*Aristotelian Society, June 2, 2008*
To appear in *the Proceedings*

# Epistemic agency and metacognition: an externalist view

Joëlle Proust

Institut Jean-Nicod (EHESS, ENS), Paris

---------------

Today's epistemologists debate about the respective roles of evidence and of subjective responsibility in a definition of knowledge.[i] It is often assumed that agents can be held responsible for the way they control their processes of knowledge acquisition. An ability to control the process by which a given belief is formed has been presented as a necessary condition for an agent's being possibly justified, rather than simply entitled, to form that belief.[ii] The question of what is involved in the control of one's mental agency, however, is rarely if ever addressed. Does control of one's perception, memory, reasoning, rely on something like introspective capacities? Or does it depend on external constraints? The first aim of this article is to explore these questions. Control of one's mental agency encompasses two kinds of reflective, evaluative operations, which together constitute metacognition. *Self-probing* predicts whether one has the cognitive resources needed for the success of some specific mental task at hand. *Post-evaluating* appreciates retrospectively whether the mental mental state that is attained as a result of a mental action conforms to the norm of adequacy for that action (section I). A second aim is to examine whether recognizing the contribution of epistemic feelings to metacognitive interventions in mental agency favors an internalist type of epistemic status for self-knowledge acquisition (section II). Section III provides arguments for rejecting internalism about metacognition; it introduces a "brain-in-the-lab" thought experiment to fuel externalist intuitions about metacognition; it discusses two possible types of strategies in favor of an externalist conception of metacognitive entitlement, based respectively on evolutionary considerations and on learning. Section IV examines a generalization of the thought experiment to twin-cases, and discusses the merit of a third externalist strategy, based on dynamic-coupling properties constituting what will be called "the viability core" of a mental task.

# I

*Mental agency and metacognition* - To understand how metacognition is a necessary feature of mental agency, it is useful to start with a definition of a mental action. The following definition (from Proust, forthcoming$_a$) emphasizes the classical "trying" criterion that is characteristic of any action:

> A *willing* or a *trying* is a mental event through which an operation from the repertoire
>
> 1) is called because of its instrumental relationship to a goal, and
>
> 2) is thereby made available to executive processes.[iii]

What characterizes a mental action, in contrast with a bodily action, is that the kind of goal, and the kind of instrumental relationship to the goal, that are selected and used to guide the action are mental rather than environmental properties. For example, a bodily action such as switching on the light may be tried because one's goal is to have the room illuminated, and because switching on the light is the standard way of producing this outcome. By analogy, a mental action such as counting the number of dots in a display may be tried because one's goal is to know the number of dots in the display, and because counting them is the standard way of knowing how many dots there are. A mental trying however is not only causally determined by having a goal and by there being standard ways of producing it; it involves an operation implementing these ways that must already be in the repertoire: I must know how to map numbers to objects in order to know how many dots there are. I can only rationally try to count – that is: try to perform an action which is very likely to be successful - *if I know how* to count. This does not mean that there are not irrational tryings, like trying to bend spoons through the force of one's mind, where there is no objective way of doing so. But these tryings are normally quickly relinquished by monitoring their consistent failure.

One might try to classify the mental actions that are in one's repertoire via the types of attitudes that are being controlled. Given that, by definition, a mental action tries to produce a given mental goal, it operates by controlling mental operations that typically produce or contribute to producing this goal.[iv] It might therefore seem that one could form classes of mental actions from every psychological attitude, by merely postulating that to each 'spontaneous' species, there corresponds a 'controlled' one. But this is clearly not a correct way of creating a taxonomy, for certain kinds of attitudes, such as perceptions and beliefs, have a mind-to-world direction of fit; being essentially receptive, and thus it seems that in an important sense, their content and psychological function are, precisely, of a non-controllable kind. Some caution is therefore needed when considering how attitudes are controlled when they can be controlled. One cannot control what one believes, but one can control which types of beliefs one forms, through control of the process of belief acquisition (by selecting one's sources of information, inquiring into their reliability, etc.). The same holds for perception: one cannot directly control what one perceives when one is facing a given display with eyes open, etc.; but one can choose to close one's eyes, or divert one's gaze. So even receptive types of attitudes can be controlled through attentional processes, and thereby gain an agentive character (without changing the fact that the attitudes still are – in fine - essentially receptive). Standard cases of epistemic actions include directed memory retrieval (in contrast with pure associative cases of memories that pop out in the mental stream), directed visualizing (in contrast with associative visualizing) and directed imagining (in contrast with passive types of imaginative episodes). Note that, in all these cases too, the subject can visualize voluntarily, try or refrain from doing it; but there is an irreducible factive aspect to such mental actions: the subject cannot change at will the content of her visualizing *p*, on pains of impairing her own mental goal, which is to visualize rather than to imagine. Other types of epistemic actions include considering alternative views (in opposition to merely understanding one view, which is not agentive), reflectively deliberating on truth, performing directed reasoning (inferential or deductive, like checking the soundness of an argument).

Although it may be useful to have a taxonomy of mental actions, epistemic agency forms a seamless continuity with non-epistemic mental agency, and with bodily action. Planning, for example, is not purely epistemic, since it includes preference management: in order to plan, one needs to know not only how the world is, but how one wants or desires the world to be; to do this, one needs to consider counterfactual possibilities, each having each a pay-off structure associated with certain attractive and undesirable features. Another form of non-epistemic agency consists in controlling one's emotions, or in using one's imaginative capacity not in order to acquire knowledge, but in order to change one's mood. In all these cases, mental agency and bodily agency form a seamless functional continuum. What explains this is the functional symmetry between an embodied mind and a 'mentalized' environment – an environment grasped as affordances rather than as objects.[v] Given that, with bodily actions, one wants to produce some change in the world, one needs to control one's attentional focus to salient affordances in the changing world (this is part of what trying to perform a bodily action consists in). In this dynamic process, one needs to perform one or more of the following actions: direct one's memorial focus to given contents, appropriately regulate one's emotions, survey one's reasonings, rehearse the instrumental steps to the goal, etc. As a consequence, acting on the world supposes the ability to proportionate one's goals to one's mental as well as one's bodily resources. Therefore, acting non-routinely on the world requires changing one's own mental properties (both epistemic and non-epistemic).[vi]

Changing one's mental properties, however, requires a very specific form of self-knowledge acquisition, in which one probes one's available mental dispositions. Before attempting to retrieve a proper name, one needs to consider whether the item is available in one's memory; before attempting to predict who will win the US Presidential election, one needs to consider whether one has the relevant evidence and has acquired the ability to form a prediction on its basis. Before one attempts to perform a complex, or time-consuming mental action, one needs to consider whether one has the necessary motivation to perform it.

In short, just as, in bodily action, one needs to consider whether one can jump over a 4-ft ditch, and usually thinks one can, although one might do it out of desperation, before deciding to do so, mental action requires an ability to predict whether one is in a position to perform it. This precondition may be missed if one fails to consider that mental action, like physical action, involves a cost; acting mentally consumes resources, it takes time; a mental action is often a key ingredient of success in the total action (for example, a flexible capacity to direct one's memory is one key for brilliant conversation or teaching). In order to decide whether one is able to act mentally in a given context, one needs to perform what I will call "self-probing" – a predictive and evaluative ability through which a thinker estimates whether a specific token of mental action can be executed, given the mental resources available to her at that moment.[vii] Its function can be presented in analogy with bodily action. Theorists of action control have shown that selecting a given command to bodily act first requires what they call "inverse modelling".[viii] When an agent forms the intention to grasp some object in the external world, she needs to select the command that will reach the object in the most efficient way given the spatial and social contexts of the action, the position of her limbs, etc. An inverse model of action thus dynamically presents the behavior that will best achieve the current intention. It responds to the unspoken question "can I do this, and how?".

Inverse modelling allows selection of one among several movements or instrumental sequences to attain one's goal. Self-probing has a similar predictive-evaluative function, the difference being that the selection process now applies to cognitive, informational states, rather than to the kinematics of bodily movement. Given the present knowledge of how memory retrieval works, it is not known whether a thinker *always* needs to select among *alternative* routes to search one's memory for a name. Clearly strategy selection sometimes occurs, for example when one tries a visualizing strategy, then opts for an alphabetical one to retrieve a name. But a more basic function for self-probing is to know whether some crucial information is in store, and how long it will take and how effortful it will be to retrieve it (or analogous procedures for learning new material, engaging in reasoning, planning, etc.). In brief, , for all cases of mental action, the mental preconditions for success may need to be checked. A mental kind of inverse modelling seems necessary to predict whether an action will be viable or not, that is: potentially successful given certain time and energy constraints and associated pay-off.

Self-probing involves establishing in a practical way whether the preconditions for a mental action hold. But a second type of evaluation needs to work retrospectively, in order to establish in a practical way whether a token of mental action has been successfully completed. To do this, a thinker needs to compare the mental property that is produced as an outcome of her mental action to her goal. What I propose to call "post-evaluation" thus checks whether a given mental action is or is not successful: was the word retrieved from memory the correct one? Was one's reasoning gapless? Did one include all the items to be counted in the final count? Again it is helpful to compare what happens at this stage with the case of bodily action. In bodily action, internal feedback – that is – the dynamic memory of prior salient steps in the consequences of the action - is compared with the present result, as presented in current perception. Is this result what was expected? To complete this evaluative step, one needs to simulate how the world should be like, and judge whether the perceived outcome fits the simulation. Does what I have in my hand feel like a glass of water (as I remember it normally feels.[ix] Similarly for mental action. Once you get to a proper name, you simulate that John's spouse is named Sue, and you see how well it fits your model of John's world. Does "Sue" sound right? "Sounding right" might again involve a form of simulation: you pretend that Sue is John's wife's first name, and see whether you have a feeling of familiarity or fluency, or rather if it does not feel exactly right. Note that this evaluation does not need to proceed through concepts. The thinker does not need to test whether the observed result falls under the conceptual representation of her mental goal. All that is needed is to perform what is called "pattern matching" between the retrieved word and the mental goal. The two types of evaluation are functionally different, insofar as their respective aims are (i) checking on the mental resources needed to perform a mental action and (ii) evaluating its outcome when performed. One is predictive, and concerns mental availability or feasibility, while the other is retrodictive, and determines whether a mental action is successful. However, they are similar in several essential respects. First, they both evaluate *success* in a strictly closed-loop, modular way. Just as an evaluation of success for a physical action only considers the hic et nunc of the action (I did jump over the ditch) without considering further consequences, metacognitive evaluation restricts itself to the question of whether a token of mental action is going to be, or has been, successfully completed: whether the name is correct, the visualization helpful, the reasoning sound, etc. It does not use – or at least does not need to use -— a rich conceptual structure to make inferences and generalizations from this evaluation. Second, they both have an *intensive, gradient-sensitive* way of evaluating, rather than a discrete judgment as to whether the mental action is possible or not, successful or not. As we shall see later, both types of evaluation crucially rely on *epistemic feelings*, i.e. affective or emotional signals. These are known, in the metacognitive literature, as feelings of knowing, of fluency, "tip of the tongue" phenomena, feelings of uncertainty, of insight, of being lost, etc. Third, in both cases, the feelings *motivate* the thinker to mentally act. In self-probing, a feeling immediately leads to execution of a mental action, or to refraining from doing so, (and selection instead of another way of responding to the current situation). In post-evaluation, the feeling leads to accepting of the action as having been successfully carried out, or to considering of corrective actions. Fourth, the question that self-probing asks is in important ways mimicked by the question that post-evaluation raises. "Do I know her name?" is echoed by "her name is most certainly Sue". This property, called "representational promiscuity" in Proust (2007), provides a strong argument for the claim that self-probing and post-evaluating are both internally related to each other and parts of the extended causal process through which mental trying is executed.

Representational promiscuity helps us understand how self-probing and post-evaluation together constitute the domain of metacognitive interventions. This domain can be described as a set of context-sensitive control cycles, in which a command is (i) probed,(ii) executed; (iii) evaluated by a comparator (i.e. by monitoring its effects in relation to standard expectations), with the outcome being (iv) a source of new commitments to act mentally. Metacognition is then a major source of epistemic motivation. If a given post-evaluation provides a "failure" answer (e.g. "Sue" feels wrong), then self-probing may be resumed on a new basis: granting that this name is wrong, do I have more choices in memory or do I find myself unable to retrieve the name? Granting that my counting is wrong, am I able to recount or should I abandon the task?[x]

These various features strongly suggest that metacognition forms a natural kind within the class of mental abilities; its general function is to evaluate the cognitive adequacy of one's dispositions to mentally act. The differences in time-orientation of the two species of metacognitive interventions are not accidental features, but rather a consequence of the rational conditions of agency itself, which requires prediction of ability and assessment of results. However, an objector might ask why metacognitive interventions do not themselves qualify as mental actions. Are not self-probing, and post-evaluating, options that can be tried or not? Successfully completed or not? If so, then why should (or indeed how could) one stop here ? Should not we also observe that self-probing needs only be performed in cases where it is rational to do so ? Therefore should we not consider that self-probing presupposes another form of probing, namely probing whether one is in a position to perform self-probing, etc. as in the famous regressus that Ryle directs against the concept of volition?[xi] Similarly, should we not have to evaluate the post-evaluation, then evaluate the second-order evaluation, and so on?

In reply to this objection, one should point out, again, that the structure of agency in the physical sense also includes as constitutive parts a self-probing and a post-evaluative phase. Neuroscience tells us that, when preparing to launch an action, the agent simulates the action. Simulating the action might contribute to selection of inverse and direct models for that particular action.[xii] It would be implausible to say that the agent performs a token of mental action whose function is to probe her ability to execute the physical action. Rather, probing one's ability to jump is analysed as a constitutive part of the associated physical action. It is a mental operation that collects the information that jumping requires: in particular, this mental operation allows an agent to know whether she can perform an action with the required parameters. Similarly, a mental action can only be rationally selected if the cognitive resources that allow the agent to complete it successfully are present; and it can only be judged successful if the agent has a way to compare its outcome with a norm of success. The core of a mental action (for example: retrieving a proper name that does not come immediately to mind) prompts the subject to first inquire whether the name can be recovered. An episode of self-probing cannot occur without an intention to perform the core of the corresponding mental action. One cannot ask oneself "can I recover this name"? for the sake of knowing whether one can, without actually trying to recover the name. [xiii]

Similarly, post-evaluating the action that has just been performed is not another mental action; for post-evaluating directly affects the agent's realizing either that the action is completed or that a new action needs to be performed. Post-evaluation cannot occur independently in this particular way, i.e. without affecting substantially the course of further action taking.[xiv] Another way of making the same point is to say that the scope of a single mental action supervenes on a functional, normative and motivational continuity between the metacognitive phases and the core mental action. We can conclude this section by saying that metacognition involves two types of evaluative intervention, respectively forward-looking and backward-looking, which do not qualify as independent mental actions, but which constitute necessary steps of every mental action.

## II

*Epistemic internalism about metacognition* - Granting that metacognitive interventions are intimately related to an evaluation of uncertainty about one's own cognitive dispositions, as opposed to uncertainty about the world, it might be tempting to take an internalist stance with respect to this form of self-knowledge. Epistemic internalism is the view that determining what one knows, what knowledge consists in, and how we can be certain that we know, are typically questions that a responsible thinker should raise. From an internalist viewpoint, furthermore, these questions can be answered on the basis of the thinker's own epistemic abilities and cognitive resources. As a consequence, a thinker is able to obtain justification for her true beliefs through introspection. Metacognitive abilities do indeed seem to provide ammunition for epistemic internalism, in that they offer both introspective access to one's mental agency and a way of evaluating its adequacy relative to criteria such as truth, efficiency and rationality. On such an internalist, Cartesian construal, a metacognitive agent has immediate, privileged, transparent access to her own mental abilities. *Immediacy* means that the access that she has to her mental contents (and in particular, to her metacognitive evaluations) does not require inference or observation of external events (in contrast with the access she may have to others' attitude contents).[xv] *Having privileged authority* refers to the fact that she alone is in a position to predict whether she will be able to perform such and such a mental action or to judge whether the outcome "looks right". The *principle of transparency* states that believing *p* entails knowing that one believes *p*. More generally, when Φ is a mental state: Φ-ing entails knowing that one Φs. A factive state such as knowledge, on this view, also qualifies for transparency. Knowing that *p* necessarily entails knowing that one knows that *p*. An internalist seems entitled to claim that transparency prevails in metacognition: forming a partial belief that *p* entails knowing the subjective strength with which that belief is entertained; similarly, self-probing whether a mental action is feasible, or post-evaluating whether it was successful, seem transparent to the agent.

These Cartesian intuitions are, finally, associated with an individualist view of the mind in which the nature of an individual's attitudes and mental contents does not depend on the physical and social environment.[xvi] This view of the mind is favored by most if not all authors with an interest in subjective uncertainty. For David Hume, for example, metacognitive awareness constitutes an inner realm separated from the world perceived and acted upon. Mental dispositions are supposed to be directly and certainly known, through metacognitive introspection,  in contrast with the uncertainty that affects perception and concept use as applied to external events.

Let us examine in more detail how internalists could back up their view that a metacognitive episode involves immediacy, privileged authority and transparency. Once a mental resource appears as instrumentally relevant in a given action (*do I remember what the object's color is?*), the subject immediately comes up with a response, that is: with an evaluation of the resource level observed, as compared to the required one. As we saw above, the metacognitive step does not consist in retrieving the object's color, but in assessing whether the task is feasible given the various constraints that apply. Similarly, in post-evaluation, a mental agent seems to be immediately aware of having or not having attained her goal - no observation or inference seems to be involved.

From an internalist viewpoint, moreover, the subject appears to have *privileged authority* for her metacognitive calls in the sense that she alone is in a position to predict whether she will be able to conduct such and such a mental operation or to judge whether the outcome "looks right". The distinctive phenomenology of epistemic feelings is a crucial internalist argument in favor of such an authority. For it is on the basis of her own epistemic experience that an agent is able to detect the availability of her mental resources (in self-probing) or the adequacy of her mental action (in post-evaluation). The agent alone is in a position to be immediately aware of her feelings and of their specific  epistemic value (for example, a tip-of-the-tongue experience). She alone can have access to the intensity of an epistemic sentiment, and predict on its basis her present disposition to mentally act, or retrospectively assess her mental action. For example, the agent has  privileged and direct access to whether she has now learnt a list of words, or not;  no one else can have such  non-inferential knowledge.

Internalists might have more problems with the principle of transparency as applied to metacognition. There is an important difference between  transparency as it is supposed to apply to metacognitive interventions and the full-blown transparency  articulated in the principle. The so-called KK principle, (KK stands for: “knowing that one knows") indeed, posits that "if one knows something, then one knows that one knows it". But the metacognitive interventions described above fail to pass such a "positive introspection" criterion. For although self-probing allows evaluation of the likelihood that one will be able to perform a mental action successfully, it does not entail "knowing reflectively that one is probing one's ability to remember *r*, to learn *p"*, etc. Similarly, post-evaluating one's mental action does not entail “knowing reflectively that one is evaluating the correctness of one's retrieval, the informational sufficiency of one's percept", etc. The reason why such entailments do not hold generally is that non-humans, who notoriously lack the ability to attribute mental states to themselves (and to others), have been found to correctly perform some types of self-probing and post-evaluating. Let us briefly survey the scientific data and their conceptual consequences.

Evidence from comparative psychology suggests that macaques and dolphins can evaluate their abilities to perceive or to remember a target stimulus, and seem to make rational decisions based on these evaluations.[xvii] When the target stimulus is difficult to remember, or is hard to discriminate perceptually from another, animals choose *not to* volunteer a response if they are offered the choice. Furthermore, their responses are more reliable when they are free to respond or not than when they are forced to provide an answer. If such a rational sensitivy to one's own reliability obeys the principle of transparency, the animals should be able to represent their own mental states of perceptual (or memory) discrimination, through higher-order representations (that is: as perceptions, or as memories). A plausible claim indeed is that the ability to extract and exploit subjective uncertainty requires the ability to represent oneself as a thinker, as well as to represent that one forms attitudes of a certain type, *and* that one's attitudes may end up being correct or not, felicitous or not, etc.. Developing in full what is needed to self-attribute a degree of confidence for some perceptual judgment, would include the following various abilities:

1. The ability to form a first-order representation, whose verbal equivalent is "*O is F*"

2. The ability to form the metarepresentation of an epistemic or a conative attitude directed at that content, such as, "I perceive (believe etc.) that *O is F*"

3. The ability to attribute to the metarepresentation a property that qualifies its relation with one's first-order representation, such as, " I perceive with uncertainty *r*, that *O is F"*

4. The ability to form a judgment qualifying my occurrent perception of *O*, such as, "I judge that I perceive, with uncertainty *r*, that *O is F*". This judgment involves the attribution

of the first-order, second order and third-order representations to myself *as one and the same* thinker of these representations, that is, to have a representation of the form:

Self, $\mathrm{PA}_{2\ (=\mathrm{judge}),}$ (self , ($\mathrm{PA}_{1(=\mathrm{perceive}\ ,\ \mathrm{with\ uncertainty\ r})}$ {*O is F*}) [xviii]

What makes this analysis attractive is that it clarifies the mental concepts that need to be in a thinker's repertoire to enable her to produce a fully explicit representation of her uncertainty, and to communicate to others her degree of confidence in the success of the corresponding mental action. What makes it deeply problematic, however, in the case of animal metacognition, is that conditions 2 to 4 are not met in most, if not all non-humans. Macaques, for example, have no mental concepts, and cannot, therefore, metarepresent that they *perceive* or that they *judge that P*.[xix] An important epistemological lesson is to be drawn from this comparative evidence: the source of predictive/evaluative practical self-knowledge produced by metacognitive interventions cannot consist in a theoretical body of social knowledge. There must be a form of pratical access to self-knowledge that allows non-mentalizers to perform various types of mental actions without needing a conceptual representation of the fact that they do so.

An epistemic internalist might respond to these considerations in two ways. First, she might claim that although some mental agents may not have all the concepts required for full transparency to hold, still KK applies whenever the agents have these concepts available. Second, she might insist that epistemic feelings do not need to be associated with conceptual contents to be efficient in guiding and motivating rational behavior. Even when self-probing for perceiving or remembering, say, fails to be luminous, and even when it may occur outside of awareness, epistemic feelings are transparent: they carry information about one's ability to perceive or to remember. This information is made available to each thinker, and tells her *whether*, and, in relevant cases, *when* to mentally act. These feelings have the function of indicating the normative status of a considered or executed mental action. But they do not carry this information[xx] as a propositional content would, by attributing a property to a particular. Rather, one might hypothesize that the information they carry is feature-like and non-conceptual.[xxi] Combining the two parries, the internalist might conclude that transparency of epistemic feelings finally holds, in the sense that agents know that they feel that they know (or can know) that *p* if they have the relevant concepts available. When agents do not have the relevant concepts, they merely feel like (are attracted to, or repelled from) performing an action that is in fact mental, but that they do not need to represent *as* mental.

To summarize, on this construal of metacognition, metacognitive awareness constitutes an inner realm separate from the world that is perceived and acted upon. Mental dispositions are immediately and certainly known, through metacognitive introspection, in contrast with the uncertainty that affects perception and concept use as applied to external events. Metacognitive episodes seem to enjoy first-person authority with respect to the evaluations that are generated. Metacognitive achievements seem to confirm that the mind is transparent to itself, as Post-Cartesians hold, and that introspection necessarily delivers true reflexive judgments concerning one's occurrent states.

## III

*Epistemic externalism about metacognition-* There are, however, externalist motivations for resisting this picture of metacognition, and the concepts of self-knowledge and justification that inspire it. Two different types of externalism object to the very idea of having privileged, transparent, and immediate access to the likely validity of one's own thought contents. *Externalism about meaning*, on the one hand, as traditionally conceived (Putnam, 1975, Burge, 1979), is the view that facts about the physical and the linguistic environment determine mental content. On this view, subjects do not have full command of the content of their mental states, because content is essentially relational.[xxii] Meaning externalism has an impact on the content of self-knowledge. Given that we have no authority with respect to the meaning of our thoughts, it would seem that we have no authority either when attributing to ourselves the content of our thoughts.[xxiii] Furthermore, we are in no position to appreciate our margin of error relative to our knowledge, because the relevant evidence, again, is in principle not available.[xxiv] *Epistemic externalism,* on the other hand, is the view that a subject does not need to know that she knows in order to be entitled to knowledge. Epistemic externalists claim that knowledge depends upon a relation between the believer and the world, and does not need to be formed as a consequence of a subject's having access to reasons for believing what she does. A basic condition for knowledge attribution is one of reliability.[xxv] A belief counts as knowledge, on this view, because it is produced by a generally reliable process, leading to a high proportion of true beliefs.

From an Epistemic externalist's viewpoint, the internalist's emphasis on epistemic feelings (as a way to account for the animal evidence, while also securing transparency, subjective authority and immediacy of metacognitive contents) is misguided. A main worry is that internalists are attributing to an agent's metacognitive abilities - narrowly construed - the disposition to access mental contents and their degree of certainty. From an internalist viewpoint, self-probing is made possible by inspecting one's feelings; these subjective feelings reliably track the cognitive adequacy of the ensuing mental action. The agent does not have to turn to the world in order to know whether her mental action is likely to succeed. Similarly, post-evaluation crucially involves feelings that reliably track objective truth or correctness, on the basis of introspection alone. But it may be objected that this explanation leaves it completely mysterious how epistemic feelings might reliably track norms such as cognitive adequacy, truth or correctness. The externalist source of the worry is that the concept of an epistemic norm (such as truth, or rationality) cannot be grounded in a strictly subjective process or ability. A norm provides a principled way of comparing one's mental contents with external constraints or facts. Therefore a substantive part of the normative explanation will be left out if one concentrates on the processes through which the subject effects the comparison. Internalists explain that epistemic feelings indicate proximity to a norm of mental actions; but they do not explain on what information this pivotal role depends, that is, what is the objective basis of the norm itself. A difficult, but major, issue, for an epistemic externalist about metacognition, is how to identify the objective facts of the matter that, beyond a subject's ken, govern norms, and explain why epistemic feelings are calibrated the way they are. In other terms, there must exist independent 'evidence' or facts to which the feelings correlate. Otherwise, one will lack any explanation for why a norm works as a constraint that a subject needs to approximate if she is to succeed in her mental actions.

Let us illustrate the incompleteness of an account that ignores the distal source on which epistemic feelings are grounded through the following "brain in the lab" experiment. Suppose that a mad scientist provides Hillary with regular spurious feedback on how she performs in a type of mental task. Whenever she performs a given *type* of mental action (such as retrieving a name, performing an arithmetic calculation, controlling her perception for accuracy, checking the soundness of her reasoning), she will receive consistently biased feedback; she will be led to believe that her mental actions of that type are always correct. For the sake of the argument, let us assume that Hillary has no way of figuring out that the feedback that she receives is systematically biased.There are several ways of exposing Hillary to biased feedback. The mad scientist can explicitly misinform her, by systematically telling her – after a block of trials - that she is performing well above average, even when it is not the case. Or, still more cunningly, the scientist can use implicit forms of spurious feedback – that is produce a kind of feedback that is extracted by Hillary in ways that she cannot consciously identify. For example, self-probing can be manipulated by the perceptual aspect of the tasks: using familiar items for new tasks misleads her into believing that these tasks are easier than they are. Another trick is to manipulate the order in which tasks of a given level of difficulty are presented. When the tasks are ordered from more difficult to less difficult, Hillary will have a misleading feeling of growing self-confidence. Priming can also be used to prompt Hillary with the correct solution, which she will believe to have found herself. The mad scientist can combine these various ways of manipulating self-confidence; he can use priming to make certain tasks easier, while offering verbal feedback to the effect that these are the difficult ones. To prevent Hillary from having the sense that she is being manipulated, there are several strategies that the mad scientist could use; he can, for example, organize the temporal pattern of the responses in a way that prevents Hillary from performing a careful post-evaluation. Alternatively, he can erase her own post-evaluations from her memory by applying for example, well-targeted transcranial magnetic stimulations each time she performs one.

After being trained in this biased way, Hillary's epistemic feelings have become highly unreliable. She now feels over-self-confident in new tasks belonging to the type that was biased. When she needs to probe whether she can quickly calculate an arithmetical operation, for example, she will tend to have the feeling that she can perform it, and, after having performed a mental action of that type, she will tend to feel that it was correct even when it is not. This thought experiment only generalizes experimental work, showing that subjects calibrate their epistemic feelings on the history of their previous results over time in mental actions of the relevant type. As Asher Koriat observes, "It is because metacognitive judgements rely on the feedback from control operations that they are generally accurate" (Koriat, 2006). As a consequence, tampering with feedback decalibrates a subject's epistemic feelings. One can thus conclude that the existence and reliability of epistemic feelings *supervene in part on* the existence and quality of the feedback provided. Therefore, the internalist case for epistemic feelings as a source of epistemic intuition loses considerably in explanatory force and credibility. Epistemic feelings are not sufficient to explain why a subject can perform accurate self-probing and post-evaluation; furthermore, epistemic feelings can be illusory, in the sense that they can systematically lead one to make irrational decisions on how to act mentally, if they have the wrong informational history.

It is worth reflecting, therefore, on the objective conditions that make epistemic feedback reliable. Saying that these objective conditions are those which produce a majority of accurate metacognitive calls would be circular, because that is merely how process-reliability is defined. The Epistemic externalist needs to uncover the objective basis that a mental agent has for inferring that her epistemic feelings reliably track a norm. Two types of externalist account have been offered for the epistemological grounding of self-awareness of physical and mental agency, which suggest applying a similar strategy to metacognition.

The first holds that metacognitive evaluations tend to be reliable because, as is the case for awareness of agency, feelings have been selected to be reliable. This is a view similar to Peacocke's explanation of the entitlement to represent oneself as the agent of an action: "States of their kind have evolved by a selection process, one which favours the occurrence of those states whose representational content is correct" (Peacocke, forthcoming). Evolution is supposed to have found the way to track correctness in metacognition, thanks to a context-sensitive process of norm-tracking. In other terms, adapting Plantinga's conditions for warrant as applied to belief,[xxvi] one might suppose that "the segment of the design plan governing the production of that belief [metacognitive operation] is aimed at the production of true beliefs [correct evaluations]". The mad scientist case, in this explanatory framework, would be accounted for by the fact that the experimental conditions modify the evolved ways of applying self-probing and post-evaluations. A subject has to be in conditions suitable for learning about the world through her actions (by its failure/success pattern), on this view, to correctly calibrate her feelings with objective norms. Hillary's metacognitive capacities are arbitratrily severed from their normal feedback. Given that the conditions associated with their proper functioning are not met, it is predictable that her metacognitive feelings will lead her astray.

This form of teleological explanation however, fails to account for the environmental, physical or social conditions that make feedback reliable. First, as Dretske (2000) observes for belief, it may explain why metacognition was reliable in the circumstances in which it evolved – in the past -, but not why it is presently reliable. Second, a selectionist explanation is ill-equipped to explain why a given capacity has been selected among competitors.[xxvii] This observation applies directly to the case of metacognition. Explaining the general reliability of metacognition by saying that there is an evolved metacognitive ability whose function is to track correctness has  low explanatory value. A genuine explanation should be causal, rather than merely teleological: it should describe the type of information that has to be available for an organism to make flexible calls, and then explain how a given ability can extract and use this information.

A tentative answer to this question, aiming to point to the relevant informational source, is considered in Carruthers (2008). On this view, metacognition uses the same kind of information as cognition does, namely objective frequencies. Here is, roughly, the argument in favor of what might be called a "reductive strategy" with respect to metacognition. Epistemic feelings and emotions have two major functions in animal cognition; they predict states of affairs relevant to survival, and they help select and guide adequate action programs. Although the associated emotional states carry information that may be relevant to self-knowledge, such information does not need to be extracted and processed for an organism to make rational decisions. Animal cases of metacognition, in particular, can be explained in strictly first-order terms, as a function of first-order beliefs and desires. Consider surprise. Being surprised seems to presuppose that one is aware of having a belief, and that one suddently realizes that one's belief is false. On this construal, surprise seems to *essentially* involve belief metarepresentation, updating and revision. Surprise, however, is not *essentially* self-attributive. Surprise results from a mismatch between an observed and an anticipated fact, and disposes the agent to update and revise its beliefs and motivations to act, without *needing* to use the concept of a false belief (nor of a desire). Non-human animals seem to experience surprise as we do, although they are unable to attribute beliefs to themselves. Surprise thus merely involves a salient, unanticipated change in the environment, and a motivation to adjust to it.

What holds for surprise also holds for other affective states whose function is to predict the likelihood that a current goal can be fulfilled. Many of the feelings related to agency, such as the sense of being in control (sense of agency) or the sense of moving one's body (sense of ownership), the sense of physical exertion, of purposiveness, of anticipated or observed success or failure, are components of basic control structures involved in first-order representation of action. They are instrumental conditions, that evaluate whether a token of action is developing successfully on the various dimensions to be monitored. It would be unparsimonious to offer a 'metacognitive' interpretation of these feelings, which are part and parcel of bodily action. So, then, what does make "metacognition" special? Metacognition merely consists in practical reasoning dealing with world uncertainty, with a gate-keeping mechanism that helps the organism to make decisions when the world becomes too unpredictable. This view thus contrasts first-order types of information processing that are abusively called "metacognitive" in animal research with conceptual forms of self-attribution in humans, which qualify as genuinely metacognitive. They involve a mindreading ability, i.e. the capacity to conceptually represent mental states in self and in others. From an externalist viewpoint, social interactions in a linguistic environment determine conceptual contents; the view is thus that a proper social and linguistic environment, in conjunction with evolved mechanisms, will explain how thinkers can have access to self-knowledge.

This reductive strategy, however, does not do better than the evolutionary strategy at accounting for the fact that metacognition is reliable. If the reductive strategy is correct, macaques and dolphins should lack the ability to decide rationally what to do when they are presented with difficult new stimuli, or when their memory is failing. Macaques and dolphins, however, present humanlike metacognitive performance, in contrast with other species, such as pigeons. These performances are not explainable in first-order terms, for reasons reviewed elsewhere.[xxviii] Furthermore, the reductive strategy should explain how metacognitive performance can be correct in human subjects with a poor mindreading ability, such as children with autism. We can conclude, then, that the reductive strategy does not successfully explain how metacognitive evaluations, and mental agency in general, can be reliably conducted.

Let us now take stock. We saw in section II that the objective ground for the reliability of the kind of predictive/evaluative practical self-knowledge produced by metacognitive interventions cannot consist in a theoretical body of social knowledge that would be linguistically conveyed to mental agents. We rejected an internalist solution in which epistemic feelings can be immediately accessed and are sufficient to guide mental agency and promote rational behavior. Section III examined two strategies that could be used to ground reliability of metacognitive feelings. The evolutionary strategy was found to be ill-equipped to respond to this question in a developmental way. The reductive strategy has also been found wanting, in that metacognition does not use the objective frequencies of external events and the associated pay-off structure to produce evaluations of feasibility and correctness. We will now attempt to articulate a third view, explaining metacognitive correctness through its tracking of a dynamic norm, exemplified although not yet articulated in our 'brain in the lab experiment'.

## IV

*External norm and dynamic coupling regularities* - As a follow-up to our 'Hillary and the mad scientist' story, let us imagine a twin-Hillary on the model of Putnam's Oscar twins. Twin Hillary is similar to Hillary in all respects, except that the feedback she receives in self-probing and post-evaluation is generated by a normal, well-calibrated comparator. For the sake of the argument, let us assume again that Hillary has no way of figuring out that the feedback she receives is systematically biased. Let us now take a given metacognitive episode, in which Hillary and twin-Hillary, having to retrieve a word from memory, both have a feeling of knowing that word. From an internalist viewpoint, they are identical twins at the time of this episode. They have the same feeling, and are similarly motivated to search their memory as a result of this feeling. In addition, they both actually produce the correct word when they have performed their directed memory search, and correctly evaluate that the word retrieved is the one they were searching. Twin-Hillary has accurate information available about her mental action and has gained and used appropriate self-knowledge; but, as a consequence of the lack of reliability of Hillary's method of acquiring self-knowledge, the latter does not have the corresponding metacognitive knowledge: had the word been difficult, she would *not* have retrieved it (although she would have been convinced by the mad scientist that she did), and she would have failed to detect her failure. Therefore it is quite contingent that she has produced the correct answer, which therefore cannot count as knowledge.

Clearly, the difference between the knowledge status of Hillary and Twin-Hillary for this specific metacognitive episode is not subjectively accessible; they have formed the same feeling, and are totally unaware that feelings can be induced through external means. This conclusion is familiar to Epistemic externalists, who claim that a subject can be entitled to self-knowledge although the subject is not in a position to know that she knows (for example, because she lacks the concept of knowledge and of self), not even to know when she knows (because when her prediction is wrong, although her ability to predict is generally reliable, she has acquired a false belief about her disposition). To understand what the source is of Twin-Hillary's entitlement to self-knowledge, in contrast to Hillary's, we need to explore further the objective basis on which metacognitive comparators depend for their reliability.

Here is a suggestion that we might call 'the dynamic strategy'. According to this new approach, the objective basis is constituted by the *architectural constraints* that universally apply to *a sustained, adaptively-controlled activity*[xxix]. Let us first explain how one can articulate norms on the basis of architectural contraints and goals. The relevant constraints, in the case of action, are those deriving from an adaptive control architecture. Any cognitively operated control operation (whether cognitive or metacognitive) can be compared with a corresponding ideal strategy, as determined by *a priori* mathematical reasoning, on the basis of the agent's utilities and costs. In Signal Detection Theory, a "norm" of decision can be computed a priori in any noise and signal+noise probability distribution for a given pay-off schedule. Let us call this kind of normativity "prescriptive normativity" (or P-normativity) as it suggests that, each time a cognitively operated control system is active, an optimal solution exists that the system *should* select. P-normativity so understood does not have to be restricted to agents who can in fact understand that their control operations have to follow specific norms. P-normativity can be approximated through the objective constraints that result from probability theory as applied to a set of stimuli.

Metacognition differs from Signal Detection in that it forms evaluations on the basis of an extended sequence of *dynamic couplings* effected in prior interactions between mental actions and monitored outcomes. So we need to examine the kind of prescriptive normativity that applies to this form of dynamic coupling, that is, the system of rules that determines the rational decision for each attempted mental action. Let us introduce a technical term borrowed from a mathematical theory for adaptive control, called 'viability theory'. Very roughly, this theory describes the norm for adaptive control as one that allows the evolution of a system to remain within the limits of a "viability core" (VC).[xxx] A VC for Φ-ing determines at each new time and portion of the regulation space where the agent stands, whether it is rational to Φ at *t*. One of the important assumptions of this mathematical theory is that, however complex the parameters that influence the shape of VC for Φ-ing, the relevant information concerns only the limit of the viability core, which can be discovered on the basis of prior feedback. Neither Hillary nor twin-Hillary know what a viability core is. But Twin Hillary has feelings based on reliable feedback. Emotions are perfectly suited to reflect holistic types of constraints, and thus are ideal for converting multiple dimensions to a single decision. For twin-Hillary, her dynamic feedback (ie. its evolution over time, with specific patterns that will not concern us here) now *explains* why she is able to *sense* where the limits of the viability core for directed recall are. Her feelings have been correctly educated, because they are based on sufficient information concerning the VC for directed recall. These educated feelings should be called 'sentiments'[xxxi]: they carry information about the viability core for active remembering. This norm is unarticulated, but it influences Twin-Hillary's mental agency through the sentiments that have been educated to track it.

Hillary, on the other hand, has no such sentiment. She indeed has a feeling, but her feeling does not carry information about where the norm lies. Our thought experiment thus allows us to distinguish cases where one is entitled to form metacognitive evaluations from cases where one is not. An agent is entitled to act on her metacognitive evaluations if her sentiments track the viability core of the corresponding mental action. If however, her feelings have never been so educated, the agent has no entitlement to metacognitive evaluation.

Given the predictive function of the viability core, it may happen that a given prediction goes wrong; this is so because coupled evolutions obey inertia principles, and are therefore open to hysteresis effects; abrupt changes that can affect the brain in unlawlike ways may bring the agent to misperceive for some time the limits of her viability core.[xxxii] But this does not mean the agent is not entitled to have formed a correct metacognitive evaluation, for she formed it under the influence of the information concerning VC. An agent with reliable feelings will in this case be entitled to form the corresponding metacognitive evaluation.

The two reasons offered for endorsing Epistemic Internalism about metacognition now appear dubious. Transparency is illusory: although a mental agent has recognizable feelings that normally dispose her to mentally act in a certain way, she is necessarily unaware of the dynamic facts that make these feelings trustworthy. The contributions of the social and the physical dynamic environment are essential to adequately calibrate the corresponding epistemic sentiments. First-person authority is jeopardized in turn, as the mental agent is not in a position to know *when* she makes adequate calls: she thus depends on others and on the world for her self-knowledge.

Let us summarize the main arguments leading to this conclusion. First, a mental agent inherits a given cognitive architecture, in which she can only properly perform mental actions when she has the disposition to perform the associated metacognitive evaluations. Second, she cannot decide which control will allow her, for example, to retrieve a memory or check an argument, for these regulations are fixed by the design of her metacognition. Regulation laws determine which outcome can be associated with which command; these laws can be practically relied upon, but they are not clearly understood by normal agents. Third, she cannot decide, either, which portion of the regulation space is accessible to her at a given time: for this depends on developmental facts about her, such as age, prior experience, etc. Fourth, the same extraneity characterizes the 'monitoring' aspect of self-knowledge: a mental agent enjoys epistemic feelings or sentiments as a result of her mental history, which crucially involves a social and physical environment. Fifth, she has no introspective way of knowing whether her epistemic feelings are correctly calibrated or not. Furthermore, she cannot recalibrate them, because this calibration depends on independent dynamic facts - the viability core for the associated regulation. Finally, the mental agent has no other choice but to trust her epistemic feelings, although she cannot calibrate them.

We can thus conclude that the ability to control one's mental agency is not itself able to be under the agent's control, for the control architecture is given to the agent and shaped by the dynamic environment in an opaque way. Therefore it is not clear that a metacognitive agent should be held responsible for her evaluations. Even an agent equipped with mindreading abilities, and able to grasp the limitations of her aptitude to govern herself, would still have, in fine, to depend on her sentiments to assess the viability of her mental actions.

*Acknowledgment*

I have benefited from useful discussions concerning some of the issues raised in this paper with my colleague Claudine Tiercelin, with Mike Martin, Sam Guttenplan and other participants of the meeting at the Aristotelian Society, and with the members of the Metacognition group at the IJN. I am deeply grateful to Dick Carter for his comments, questions, and linguistic suggestions. This article was written as part of the European Science Foundation EUROCORES Programme CNCC, and was supported by funds from CNRS and the EC Sixth Framework Programme under Contract no. ERAS-CT-2003-980409.

## References

Alston, W.P. 2005. *Beyond "Justification",* Ithaca & London: Cornell University Press.

Aubin, J. P. 1991. *Viability theory*. Heidelberg: Birkhaüser.

Burge, T. 1986. Cartesian Error and The Objectivity of Perception, in T. Burge, *Foundations of mind, Philosophical Essays vol II,* Oxford: Oxford University Press 192-207.

Carruthers, P. 2008. Meta-cognition in Animals: a Skeptical Look. *Mind and Language*, 23, 1, 58-89.

Clark, A. & Chalmers, D. 1998. The extended mind, *Analysis*, 58, 10-23.

Cussins, A. 1992. Content, Embodiment and Objectivity: The Theory of Cognitive Trails, *Mind,* 101, 651-688.

Decety, J. 2001. Neurophysiological Evidence for Simulation of Action, in J. Dokic, J. & J. Proust, (eds.), *Simulation and knowledge of action,* Amsterdam : John Benjamins.

Dretske, F. 1988. *Explaining Behavior, Reasons in a World of Causes,* Cambridge, MIT Press.

Dretske,F. 2000. Entitlement: Epistemic Rights without epistemic duties? *Philosophy and Phenomenological Research*, 60, 3, 591-606.

Goldman, A. 1979. "What is Justified Belief?" in G. Pappas, (ed.), *Justification and Knowledge: New Studies in Epistemology,*Dordrecht: Reidel,1-23.

Greco, J. 2001. "Virtues and Rules in Epistemology," in *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*, Abrol Fairweather and Linda Zagzebski, eds., Oxford: Oxford University Press.

Hookway, C. 2003. Affective States and Epistemic Immediacy, *Metaphilosophy*, 34, 1-2 78-96.

Hookway, C. 2008. Epistemic immediacy, doubt and anxiety: on the role of affective states in epistemic evaluation, in Kuenzle & Doguouglou(eds.), Epistemology and the Emotions, forthcoming.

Hume, D. [1739-1740]. 1978. *A Treatise of Human Nature*. Oxford: Oxford University Press.

Koriat, A. 2000, The Feeling of Knowing: some metatheoretical Implications for Consciousness and Control. *Consciousness and Cognition*, 9, 149-171.

Koriat, A., Ma'ayan, H., Nussinson, R. 2006. The Intricate Relationships Between Monitoring and Control in Metacognition: Lessons for the Cause-and-Effect Relation Between Subjective Experience and Behavior. *Journal of Experimental Psychology: General*, 135,1, 36-69.

Kornell, N., Son, L. K., Terrace, H. S., 2007. Transfer of Metacognitive Skills and Hint Seeking in Monkeys, *Psychological Science*, 18, 1, pp. 64-71.

Naccache, L., Dehaene, S., Cohen, L., Habert, M.-O., Guichart-Gomez, E., Galanaud, D. & Willer, J.-C. (2005). Effortless control: executive attention and conscious feeling of mental effort are dissociable. *Neuropsychologia*, 43: 1318-1328.

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

Plantinga, A. 1993. *Warrant and proper function*, New York: Oxford University Press.

Proust, J. 2001. A plea for mental acts, *Synthese*, 2001, 129, 105-128

Proust, J. 2006. Rationality and metacognition in non-human animals, in S. Hurley & M. Nudds (eds.), *Rational Animals ?,* Oxford, Oxford University Press, 247-274.

Proust, J. 2007. Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition ? *Synthese*, 2007, 2, 271-295.

Proust, J. 2008. Is there a sense of agency for thought ? in L. O'Brien and M. Soteriou (eds.) *Mental Action* Oxford, Oxford University Press.

Proust, J. forthcoming. Which representational format for metacognition? in R. Lurz (ed.) *Animal cognition,* Oxford: Oxford University Press.

Putnam, H. (1975/1985) "The meaning of 'meaning'". *In Philosophical Papers, Vol. 2: Mind, Language and Reality*. Cambridge 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 Monitoriing, Journal of Experimental psychology: general, 135, 2, 282-297.

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.

Sober, E. 1984 . *The Nature of Selection*, Chicago, University of Chicago Press.

Tiercelin, C. 2005. *Le doute en question : Parades pragmatistes au défi sceptique*. Paris: Editions de l'Eclat.

Tye, M. and McLaughlin, B. 1998. 'Externalism, Twin Earth, and Self-Knowledge', in *Knowing our Own Minds*, C. Wright, B. Smith, and C. MacDonald (eds.), Oxford: Clarendon Press, 285-320.

Williamson, T. 2000. *Knowledge and its limits*. Oxford: Oxford University Press.

Wolpert, D. M. & Kawato, M. 1998. Multiple paired forward and inverse models for motor control, *Neural Networks*,11, 7-8, pp.1317-1329.

[i] See Alston (2005).

[ii] Dretske (2000). For a defense of virtue epistemology, see Greco (2001).

[iii] Another definition based on "trying" is provided in Peacocke (forthcoming).The particular definition one uses does not affect the argument concerning metacognition that is proposed here.

[iv] On the distinction between mental operation and mental action, see Proust (2001).

[v] On this symmetry, as expressed in a feature-placing representational format, see Cussins (1992) and Proust (forthcoming).

[vi] The distinction between habitual and innovative? action is crucial to understanding the difference between humans and non-humans. Habit selects programs of action in which the necessary attentional levels are learnt over time. New actions, however, require from the agent an additional form of flexibility, having to do with self-control.

[vii] A number of mental or physical actions are not performed properly, because self-probing is not conducted adequately. How careful self-probing should be depends on the type of outcome that the agent wants to obtain (or avoid), given her prior experience of performing actions of that type. The consequences of a bad decision, in both cases, can, in certain contexts, be severely detrimental to the agent and/or to others (The Chernobyl disaster is a direct consequence of a repeated failure to conduct adequate self-probing, both at the individual and at the institutional levels).

[viii] Cf. Wolpert and Kawato (1998).

[ix] Obviously, one does not need to use words to conduct post-evaluation: expected sensory feedback is represented in non-verbal format. See Proust (forthcoming).

[x] An additional common feature is that the evaluation is conducted in both cases through *self-simulation*. Self-simulation is a dynamic process that makes available previous experience to the comparator in a format immediately relevant to the present context. This feature may not belong to the essence of self-knowledge, but to the causal processes that it involves.

[xi] See Proust (2001) for a discussion of Ryle's arguments.

[xii] See Decety (2001), Wolpert & Kawato (1998).

[xiii] This impossibility is related to the fact that metacognition is an engaged, simulatory process, in contrast with 'shallow' access to self-knowledge, which metarepresentational attribution provides. On this distinction, see Proust (2007).

[xiv] Obviously one can evaluate the physical action of someone else, or even one's own mental action, in a detached way; but this detached evaluation, performed in a third-person kind of way, does not qualify as metacognition, because it uses concepts and metarepresentations, and does not need to immediately promote further rational decisions to act or not.

[xv] Immediacy of access is a precondition for immediacy of justification. As Hookway (2008) emphasizes, there are two such notions: a belief can be immediately justified either when its justification does not depend upon the believer being able to offer reasons or arguments in its support, or when it does not depend upon other information the agent possesses about the world at all. The kind of immediacy that is relevant to the strong internalist type of justification is the second kind.

[xvi] For a critical approach, see Burge (1986).

[xvii] Smith et al., 2003, 2006.

[xviii] Should a representation of "self" be included to fill-in one of the arguments in the two-place predicate PA2, as suggested in 4 ? Given that PA2 is not itself in the scope of a higher metarepresentation, it can be argued that the judging self does not need to be fully represented. Some reflexive marker should, however, bind PA$_2$ and PA$_{1.}$

[xix] Cf. Proust (2007), for additional reasons in favor of the view that metacognition does not have a metarepresentational structure.

[xx] Certain dimensions of the feeling carry non-conceptual information about features such as: the feasibility of the mental action (within reach or not), the temporal pattern of such feasibility, the effort involved (easy, difficult), and the urgency of the mental action considered. A high value on each dimension seems to predict high subjective confidence and likelihood of successful execution of the mental action considered or performed.

[xxi] This claim is defended in (Proust, forthcoming).

[xxii] Another form of content externalism takes the active contribution of the environment to cognitive processing (Clark & Chalmers, 1998) to be partly constitutive of meaning. This "active externalism" will not be discussed here.

[xxiii] Tye & McLaughlin,1998.

[xxiv] Williamson (2000)

[xxv] See Goldman (1979).

[xxvi] Plantinga (1993), 46-47.

[xxvii] See Sober (1984), Dretske (1988).

[xxviii] See Proust (forthcoming).

[xxix] Control systems involve a loop in which a command is selected and sent to an effector, which in turn regulates further control states. Devices that use information as a specific causal medium between regulating and regulated subsystems are called "adaptive control systems". See Proust (2006).

[xxx] See Aubin (1991).

[xxxi] I am relying here on Claudine Tiercelin's Peircian suggestion, in Tiercelin (2005). Hookway (2008), also following Peirce, similarly contrasts raw feeling with educated sentiment.

[xxxii] A hysteresis effect occurs when a system's output is determined in part by the path that the input followed before it reached its current value.