reflective consciousness* (pp. 242–271). New York: Oxford University Press.
Smith, J. D., Beran, M. J., Redford, J. S., & Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General, 135*(2), 282–297.
Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin. *Tursiops truncatus Journal of Experimental Psychology: General, 124*, 391–408.
Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences, 26*(3), 317–373.
Sperber, D. (1997). Intuitive and reflective beliefs. *Mind and Language, 12*(1), 67–83.
Stich, S. (1978). Beliefs and subdoxastic states. *Philosophy of Science, 45*, 499–518.
Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience, 6*, 363–375.
Tager-Flusberg, H. (2000). Language and understanding minds: connections in autism. In S. Baron-Cohen, H. Tager-Flusberg & D. J. Cohen (Eds.), *Understanding Other Minds* (pp. 124–149). Oxford: Oxford University Press.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Zaitchik, D. (1990). When representations conflict with reality: The pre-schooler's problem with false belief and 'false' photographs. *Cognition, 35*, 41–68.