Animal Behavior and Cognition

ABC 2019, 6(4):309-328
DOI: https://doi.org/10.26451/abc.06.04.10.2019

From Comparative Studies to Interdisciplinary Research on Metacognition

Joëlle Proust

Institut Jean Nicod

Corresponding author (Email: joelle.proust@gmail.com)

Citation – Proust, J. (2019). From comparative studies to interdisciplinary research on metacognition. *Animal Behavior and Cognition, 6(4),* 309-328. https://doi.org/10.26451/abc.06.04.10.2019

Abstract – The goal of this article is to critically examine the notion of metacognition, based on comparative, developmental and neuroscientific publications. A number of researchers define "metacognition" as "knowing what one knows." Others define it more broadly as a set of abilities allowing an individual to control and monitor his/her own cognitive activity" – where "cognitive activity" is taken to mean "activity with an informational goal." Developmental, neuroscientific and comparative studies, however, show that cognitive agents can pursue informational goals and reliably monitor them without representing their own mental states as mental states: they enjoy "procedural" metacognition. Various objections raised in the literature against this hypothesis are discussed, such as the kind of reinforcement at work in metacognition, and the role of metacognitive awareness in human and nonhuman decision-making. Finally, Peter Carruthers' first-order account of the comparative and developmental evidence of metacognition in terms of "basic questioning" is compared with the account in terms of procedural metacognition.

Keywords – Metacognitive development, Fluency, Procedural metacognition, Reinforcement learning, Non-human self-awareness, Evolution of consciousness

In the last two decades the field of metacognitive studies has undergone a succession of insights and revisions in which animal evidence has played a central role. In spite of the pervasive influence of comparative studies on cognitive science, there is still a residual scientific gap to be bridged between animal and human metacognition. Only some developmental psychologists take nonhuman studies into account in their own research on children's metacognition. Reciprocally, animal metacognition researchers themselves rarely consider what developmental studies or neuroscience have to say about the uncertainty response and its mechanisms.

The goal of this article is to compare the respective contributions to the study of metacognition of various areas of cognitive science.

Two Definitions of Metacognition

Over the history of science, stylistic preferences for one term over another often prevail at the expense of mutual understanding and knowledge acquisition. This is the case for metacognitive studies. The current lack of consensus about metacognition in developmental and comparative studies can be explained in part by its history and the terminological variations associated with successive theories. John

Flavell's early developmental research implanted the view that metacognition gradually develops from non-metacognitive toddlers to metacognitive adolescents. Under his influence, metacognitive abilities were seen to emerge only in late preschool years (Flavell, Green, & Flavell, 1995; Lockl & Schneider, 2002). A consensual justification of this finding was that mental states cannot be accessed without being explicitly represented as mental states, that is: without a capacity to read one's own mind – a capacity that emerges at around 5 years of age. Developmental psychologists based their conviction on the converging evidence that, *in verbal tests*, 3-year-old children are unable to reliably report what they know or do not know, or assess their uncertainty (Lockl & Schneider, 2002). Animal studies of metacognition, however, have demonstrated that nonverbal metacognition is present in nonhuman animals. Experimental work by Hampton (2001), Inman and Shettleworth (1999), Kornell, Son, and Terrace (2007), Shields (1999), and Smith, Shields, and Washburn (2003) found a mixed reception from developmental psychologists.

A number of developmentalists have resisted this evidence to this day because it looks incompatible both with Flavell's theory of metacognition, and with the etymology of the word "metacognition" (Perner, 2012), taken as the ability to "reflect on one's own mental processes," or to "know what one knows." *Knowing* what one knows, or *reflecting about* one's mental processes, however, require *possessing concepts* – of mental states, of *knowledge* – and assessing beliefs as being justified and true (or not).

Why does this line of reasoning look tempting? It is arguable that, given its Greek prefix, the *word* "metacognition" *suggests* a second-order reading, in the sense in which metarepresenting one's beliefs is a second-order representation.[1] A second-order representation (i.e., a *metarepresentation*) is a representation that is *about* a first-order *representation* – about its meaning. Any metarepresentation refers to the *meaning* (*or semantic content*) of a proposition. For example, I represent myself as forming a first-order belief such as "I solved this problem." Here I form a judgment not only about a problem being solved. The belief being expressed is that *I, myself*, with salient properties in mind such as age, gender, social origin etc., solved it.

The decision to use the term "metacognition," however, was made in the last century in an entirely different scientific context. At the time, the most significant findings were those of Joseph Hart (1965) about what *he* called "memory monitoring.[2] John Flavell coined the term *metamemory* to refer to Hart's findings, maybe on the model of the term "metalanguage" as used in philosophical logic.[3] It seemed obvious to him that a feeling of knowing is a judgment about what one knows, hence a second-order representation about a first-order knowledge state. Claiming in our present scientific context that metacognition is *by definition* a second-order state, however, has the unfortunate consequence of making theory a matter of definition, thus considerably hampering new research. This chapter tries to contribute clarification to a terminologically confused domain.

As just indicated, metarepresenting requires using concepts such as self-concept, and reference to the *semantic* components of the first-order representation. This is what dual-process theorists call "conceptual" metacognition (see Proust, 2017). The *word* "metacognition" does not make it *factually or demonstrably* second-order, anymore than the terminological distinction between "procedural metacognition" and "conceptual metacognition" holds as an empirical validation of this contrast.

A functional, non-question-begging *definition* of metacognition states that it consists in a set of abilities allowing an individual to control and monitor his/her own cognitive activity (Nelson & Narens,

---

[1] As observed by Richard Carter (personal communication), "meta-" in ancient Greek means "after," "over," "above," "between," not: "*about.*" The suggestion, then, does not strictly speaking stem from the etymology of "meta." When the *word* emerged in publications by Flavell (who used the terms "metamory" and "metacognition" in his 1979 article), the domain covered by these terms was considered unclear. Robert Bjork reports his own impressions as follows: "I remember thinking that early definitions of metacognition and metamemory, such as "thinking about thinking" or "knowing about knowing" referred to philosophical matters that might make for an interesting discussion at a cocktail party but not rigorous or productive research – so that, too, may have been a factor in my not seeing the importance of understanding metacognitive processes" (Bjork, 2016, p. 1).

[2] On the history of the field called "metacognition." See Dunlosky & Metcalfe (2009), p. 9-34.

[3] Tarski (1936) showed that truth relative to a formal language cannot be defined within that language itself, but within a *metalanguage*, describing the symbols and rules of the first-order language.

1992. Substantial evidence has been collected in the last forty years indicating that conceptual metacognition does not exhaust the mechanisms of control and monitoring that constitute metacognition. Experiments performed by Asher Koriat and his group have shown that the mechanisms subserving metacognitive evaluations of the most ordinary sort, such as evaluating one's ability to remember an item, although they bear on first-order states, *do not refer to them*. Control and monitoring of one's cognition can be conducted without any reference to content, and in fact depend mostly on "vehicle information" (in Koriat's terms) – e.g., the number of alternative representations activated by the question handled, or the temporal features of the activity (onset delay, total duration, etc.) (e.g., Koriat, 1993, Koriat & Ackerman, 2010). These dimensions are captured by the notion of processing *fluency*, applied to various sections of the cognitive activity being monitored (discussed further below).

Evidence such as this, in combination with comparative data, have led theorists to hypothesize that nonhumans and very young children can form metacognitive evaluations even though they are not able to verbally report their uncertainty, or "think" about it as human adults do. A main tenet of dual-process theorists, then, is that there are two forms of metacognition: procedural metacognition and conceptual metacognition (Koriat & Levy-Sadot, 1999; Shea et al., 2014). While the first is based on implicit associative heuristics, the other is based on explicit symbolic representations. This kind of distinction extends beyond metacognition to a number of fields such as human reasoning, social understanding, and instrumental decision-making. Cognitive agents can rely on either low-effort, or effortful processing, which are also labeled "associative" or "system 1" processes and "rule-based" or "system 2" processes (see Chaiken & Trope, 1999). In the case of metacognition, the processes involved in each system differ along a number of dimensions, most notably in relation to informational sources, type of awareness, availability for verbal report, and action guidance (Proust, 2013, 2017).

A Short Glossary

In order to clarify the contrast between the two kinds of theories of metacognition currently on offer, readers may find it useful, at this point, to have a glossary of the main concepts that have been introduced in dual-process theories.

- *Procedural metacognition* and *implicit metacognition* are terms widely used in psychological and philosophical studies of metacognition. Both terms refer to the set of processes allowing an agent to control and monitor a first order cognitive action without representing it conceptually. Metacognition is "procedural" when it results from nonconceptual predictive processes. It is "implicit" to the extent that these predictive processes are selected independently of conscious awareness, even though they generate conscious feelings.
- *Metacognitive feelings* have been discussed in detail in the philosophical and empirical literature, by Koriat (1993), Schwarz and Clore (1996), Schwartz, Benjamin, and Bjork, (1997) and many others. The term refers to the affects produced as a result of engaging in a cognitive task. It has been shown that these affects provide internal feedback on the cognitive activity, which in turn guides decision making, i.e., metacognitive control.
- *Metacognitive information* refers to the information (sets of cues, heuristics) that triggers a given feeling or a given judgment (Schwarz, 2011).
- *Metacognitive sensitivity* refers to the calibration of the monitoring mechanisms (Pieschl, 2009). Monitoring one's confidence relies on tracking reliably over time the degree to which one's predictive feelings are corroborated by the rate of success.
- *Conceptual metacognition* refers to the kind of self-evaluation that depends on judgments based on beliefs and theories about the self, the task, and the competences that it involves or seems to involve.

Dissociating Procedural from Declarative Metacognition in Children

Some developmental psychologists, in the last decade, have started to take inspiration from animal studies to explore young children's metacognition from a procedural angle (Balcomb & Gerken, 2008; Bernard, Proust & Clément, 2014; Goupil & Kouider, 2016; Kim, Paulus, Sodian, & Proust, 2016; Paulus, Proust, & Sodian, 2013; Vo, Li, Kornell, Pouget, & Cantlon, 2014). Evidence for procedural metacognition in non-human primates suggests that human children should similarly be able to monitor their confidence level in a given perceptual or memory task, and to control their decision on this basis – independently of their ability to theorize about their own minds. Researchers have reasoned, in addition, that non-verbal children need mechanisms for selecting what to learn, and from whom – discriminating novel from familiar items, assessing what they remember, and selecting reliable informers. Implicit forms of human metacognition, then, have been hypothesized to precede children's ability to attribute mental states to themselves.

Developmental Studies of Procedural Metacognition: A Short Survey

Evidence for procedural metacognition in very young children was first collected by Frances Balcomb and LouAnn Gerken (Balcomb & Gerken, 2008). They presented to young children a non-verbal opting-out task replicating the memory-monitoring paradigm first used by Shields (1999) and Smith et al. (2003). Three-year old children (mean age: 3 years and 6 months) learned a set of paired associates, and were given a recognition memory test, with an option to skip uncertain trials. The authors found that accuracy for accepted items was higher than for skipped items (as shown by a subsequent forced-choice recognition test). This finding indicates that the children used metacognitive information to flexibly adjust their decision to item difficulty.

How did the authors explain their own findings? They hypothesised that children might use affective markers discriminating items that they could/could not remember.[4] A "threshold account" was used to explain why easy items may lead to accepting the memory test. This type of hypothesis had already been proposed by Zajonc (1968). Repeated exposure to a stimulus changes both its speed of processing and its pleasurability – what is now called its "processing fluency." Zajonc speculated that a threshold comparator mechanism takes processing properties as input, and elicits positive or negative feelings as output. This mechanism, then, is taken to explain how agents "know [recognize] what they know," even before they fully understand what knowledge is.

This type of account has been documented in adult metacognitive studies, following pioneering research by Koriat (1993) and Schwarz et al. (1991). It assumes that specific fluency-dependent feelings, such as feelings of knowing, feelings of being right, feelings of uncertainty, etc., serve as an interface between non-conscious predictive cues, on the one hand, and a conscious decision such as accepting or rejecting a task proposal or a task outcome, on the other hand. Koriat (2000) generalized it in terms of what he calls "the cross-over principle": metacognitive feelings are conscious affective signals that are generated by non-conscious heuristics. Their being both conscious and affective signals explains why they are able to flexibly influence decision-making. Feelings, however, tell an agent what to do independently of any capacity to express verbally her own uncertainty. Being a pioneer of dual-process theories of metacognition, Koriat, then, does *not* defend a "declarative" view of metacognition, as is occasionally reported (Smith, Beran, Couchman, Coutinho, & Boomer, 2009). According to him, affective experience is a main source of information that guides cognitive decisions in procedural metacognition. Concept-based beliefs and theories about one's own mental abilities are a second source of

---

[4] A description of Balcomb and Gerken (2008) and Goupil and Kouider (2016) has been offered by Peter Carruthers (2019) to the effect that the authors defend a view of metacognition in terms of *awareness* of their own ignorance" and "wanting to know." This second-order description is misleading, because these authors interpret children's performances in terms of an affective theory. The divergence actually consists in the way in which the word “metacognition” is defined, not in the underlying theory. More on this below.

decision-making, used by humans both to attribute knowledge to self and others, and to interpret their metacognitive feelings.

Thus the causal role of feelings in so-called "procedural" metacognition is well documented. There is now massive evidence that feelings of uncertainty are generated subpersonally by non-conscious predictive heuristics, which are in turn selected by reinforcement learning. Granting the central importance of this view in comparative metacognition, it may be worth analysing reinforcement more closely.

Two Kinds of Reinforcement Learning

The reinforcement learning principle posits that biological agents continuously adapt their behavior based on the consequences of their actions (Sutton & Barto, 1998). The consequences of an individual's actions can be predicted, however, not only through external feedback (delivered by the environment), but also through internal feedback (delivered by the activity itself). In a model of *internal reinforcement learning*, confidence prediction errors serve as signals for a mismatch between the current level of confidence and a running average of previous confidence-outcome pairings (expected confidence) (Daniel & Pollmann, 2012). The processing cues that have been associated in past activity with later success are automatically memorized as a function of their specific statistic reliability. It is currently proposed that the significant neural correlates of confidence prediction errors involve mesolimbic brain areas such as the ventral striatum and the ventral tegmental area (Daniel & Pollmann, 2012; Guggenmos, Wilbertz, Hebart, & Sterzer, 2016).

In internal forms of reinforcement, processing cues can be characterized as "mental," as "private," or as "neural" The qualification of "neural" originates in the fact that they are derived from processing information, not from content analysis. The qualification of mental is justified to the extent that discrepancy between expected and observed internal feedback is the informational source of so-called metacognitive feelings (feelings of familiarity, feelings of knowing, feelings of correctness, etc.), which in turn guide decision-making.

In summary: Experiments in developmental studies have been designed in the framework offered by the affective theory of metacognition in the first decade of this millenium. This form of metacognition is taken to be common to humans and non-humans. It is "experience-based," in the sense that non-conscious heuristics guide decision-making through graded metacognitive feelings (for a detailed defense, see Proust, 2013).

If this hypothesis is correct, there should be functional dissociations, in human adults and children, between evaluating what to do – whether prospectively (opt out, accept a task) or retrospectively (wager) – and attributing to oneself a state of knowledge. A few studies were conducted in following years with the goal of demonstrating the existence of systematic dissociations between procedural and declarative metacognition in young children. Because such dissociations buttress the case for implicit animal metacognition, it is worth summarizing the relevant data.

Dissociations Between Metacognitive Evaluations in Children

As indicated earlier, there is still a debate among developmental psychologists about the role of self-attribution in metacognition. In an experiment conducted on 3-, 4-, and 5-year-olds, Lyons and Ghetti (2013) found *parallel* patterns of opting-out responses and verbal confidence reports, consistent with their own view that metacognition requires self-attribution. In this particular study, however, the authors had introduced a pre-test phase in which the children were taught how to associate their feelings of uncertainty with appropriate verbal reports. This prompted Bernard et al. (2014) to compare the pattern of opt-out strategies in 3-year-olds in a visual discrimination task with participants' responses in false belief tasks, when no prior training was included. It was found that children as young as 3 years of age were able to adaptively accept or skip a trial. No correlations were found between children's procedural metacognition and their performances in false belief tests.

Another study directly compared implicit confidence responses and explicit confidence reports (Paulus et al., 2013) in the same trials. It was hypothesized that fixation time on a confidence scale of five smileys would express implicit uncertainty monitoring in 3.5-year-old children. In a learning phase, children had to learn animal-object pairs. In the test phase, children had to perform a recognition task and choose the correct associate for a given target among four possible answers. Children's explicit confidence judgments and their fixation time allocation on the confidence scale were collected in each trial. A strong contrast was found between implicit monitoring and explicit report. Explicit confidence judgments were unreliable – they did not differ for remembered as compared to non-remembered items. In contrast, children's fixation patterns on the confidence scale correlated with their actual performance. Children looked longer at smileys expressing high confidence ratings when they had correctly remembered the associated item.

A crucial test of the independence of control-and-monitoring (procedural) metacognition and self-attribution of mental states involves evidence for metacognitive sensitivity in *non-verbal* children. This was the goal of a study by Goupil and Kouider (2016). They tested whether two core metacognitive processes, decision confidence and error monitoring, were already present in 12- and 18-month-old infants. The experimental paradigm consisted in familiarizing chidren with a situation in which the experimenter hides a toy in one of two opaque boxes; the children were subsequently asked to point to the box where they thought the toy was hidden. In the test phase, the two boxes were masked by a curtain right after the hiding for variable delays. When the curtain opened, participants were asked to point to the box they wanted to explore. Pointing, in this task, was the first-order decision.

The implicit response for decision confidence was the infant's post-decision search duration (i.e., persistence): how long is an infant willing to search for a toy at a given location? Persistence in this experiment could be measured because the hidden toy was actually unreachable – hidden in a pocket inside the box.[5]

This paradigm may be more specifically relevant to comparative studies, to the extent that it pries apart an explanation of persistence in terms of confidence and an explanation in terms of the strength of memory traces (a first-order cause of action: infants search longer when their memory about the toy location is stronger). While the memory trace account predicts that persistence varies only with delay, the metacognitive account predicts that persistence covaries with accuracy for above chance-level performance, within the limits of the memory available for the task. Evidence for an *interaction* between delay and accuracy was consistent with the metacognitive account.

A similar paradigm was also used by Goupil, Romand-Monnier, and Kouider (2016) to demonstrate that 20-month old children can ask adults for help in a strategic way, in order to decline the choices which they assess as too difficult. This adaptive use of metacognition in communication contrasts, again, with the fact that a verbal report of their own mental states is unavailable to 20-month olds.

The developmental experiments summarized above have explored either opting-out responses, or retrospective confidence evaluations about a prior response, and found them in very young children. Granting that young children are implicitly sensitive to their own uncertainty, in the absence of external feedback, this sensitivity has to stem from reinforcement by internal feedback.

Animal Metacognition: Stimulus Reinforcement VS Activity-based Threshold Reinforcement

Defining Animal Metacognition

How can one explain that nonhumans and human children have similar performances in evaluating what they can remember? The response is: by defining metacognition in an empirically adequate, but not theory-laden way. We don't need to know what animals "think about," we only need to

[5] Post-decision persistence had been found to covary with confidence in rats by Kepecs and his group (Kepecs 2014; Kepecs et al., 2008). This study will be summarized below.

observe which trials they accept to perform, or how they wager about a cognitive outcome. As comparative psychologists have made clear, there are specific conditions that need to be fulfilled to count as involving animal metacognition. Robert Hampton (2009) offers a framework for a careful operationalization of metacognitive evidence in non-humans. It includes four rules bearing on task structure, completed by three negative constraints, which together are meant to characterize "endogenous metacognition" (i.e., a capacity of self-evaluation generated by the animal's own cognitive activity, rather than by task-specific associations available to an external observer):

Rules 1-4 on task structure:

1. There must be a primary behavior that can be scored for its *accuracy*.
2. *Variation* in performance (i.e., uncertainty about outcome) must be present.
3. A secondary behavior, whose goal is to *regulate* the primary behavior, must be elicited in the animal.
4. This secondary behavior must be shown to benefit performance in the primary task (for example, animals must decline tests that they would otherwise have failed).

Rules 5-7 on acceptable mechanisms:

5. The metacognitive responses must not be based on response competition (where perceptually presented stimuli are merely selected on the basis of their comparative attraction).
6. They must not be based on environmental cue association.
7. They must not be based on behavioral cue associations, i.e., "ancillary responses" such as hesitation, or response latency.

This set of rules above has also been applied to developmental experiments. The first four rules operationalize the contrast between metacognitive and cognitive tasks. The last three rules operationalize the contrast between external and internal forms of reinforcement. For example, response competition would involve external types of reinforcement, such as anticipated reward. A hidden "Clever Hans effect," implicit in condition 6, would also involve a form of external conditioning. Note that neural evidence enables researchers to pry apart external and internal forms of reinforcement. Response latency, for example, is recognized to be an important predictive cue for uncertainty. This signal, however, has been shown by human studies to be neural in its source, and mental – or private – in its affective correlate, rather than merely behavioral.

In conclusion, Hampton's rules have each prompted experimental tests across the multiple methods provided by interdisciplinary research. In combination, these tests offer evidence that internal feedback is at work. Let us examine, however, a first-order reductionist proposal that has been presented as an alternative to a metacognitive interpretation of animal performance.

The Stimulus Reinforcement Objection

Animal responses presented as "metacognitive" or "confidence-based" have been suspected, within comparative psychology, to be explainable by associative or external reinforcement learning mechanisms (Jozefowiez, Staddon, & Cerutti, 2009; Le Pelley, 2012;). In the latter study, a simple learning model, BEM (Behavioral Economic Model), is presented, that assumes, alternatively to a metacognitive account of the animals' decision to opt out, that behaviors associated with the higher payoff tend to be produced when the perception of the triggering stimulus is noisy. One main reason offered in favor of BEM is that, in contrast to the metacognitive account, a direct reinforcement model specifies the learning processes involved, and is able to generate a computable theory. The authors have a point, when they claim that attributing to rhesus monkeys a competence like "theory of mind" or "insight," – whose mechanisms remain highly mysterious – fails to provide a scientifically acceptable account. Denying the role of reinforcement in metacognitive learning is similarly unclear (Smith et al., 2009). In this section, an alternative way of describing the difficulty is proposed, that builds on the views presented above.

Reinforcement, in metacognition, has a prominent function. But this function does not, or not only, consist in reward maximization, but in efficiency maximization.

Reinforcement in Metacognition

In order to fully understand how reinforcement leads to deciding what to do in a metacognitive task as defined by Hampton (2009), one must explain how internal and external feedback are strategically integrated into a single decision-making process (Goldsmith & Koriat, 2008). To perform a metacognitive evaluation about feasibility, for example, an agent needs to assess task difficulty, to predict his/her own efficiency in the present context, and assess the likely outcome (reward or penalty) as a function of his/her subjective uncertainty. In other terms, an agent learns how to minimize the discrepancy between expected and observed feedback values.

1. The agent must learn over time how to adaptively assess *his global likely chances* of success in a type of task. This means that a *calibration standard* must be acquired and continuously revised (Baranski & Petrusic, 1994; Loussouarn, Gabriel, & Proust, 2011).
2. The agent must learn how to adaptively adjust, for *each specific trial,* his confidence to his predicted cognitive performance (a *resolution* standard must be acquired and continuously revised as a function of the feedback generated in former trials).
3. The agent must learn how to set up *a decision criterion* that allows him to strategically accept or decline a task or response in each specific trial – with gains *depending both on incentives and efficiency i*n a trial (Goldsmith & Koriat, 2008). This decision criterion can be applied either in decisions to accept or reject the task (predictive metacognition, tested in free-choice, opt-out paradigms versus forced-choice paradigms), or in decisions for wagering (i.e., accepting or rejecting a trial outcome as tested in post-decisional paradigms).

Note that these forms of learning do not coincide with *stimulus* reinforcement (which only partly contributes to step 3). What is reinforced in steps 1 and 2 is the standard predicting success for the informational dimensions relevant to metacognition (calibration and resolution). Integrating these two steps with strategic considerations such as expected reward and potential penalty requires a form of more complex reinforcement learning balancing efficiency, total gains, and potential losses.

Are these various steps in assessing an action value with respect to a standard relevant to animal studies of metacognition? We can easily see that they are, and have been explored in the comparative literature.

Step 1 (confidence in a given ability) determines task-engagement. Nonhuman animals need to practically assess their competence level in a given type of task such as assessing perceptual evidence for a predator, or remembering foraging sites. In comparative studies, confidence in ability is secured by prior training in various tasks.

Step 2 (reliance on resolution standards) is the core function of sensitivity to subjective uncertainty. Animals also need to attend to internal cues or feelings predicting, in a given trial, their likely chances in succeeding. Resolution is measured when uncertainty is retrospectively compared, trial by trial, with accuracy. Fragility in resolution has been investigated, for example, by Smith, Coutinho, Church, and Beran, (2013).

Step 3 (integrating expected reward and subjective uncertainty in decision-making) is obviously a major constraint for efficient effort expenditure. Demonstrating how this integration occurs requires experimental paradigms that pry apart the roles of expected reward and subjective uncertainty, as requested by Jozefowiecz et al. (2009) and Le Pelley (2012). Human experimental paradigms have demonstrated that, at a fixed level of uncertainty in a memory task, a human participant will decline or accept a trial as a function of the pay-off schedule ("the goal-driven" aspect of assessment), while preserving the comparative degree of confidence for each test item across schedules ("the data driven" aspect of assessment, Koriat, Ma'ayan, & Nussinson, 2006; on this contrast, see discussion below).

Animal studies of metacognition similarly explored whether rewards modulate uncertain responses, by allowing both incentives and task difficulty to vary across trials. This was done by

Zakrzewski, Perdue, Beran, Church, and Smith (2014) in a paradigm where responses were "forced" (no opt-out allowed), but uncertainty could be measured by a decision guided by anticipated failure or success. Monkeys were trained in a sparse-dense classification task allowing animals to accumulate on-screen tokens for successful trials. Because failure caused a loss of all the accumulated tokens, a higher level of accumulation correlated with a higher risk. Monkeys could decide on each trial, after presentation of the stimulus, whether to directly classify it or to "cash out" what they had accrued before making a classification. Because cashing out did not remove a trial, this task was not an opt-out task. Cashing out, however, involved aversive time-out. Animals massively cashed out when they were uncertain in a trial. Animals' decisions, then, cannot be explained by aversive stimulus features (such as difficulty: an objection often directed at opt-out paradigms). Here the animals preferred time-out to loss of accumulated benefits. The best explanation is that, just as in human strategic step 3 metacognition, animals' decisions integrate reward and subjective uncertainty. Anticipated task difficulty is combined with risk of losing the accumulated tokens to elicit momentary cashing out.

Another angle on step 3 (strategic metacognition) is offered by a study by Beran et al. (2015), where a post-decisional paradigm was used to test subjective uncertainty in chimpanzees. The chimpanzees had a matching-to-sample computerized task to complete. The reward for correct responses was distantly provided food: animals had to leave the computer apparatus to collect it. If not reached on time, the food was lost and not recoverable. Crucially, a delayed auditory stimulus announced that food had been deposited, which made food collection still possible, but much more effortful. Animals' confidence in their prior response, in this paradigm, is manifested by how readily they will move through space to collect food at the dispenser before the auditory signal is produced. The participants were shown to flexibly adjust their early movement behavior to their own uncertainty in combination with the risk incurred at any trial of a given difficulty level, and also used the less effortful way to collect reward in association with internal feedback (their flexible anticipation of success).

These experiments make it clear that the kind of reinforcement that explains flexible decision-making must integrate two different sources of information: subjective uncertainty and reward. A similar conclusion has been drawn from studies pairing behavioral decisions and neural correlates. A classic neuroscientific study by Kepecs and Mainen (2012) used a post-decisional response to elicit spontaneous expressions of confidence in rats. The first-order task involved categorizing a dominant component in an olfactive mixture. Reward was delayed by several seconds. During this interval, rats were given the option to restart the trial by re-entering the odour sampling port. First-order and metacognitive decisions were collected in each trial. The time animals were willing to wait to re-enter the sampling port ("restarting") was an indication of their decision confidence. The response patterns spoke against the differential role of reward in persistence: granting that one and the same stimulus could elicit both correct and error choices, which were respectively associated with different uncertainty assessments, confidence responses could not be explained by *stimulus* reinforcement learning.

The neural representations of decision uncertainty offered additional arguments in favor of implicit metacognitive information being used by rats to restart or abort a new trial. 20% of the orbitofrontal cortex (OFC) neurons were found to predict decision outcome. Importantly, their activation correlated not with predicted reward, but with choice accuracy. Therefore, the signals observed in OFC neurons could not be readily explained as reward expectancy based on either a simple average stimulus–reward association – or more complex predictions based on external reinforcement history. Rats' behaviors at the port, however, reflected the predictive information displayed in their OFC neurons. The animals adaptively aborted uncertain trials. Which set of predictive cues extracted from current activity will be used depends on the kind of task under evaluation. It generally includes variations in onset delay, intensity (firing rates are higher at chance performance), and in the coherence of the neural activity in a given sequence of the cognitive task (Kiani & Shadlen, 2009). Another study has shown that frontopolar cortex activity in monkeys tracks differences between incorrect and correct trials before receipt of feedback rather than differences in reward (Tsujimoto, Genovesio, & Wise, 2010). In summary, the studies reviewed in this section are incompatible with the view that metacognitive responses are generated by first-order reinforcement – learning the reward value of an external stimulus. It is also incompatible

with the view that metacognition is "beyond reinforcement" (Couchman, Coutinho, Beran, & Smith 2010). Internal reinforcement is used to establish response thresholds on the basis of past observed accurate/inaccurate outcomes. Calibration of confidence depends on it. A predictive neural structure has accordingly been identified in the OFC, whose function is to predict success or failure in a cognitive task independently of reward (Kepecs, Uchida, Zariwala, & Mainen, 2008).

Metacognitive Awareness in Non-humans

The history of metacognitive studies sketched out above partly explains why there is still no agreement on a definition of metacognition able to serve the interests of cognitive science. Procedural notions are more appealing to theorists who are sensitive to the flexibility that metacognition introduces in animal and human decision-making. Declarative notions are inspired by the view that metacognition requires mindreading, or is uniquely human (Carruthers, 2008; Perner, 2012). Independent of this divergence, a major objection against the view that non-humans can be metacognitive has to do with the "awareness" issue. Assuming that metacognition consists in monitoring one's confidence level in order to control one's behavior, is the self-regulative ability demonstrated to guide decision in nonhumans comparable with self-awareness in humans? Some theorists express doubt about this issue – including those comparative scientists who pioneered studies in animal metacognition. What are the reasons that may lead to doubt that nonhumans are genuinely metacognitive (Metcalfe & Kober, 2005)?

The most skeptical among the comparative scientists emphasize the functional differences between reward-driven behavior perceived as characteristic of nonhumans, and value-driven, reflective action-planning in humans. In short, humans are persons, not only agents. They are conscious of the implications of their decisions. They are able to plan, to doubt, to reconsider, to integrate moral with epistemic and economic values. On the other hand, nonverbal animals do not have to justify to others their decisions and courses of action. This explains why they do not need to represent themselves and others as having beliefs (they can, however, implicitly represent desires and intentions). How does this lead to the position that genuine metacognition is indissociable from having self-reflective consciousness? This inference is based on two assumptions. The first is that humans have multipurpose introspective access to who they are. The second is that self-awareness must be a higher-level, "second-order" competence. These two assumptions need to be closely scrutinized.

Unity or Plurality in Self-awareness?

A commonsense view of self-awareness is that one introspects one's own thinking with one's "inner eye," to collect evidence about the person one is, and the thoughts one has. This view has inspired some scientists to take reflective consciousness to be testable in a variety of tasks and circumstances. For example, Gallup's mirror-recognition test is claimed to reveal the ability of an animal to discriminate "itself" from others; episodic memory tests are seen as exploring the capacity to "project one's own self" into the past or the future, social cognition tests as based on introspecting one's own beliefs and desires. Finally metacognition is interpreted as reflectively conscious to the extent that it leads to attributing to "oneself" one's own epistemic states (such as knowledge, doubt, etc.).

There are three main problems with an introspective view (Proust, 2017). First, a self is not a physical entity, it is *a representation,* or even, as will be elucidated, a set of incompatible representations (Metzinger, 2004). Second, the "inner eye" metaphor is a misleading intuition. Constructing and activating a context-relevant self-representation has nothing to do with contemplating internal properties; it is, rather, an inferential activity (Carruthers, 2011). There is *accordingly no common source of information* ("oneself") between the set of inferences constructed and employed across modes of self-awareness. Thirdly, each of us constructs and uses a number of self-representations, each driving decisions to act in specific contexts (Yan & Oyserman, 2018). In this process, there is no unified self in charge, and no all-encompassing self-representation available, as also pointed out by Dan Dennett (1993) in his "multiple draft" metaphor.

In summary, the tests listed above draw on very different types of self-representations, such as – among others – the ecological self, the sense of ownership, the sense of agency, and the narrative self (for a review, see Gallagher, 2000). The proposal that full-blown metacognitive awareness occurs at the presumably higher level of narrative self-representations conflicts with the "multiple draft" model of consciousness.

Is Self-awareness a Higher-level, "Second-order" Competence?

The good news for making progress in our discussion is convergent evidence that, in humans, there are forms of conscious sensitivity that do not require self-attribution. The point can be made that nonhumans have a conscious sensitivity in monitoring their cognitive actions comparable to humans. As discussed above, Koriat (1993, 2000) defends the claim that consciousness is functionally required in a flexible cognitive control system. Metacognitive feelings are generated at the junction between "implicit antecedents" – unconsciously extracted predictive heuristics – and "explicit consequences," such as deciding what to do, reporting what one remembers, etc. The "cross-over principle" states that metacognitive feelings *need* to be conscious to secure flexibility in our decisions to act.

Subsequent research in neuroscience offers a related argument. Consciousness is required for integrating all the values of action in a given situation (see step 3 in “Reinforcement in Metacognition” above). Feelings provide "a common currency" in the process of decision-making (Sugrue, Corrado, & Newsome, 2005). On this view, metacognitive feelings are subjectively experienced by humans as an affective relation with a current cognitive goal ("this is difficult," "this looks correct"), rather than merely as a property of the self ("I am really good at that").

These functional constraints should apply to nonhumans for at least three reasons. First, as far as metaperception and metamemory are concerned, nonhumans’ metacognitive performances are functionally similar to those of humans. Second, their integration-behavior suggests the application of a common currency principle. As shown earlier, step 3 assessments require, in real-world situations, the integration of expected benefits, risks, and confidence associated with various task dimensions. Capuchins have been found to be able to integrate values in economic decision-making (Padoa-Schioppa, Jandolo, & Visalberghi, 2006). Rhesus monkeys and chimpanzees have also been found to be able to integrate metacognitive value (relative certainty) and economic value (reward) as reviewed above p. 187 q(Beran et al., 2015). It is arguable that conscious feelings are needed to mediate such integration, in order to provide a common currency among various values (Proust, 2017). Thirdly, there is a striking anatomical homology (across primates) or analogy (in rodents) in the neural structures subserving metacognition. The neural correlates in rats' metacognition suggest that they control their cognitive behavior with the same orbito-frontal areas that are used by humans. It would be unparsimonious, then, to hypothesize that only humans use their conscious experience of the task to guide their decision-making. There are more specific arguments in favour of this hypothesis, however, drawing on the evolution of cognition.

The Evolution of Consciousness

It is highly consensual in this area of research that "emotion is an evolutionary extension of homeostasis, that cognition is an extension of emotion, and that the brain is organized to achieve the seamless integration of homeostasis, emotion, and cognition." (Watt, 2005, p. 85). In this picture, a primitive form of consciousness is present as soon as emotions have appeared, in evolution, to adaptively predict challenges to homeostasis. Homeostasis, as exemplified in the self-regulation of temperature, relies on non-conscious processes. Conscious emotions were selected with the function of predicting (and adaptively responding to) prototypical homeostatic challenges. Conscious hunger, conscious pain, and conscious thirst, were primary conscious states with the function of optimizing homeostasis and preparing adequate response programs. Further emotions developed to enhance control in other areas. On this view, then, reinforcement should not be dissociated from conscious affects, such as pleasure or fear (Panksepp,

2005).[6] Numerous anatomical arguments through homology (within mammals) and analogy (between mammals and birds) have been offered, that cannot be summarized here (see Baars, 2005; Denton, McKinley, Farrell & Egan, 2009; Merker 2005; Watt, 2005).

Subjectivity

The following objection could be raised against our continuist proposal about consciousness. Do not animals use the emotions they have in *behaving* adaptively, rather than in *experiencing* them as something that *subjectively matters*? Responses to this question are emerging from current neuroscientific research. The first part of the answer consists in establishing the neural correlates and informational source of subjectivity in humans, the second part in generalizing these findings to nonhumans.

In humans, a set of neural mechanisms has been hypothesized to tag selfness either *implicitly* in the conscious experience "I" have, or *explicitly*, in thoughts "about me" (Park & Tallon-Baudry, 2014). Note that this contrast (I/me) overlaps with the contrast between experience-based and concept-based metacognition. In both cases, a "subjective frame" is used to qualify the self-relevance of incoming or reafferent information (i.e., conscious aspects of subjectivity). The current status of the subjective frame originates in the viscera, including heart and guts. There is evidence that heartbeat-evoked responses in the ventral precuneus covary with the self-relatedness of spontaneous thoughts (experiential self). Heartbeat-evoked responses in the ventromedial Prefrontal Cortex, in contrast, covary with explicit representations of oneself (Babo-Rebelo, Richter, & Tallon-Baudry, 2016).

It is currently hypothesized that physiological and cognitive functions contributing to the subjective frame are both served and integrated by the so-called "default-mode network" (DMN) that is engaged when organisms *do not* pursue external goals (Raichle, 2015). It increases activity in passive states relative to active periods of engagement with the environment. It is hypothesized that it performs "internally-directed cognition," and might help connect autonomic and cognitive regulation. It has been shown to include the anterior and posterior cingulate cortex, the medial and lateral parietal cortex, the medial prefrontal cortex, and distributed activity in the heteromodal areas. Again, subjective consciousness might here play the role predicted by the "cross-over principle" – from heartbeats and gut changes to subjective feelings of confidence, of safety, made conscious through distributed somatosensory cues, as already proposed by Damasio (1996).

Relevant to our present discussion is the finding that nonhuman brains present neural structures analogous to those that constitute the human DMN (Mantini et al., 2011). In monkeys, the DMN activates a set of interconnected subsystems that converge on hubs including the posterior cingulate cortex (PCC) and regions within prefrontal cortex (PFC). It is speculated that monkeys might engage in forms of spontaneous cognition detached from the external environment during idle moments. This speculation is based on "the observation that the anatomy of the DMN in monkeys includes heteromodal association areas and not sensory regions." A structure similar to the DMN has also been found in rats. It is hypothesized that the function of this network is to integrate affective, subjective information with multimodal sensory inputs to adjust behavior (Lu et al., 2012). These findings suggest that a subjective experience, in nonhumans, is manifested in neural correlates even when no external goal is presently pursued. The presence of affective, subjective information at rest, however, does not restrict the functional role of subjectivity to passive episodes. Available evidence suggests rather that nonhumans also have a conscious sense of the self-relevance of various opportunities, risks, and courses of action.

A further consequence to be drawn from interdisciplinary research is that consciousness does not reduce to declarative self-awareness ("thoughts about me"). Basic consciousness rather consists in "I thoughts." It is exemplified by metacognitive sensitivity in human infants and in nonhumans. One of its main functions is to integrate the values of opportunities and risks in assessing the environment. Such

[6] For a dissenting view, see LeDoux and Brown (2017).

integration in turn renders decision-making more flexible and adaptive. It might also plausibly play an important role in learning and recall.

In summary, some interdisciplinary research converges on the view that consciousness does not reduce to declarative self-awareness ("thoughts about me"). Basic consciousness rather consists in "I thoughts." It is exemplified by metacognitive sensitivity in human infants and in nonhumans. One of its main functions is to integrate the values of opportunities and risks in assessing the environment. Such integration in turn makes decision-making more flexible and adaptive. It might also plausibly play an important role in learning and recall.

Objections and Alternative Proposal: Basic Questioning (BQ)

A Philosopher's Objection

In a series of studies, the philosopher Peter Carruthers has offered a detailed criticism of procedural metacognition (see in particular, Carruthers, 2008, 2009, 2011). These publications question the validity of an interpretation of animal uncertainty responses as expressions of *second-order* metacognition, using arguments similar to those discussed above. Carruthers (2017, 2018, 2019) offers a positive theory of non-conceptual epistemic emotions, such as the feeling of uncertainty. Against the view that these emotions are *metacognitive* feelings, as defended in metacognitive studies (which, as we have seen, do not restrict metacognition to its second-order theorizing), it is claimed they are "first-order" phenomena related to a preverbal attitude of "questioning the world" available to non-humans as well as to humans. Given the limits of the present article, we will concentrate on Carruthers' positive proposal.

In Carruthers (2017, 2018), feelings of surprise, interest, curiosity, and uncertainty are claimed to be present in non-humans. In contrast to the metacognitive literature, however, Carruthers' proposal is that epistemic emotions *do not* have a monitoring function: they do not tell agents anything about the precision or availability of their curent perception or memory. Concerning the claim that feelings do not *metarepresent* cognitive states, everybody should agree with Carruthers: metarepresentation is a representation referring to first-order representational *content* (see above), and feelings do not have this function. But Carruthers claims in addition that they do not monitor cognitive efficiency, that their function is rather strictly "first-order." They track external affordances, and depend on external reinforcement. Let's see how these claims are articulated and defended.

On Carruthers' view, the attitude called "basic questioning" (BQ) or "curiosity" is an affective state that is desire-like. Its function is to motivate an agent to reach a specific cognitive state, such as knowing what that bird is. Questioning your guide in order to identify a bird does not require metarepresenting your guide as having knowledge. It is merely a direct way for you to obtain information. There is no questioning behavior in nonhumans, but the author observes that nonhumans also are occasionally curious. For example, starving bees are curious to know where nectar is available. How is this possible? Animal curiosity is hypothesized to consist in a preverbal affective attitude *whose content is a question* (Carruthers, 2018, p. 7). Carruthers' hypothesis is that nonhumans, just as humans, can wonder about something being the case. This basic form of questioning does not involve metacognition (i.e., second-order metacognition, as Carruthers understands the term), because curiosity can be "de re" rather than "de dicto." This means that animals can aim to get knowledge without using or having the concept of knowledge. Prelinguistic attitudes of questioning (such as feelings of curiosity, of uncertainty, of surprise) are just as basic as are affective states like fear. In nonhumans as in humans, curiosity is claimed to activate a set of investigative behaviors with no reliance on concepts related to knowledge acquisition. "Curiosity, like other affective attitudes such as fear and anger, is apt to motivate directly (without any need for executive decision making) forms of action that are designed to alleviate the affective state in question (that is, to extinguish curiosity)." (Carruthers, 2019, p. 7).

There is much in this view that procedural theorists can assent to. Just as is claimed by the procedural theory of metacognition, informational availability seems to be what basic questioning is all about: determining "de re" "cognitive affordances" that can be relied upon, in a given context, for the

pursuit of further goals, and selecting an action that makes them available (see Proust, 2013, 2015a, 2016). To many theorists, the notion of an epistemic emotion appears to be mainly a terminological alternative to the notion of a metacognitive feeling. Another point of convergence is that epistemic emotions have, as their function, to elicit mental actions, such as holding attention to a cognitive target (Carruthers, 2017). The divergence between this view and theorizing about procedural metacognition is that the latter takes *metacognition* to control cognitive actions of this kind and to monitor feedback from them. Carruthers's view is that *first-order emotions* control cognitive actions, but do not monitor them. This view is procedural because affective information elicits basic questioning. It is *non-metacognitive* because curiosity is claimed *not to* involve internal monitoring of one's informational states. The influence of affect on behavior is, instead, directly effected through external reinforcement.

In Carruthers and Ritchie (2012), the point was made in terms of subjective emotions: "What is it that one feels bad about, when one feels uncertain? Is it about the likelihood of a cognitive perceptual or memory task being correct/incorrect? Is it not, rather, about the likelihood of a reward being missed/obtained?" However, Carruthers (2019) admits that "curiosity-satisfaction *is* directly rewarding in animals (and hence presumably in human infants likewise)." Based on Blanchard, Hayden, & Bromberg-Martin (2015), Bromberg-Martin and Hikosaka (2009), and Gipson, Alessandri, Miller, and Zentall (2009), it is now claimed that "animals will choose an option that reliably signals whether or not a food-reward is coming a few seconds later, even though this choice has no impact on the likelihood of the reward, and even though the animal knows that selecting the informative-option will reduce the size of the eventual reward, if it comes" (Gipson et al. p. 9). Hence the restriction imposed on external reinforcement now seems to be dropped: the prospect of information is not reducible to the prospect of food.

Conclusions

It is presently clear that a curious agent seeks information for its intrinsic rather than extrinsic value (as defended by Loewenstein, 1994; Kidd & Hayden, 2015). A theoretical move away from a one-level account of valence – exclusively based on external reinforcement – is a welcome feature of Carruthers' theorizing on curiosity. A theory that fails to recognize, in non-humans, a sensitivity to informational states that is distinct from reward expectation would directly conflict with the comparative evidence reviewed above.

To summarize: the cues that modulate a decision to act cognitively have been found to be based *both* on predicted accuracy and predicted reward. Neither does monitoring one's uncertainty in remembering depend on a direct relation between working memory strength and behavior as claimed in Carruthers (2017). This is shown by Goupil and Kouider (2016), Zakrzewski et al. (2014) and Kepecs and Mainen (2012), to name just a few of the relevant studies cited previously.

In order to assess Carruthers' proposal, it may be useful to describe the consequences of the duality of reward types (incentives and information) seen from the viewpoint of procedural metacognition. Expected reward or penalty (i.e., incentives) clearly affects *the amount of effort* expended, which enhances the probability of success of the outcome. Thus, metacognitive control is *modulated* by expected external reward. To the extent that agents scale up their confidence level as a function of their own invested effort (Koriat, Ackerman, Adiv, Lockl, & Schneider, 2014; Koriat & Nussinson, 2009), expected incentives also affect subjective confidence in expected success. The role of incentives in confidence is thus said to be "goal-driven" (or "control-based"). Another dimension of confidence, however, is *based on* the actual feedback collected from the task, also called "data-driven" monitoring. You may be quite certain of succeeding in a task if you put a lot of effort into it, out of your expectation of a high extrinsic reward, but also because the internal feedback received while completing the task has predicted success.

Can this important finding be accommodated in the framework of "basic questioning"? It might be, if this attitude includes an evaluative feature in its content, such as a gradient of expected success and a gradient of feedback divergence. It is unclear, at this point, whether a BQ attitude is an evaluation, or is

rather elicited by an evaluation.[7] If, as is claimed in Carruthers (2018), curiosity is triggered by the recognition of an informational gap, is this gap assessed (and when)? On the basis of what standard? Does a curious animal merely detect its own ignorance in polar terms (knowing/ignoring), or does it monitor the level of its uncertainty against a given threshold, as proposed in the empirical literature (Gottlieb, Hayden, & Bromberg-Martin, 2013). What is the valence of curiosity as an affective state: aversive, as proposed by reduction accounts, or pleasurable, as proposed by induction accounts?

These are important issues that the BQ model might address in the future. The mechanisms allowing interrogative control to be triggered, monitored, and recalibrated over time need be articulated. It may well be that this extension will turn curiosity into a monitoring state of the "procedural metacognitive" variety (ignoring for now the terminological issue).

The Objection of Metacognitive Ubiquity

A main reason that Peter Carruthers offers for rejecting a metacognitive-procedural interpretation of BQ attitudes is that it would trivialize the value and evolutionary significance of epistemic emotions (Carruthers, 2017, p. 64).[8] Carruthers' worry is that if predictive control and monitoring through a forward model qualifies as implicit metacognition, then *all kinds of cognitive subsystems qualify as metacognitive*, such as the motor system, the subsystem integrating plurimodal perception and the system of emotions. Is this true?

An uncontroversial functional definition of the word *metacognition* is that it refers to the set of abilities allowing an individual to control and monitor *his/her own cognitive activity* (Nelson & Narens, 1992. This definition makes it clear that metacognition 1) is a *person-level* control and 2) that this control has to apply to decisions associated with *cognitive* activities, such as remembering and learning, seeking information, etc. Later research has successively extended the realm of cognitive activities apt to be controlled to perceptual discrimination (Levin, 2004), language learning (Anderson, 2012), problem solving (Thompson & Johnson, 2014), mathematical calculus (Reder & Ritter, 1992) and artistic production (Schwartz, 2017).

This definition of metacognition shows that the trivialization worry does not resist scrutiny. What distinguishes procedural metacognition from automatic monitoring (as it occurs in perceptual and motor activity) is that the nonconscious heuristics relevant to metacognitive predictions do not blindly drive decision-making: they elicit conscious feelings that agents can use (or decide not to use) to select and revise their cognitive actions (see above discussions of the cross-over principle, and on conscious awaress of subjective sensitivity).

In contrast, the kind of feedback used to integrate multimodal perceptual inputs as a function of their respective informational quality is altogether *subpersonal*. Automatic mechanisms such as this have the evolutionary function of securing swift, low-cost informational efficiency. As we have seen, metacognition improves efficiency at a cost. It tends to step in when an agent is confronted with unexpected cognitive obstacles or processing difficulties (Proust, 2015b).

Motor control and monitoring differ from multimodal perception because the latter involves a rich and diverse conscious phenomenology, which overlaps with metacognition in certain respects (for example the pleasurability of fluency; Pacherie, 2008). This phenomenology, however, does not generally serve informational goals. Leaving aside the case of dancers, or actors who need to predict and assess their accuracy, agents (particularly nonhumans and young children) are not interested in the *information*

---

[7] For defense of the contrast between an evaluative and a declarative attitude, see Gawronski & Bodenhausen, (2006), Proust (2015a).

[8] The objection is summarized as follows in Carruthers, 2018, footnote 6 p. 10: "It might be granted that curiosity (and questioning attitudes more generally) need not be explicitly metacognitive. One does not have to represent one's ignorance as such in order to be curious. But it might be said that curiosity is implicitly metacognitive, nevertheless (Proust, 2013). This is because it requires agents to monitor their own states of knowledge, detecting and responding appropriately to a state of ignorance. *You can describe this as a form of metacognition if you like, but it completely trivializes the notion"* (emphasis is mine).

that their sensorimotor activity either requires or provides to themselves and others. Agents (generally) don't ask themselves: "do I know enough of the situation in order to jump?" They ask rather: "can I jump?" Obviously, however, cognitive acts are often needed while planning some motor activity; to that extent, metacognition can work as a precondition of motor activity. But these articulations are an indication that sensitivity to information serves extrinsic goals, which is an important function of metacognition.

In summary, Peter Carruthers attempted to interpret comparative evidence for uncertainty responses in nonhumans, and for sensitivity to information accuracy in young children in terms of a specific attitude labeled "basic questioning." He discarded a metacognitive interpretation both as a terminological decision for what counts as metacognitive, and a rejection of the trivialization based on a liberal criterion of internal prediction. This discussion rebuts the latter charge. The causal involvement of forward models is not sufficient to turn a given evaluation or prediction into a metacognitive one. In consequence, the trivialization objection does not apply. Granting this response, the attitude of "basic questioning," when fully articulated with its informational source, might be very close to the phenomenon studied under the label of procedural metacognition: aside from a terminological preference, a metacognitive account can in principle do the same job than basic questioning does, except that it has much broader explanatory scope.

Summary

Interdisciplinary research on metacognition has been hampered by a lack of convergence on terminology. Conflating a definition with a causal explanation has generated serious confusions. In comparative research, the keyword "uncertainty response" is associated with a variety of theories, whose interest is in part obscured by variation in the underlying definitions. Similarly, in developmental studies, the keyword "knowing what one knows" is used either as an inclusive cue to a phenomenon of common interest, or as an exclusive definition inapplicable to non-conceptual forms of knowledge sensitivity. These confusions have made it more difficult, in recent years, to state the hypotheses being defended in terms acceptable to the community at large.

The recent interdisciplinary research reviewed above suggests that, in spite of conflicting definitions, the consensus on the nature of the phenomenon under study is growing. Main points of convergence include the role of internal reinforcement learning and the dissociation in children's development between implicit metacognitive evaluation and explicit ability to report. Points of divergence, however, concern the difference between a "first-order" account and a procedural account of metacognition. Here, too, one can suspect that terminology is a source of misunderstanding. If "first-order" qualifies a direct relation between decision to act and outcome, some will see internal feedback as "first-order," while others will see it as "second-order," arguing that this feedback in fact refers to what the agent knows he knows. Here again, definition and theorizing are conflated. The most acute divergence, however, will be raised by the hypothesis that metacognition requires a form of affective awareness. Should this form of awareness be restricted to organisms having access to second-order metacognition? Evidence reviewed in the section Metacognitive Awareness in Non-humans suggests a negative answer. Conscious experience might be a precondition for integrating several affects in prediction or evaluation of an outcome. Comparative and neuroscientific research suggests that some nonhumans also deploy off-line subjective consciousness as we do. A next important step in metacognitive research will be to determine whether or not non-humans have a conscious experience of uncertainty just as humans do.

Acknowledgements

All my thanks to Mike Beran, Richard Carter, Louise Goupil, Catherine Tallon-Baudry and an anonymous reviewer for their critical observations and suggestions on a former draft. Special thanks to Richard Carter for his linguistic revision of the text.

References

Anderson, N. J. (2012). Metacognition: Awareness of language learning. In *Psychology for language learning* (pp. 169–187). London: Palgrave Macmillan.

Baars, B. J. (2005). Subjective experience is probably not limited to humans: The evidence from neurobiology and behavior. *Consciousness and Cognition*, *14*, 7–21.

Babo-Rebelo, M., Richter, C. G., & Tallon-Baudry, C. (2016). Neural responses to heartbeats in the default network encode the self in spontaneous thoughts. *Journal of Neuroscience*, *36*, 7829–7840.

Balcomb, F. K., & Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, *11*, 750–760.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*, 412–428.

Beran, M. J., Perdue, B. M., Futch, S. E., Smith, J. D., Evans, T. A., & Parrish, A. E. (2015). Go when you know: Chimpanzees' confidence movements reflect their responses in a computerized memory task. *Cognition*, *142*, 236–246.

Bernard, S., Proust, J., & Clément, F. (2014). The medium helps the message: Early sensitivity to auditory fluency in children's endorsement of statements. *Frontiers in Psychology*, 5, 1412.

Bjork, R. (2016). Prologue: Some metacomments on metamemory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 1–3). Oxford: Oxford University Press.

Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron*, *85*, 602–614.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*, 119–126.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language*, 23, 58–89.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*, 121–138.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Carruthers, P. (2017). Are epistemic emotions metacognitive? *Philosophical Psychology*, *30*, 58–78.

Carruthers, P. (2018). Basic questions. *Mind & Language*, *33*, 130–147.

Carruthers, P. (2019). Questions in development. Unpublished manuscript.

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

Chaiken, S. & Trope Y. (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1413–1420.

Daniel, R., & Pollmann, S. (2012). Striatal activations signal prediction errors on confidence in the absence of external feedback. *Neuroimage*, *59*, 3457–3467.

Dennett, D. C. (1993). *Consciousness explained*. Boston: Little, Brown and Co.

Denton, D. A., McKinley, M. J., Farrell, M., & Egan, G. F. (2009). The role of primordial emotions in the evolutionary origin of consciousness. *Consciousness and Cognition*, *18*, 500–514.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage Publications.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development, 60* (1, Serial No. 243).

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, *4*, 14–21.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731.

Gipson, C. D., Alessandri, J. J., Miller, H. C., & Zentall, T. R. (2009). Preference for 50% reinforcement over 75% reinforcement by pigeons. *Learning & Behavior*, *37*, 289–298.

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* (pp. 1–60). London: Academic Press.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*, 585–593.

Goupil, L., & Kouider, S. (2016). Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Current Biology*, *26*, 3038–3045.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences*, *113*, 3492–3496.

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife*, *5*, e13388.

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

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

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*, 208-216.

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

Jozefowiez J., Staddon J. E. R., & Cerutti D. T. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition & Behavior Reviews, 4,* 29–39.

Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron, 84,* 190–201.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1322–1337.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455,* 227–231.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324,* 759–764.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*, 449–460.

Kim, S., Paulus, M., Sodian, B., & Proust, J. (2016). Young children's sensitivity to their own ignorance in informing others. *PloS One*, *11*, e0152595.

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

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*, 149–171.

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition*, *19*, 251–264.

Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General*, *143*, 386–403.

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 483–502). New York: Guilford Press.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relation-ships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135, 36–69.

Koriat, A., & Nussinson, R. (2009). Attributing study effort to data-driven and goal-driven effects: Implications for metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1338-1343.

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

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*, E2016–E2025.

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 686–708.

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

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development*, *26*, 327–333.

Loussouarn, A., Gabriel, D., & Proust, J. (2011). Exploring the informational sources of metaperception: The case of change blindness blindness. *Consciousness and Cognition*, *20*, 1489–1501.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*, 75–98.

Lu, H., Zou, Q., Gu, H., Raichle, M. E., Stein, E. A., & Yang, Y. (2012). Rat brains also have a default mode network. *Proceedings of the National Academy of Sciences*, *109*, 3979–3984.

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Development*, *84*, 726–736.

Mantini, D., Gerits, A., Nelissen, K., Durand, J. B., Joly, O., …Vanduffel, W. (2011). Default mode of brain function in monkeys. *Journal of Neuroscience*, *31*, 12954–12962.

Merker, B. (2005). The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, *14*, 89-114.

Metcalfe, J., & Kober, H. (2005). Self-reflective consciousness and the projectable self. In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition: Origins of self-reflective consciousness* (pp. 57–83). New York: Oxford University Press.

Metzinger, T. (2004). *Being no-one. The self-model theory of subjectivity*. Cambridge, MA: Bradford Books, MIT Press.

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

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179-217.

Padoa-Schioppa, C., Jandolo, L., & Visalberghi, E. (2006). Multi-stage mental process for economic choice in capuchins. *Cognition*, *99*, 389–395.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, *14*, 30–80.

Park, H. D., & Tallon-Baudry, C. (2014). The neural subjective frame: From bodily signals to perceptual consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130208.

Paulus, M., Proust, J. and Sodian, B. (2013). Examining implicit metacognition in 3.5-year-old children: An eye-tracking and pupillometric study. *Frontiers in Psychology*, 4, 145.

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

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, *4*, 3–31.

Proust, J. (2013). *The philosophy of metacognition. Mental agency and self-awareness.* Oxford: Oxford University Press.

Proust, J. (2015a). The representational structure of feelings. In T. Metzinger & J. M. Windt (Eds.), *OpenMind* (Article 25). Frankfurt am Main.

Proust, J. (2015b). Time and action: Impulsivity, habit, strategy? *Review of Philosophy and Psychology, 6*, 717–743.

Proust, J. (2016). The evolution of communication and metacommunication in primates. *Mind and Language, 31,* 177–203.

Proust, J. (2017). Non-human metacognition. In K. Andrews & J. Beck (Eds.), *Routledge handbook of philosophy of animal minds* (pp. 142-153). New York: Routledge.

Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, *38*, 433–447.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 435–451.

Schwarz, N. (2011). Feelings-as-information theory. *Handbook of Theories of Social Psychology*, *1*, 289–308.

Schwarz, N. (2017). Of fluency, beauty, and truth: Inferences from metacognitive experiences. In J. Proust & M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach*, pp. 25-46. New York: Oxford University Press.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*, 195–202.

Schwarz, N., & Clore, G. L. (1996). Feelings and phenomenal experiences. *Social Psychology: Handbook of Basic Principles*, *2*, 385–407.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, *6*, 132–137.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*, 186–193.

Shields, W. (1999). *Nonverbal judgments of knowing by humans and rhesus monkeys*. Buffalo, NY: State University

of New York at Buffalo.
Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V., & Boomer, J. B. (2009). Animal metacognition: Problems and prospects. *Comparative Cognition & Behavior Reviews, 4,* 40–53.
Smith, J. D., Coutinho, M. V., Church, B. A., & Beran, M. J. (2013). Executive-attentional uncertainty responses by rhesus macaques (*Macaca mulatta*). *Journal of Experimental Psychology: General*, *142*, 458–475.
Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, *26*, 317–339.
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.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Tarski, A. (1936). Der Wahrheit's Begriff in den formalisierten Sprachen, (1936). *Studia Philosophica, I,* 261–405.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*, 215–244.
Tsujimoto, S., Genovesio, A., & Wise, S. P. (2010). Evaluating self-generated decisions in frontal pole cortex ofmonkeys. *Nature Neuroscience*, *13*, 120–126.
Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young children bet on their numerical skills: Metacognition in the numerical domain. *Psychological Science*, *25*, 1712-1721.
Watt, D. F. (2005). Panksepp's common sense view of affective neuroscience is not the commonsense view in large areas of neuroscience. *Consciousness and Cognition*, *14*, 81–88.
Yan, V. X., & Oyserman, D. (2018). The culture—identity—metacognition interface. In J. Proust & M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach* (pp. 225-244). Oxford: Oxford University Press.
Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–27.
Zakrzewski, A. C., Perdue, B. M., Beran, M. J., Church, B. A., & Smith, J. D. (2014). Cashing out: The decisional flexibility of uncertainty responses in rhesus macaques (*Macaca mulatta*) and humans (*Homo sapiens*). *Journal of Experimental Psychology: Animal Learning and Cognition*, *40*, 490–501.