Metacognitive Feelings, Self-Ascriptions and Mental Actions

Santiago Arango-Muñoz

Abstract: The main aim of this paper is to clarify the relation between epistemic feelings, mental action, and self-ascription. Acting mentally and/or thinking about one's mental states are two possible outcomes of epistemic or metacognitive feelings. Our mental actions are often guided by our E-feelings, such as when we check what we just saw based on a feeling of visual uncertainty; but thought about our own perceptual states and capacities can also be triggered by the same E-feelings. The first section of the paper presents Dokic's argument for the insufficiency of the "ascent routine" to account for non-transparent cases of self-ascription, as well as his account of E-feelings. The second section then presents a two-level model of metacognition that builds on Dokic's account and my own view of the issue. The two-level model links E-feelings to a mind-reading capacity in order to account for non-transparent self-ascriptions. Finally, the third section develops a deeper characterization of the relation among E-feelings, mental action, and self-ascription of mental states based on epistemic rules. In the context of self-knowledge, these remarks suggest the existence of means of forming self-ascriptions other than the ascent routine.

1. *Introduction*

Our bodily as well as our mental behavior is caused not only by cognitive mental states, such as beliefs and desires, but also by phenomenal experiences, such as emotions and feelings. We often act guided by our feelings, such as when we check what we just saw based on a feeling of visual uncertainty; but we can also start thinking about our own perceptual states and capacities based on the same feelings. Mentally acting and/or thinking about one's mental states are two possible outcomes of epistemic or metacognitive feelings (henceforth E-feelings), but they need not always go together: sometimes, we act mentally without thinking about it, and sometimes we think about our cognitive processes without acting upon them (see §3).

E-feelings are phenomenal experiences that point towards mental capacities, processes, and dispositions of the subject, such as knowledge, ignorance, or uncertainty (de Sousa 2008; Dokic 2012; Arango-Muñoz 2013a).

ISSN (print) 2281-8618-ETS

The following are some instances of E-feelings: the feeling of knowing (henceforth FOK; Reder 1987, 1996; Koriat 1993, 2000), the feeling of confidence (Koriat,2008; Brewer and Sampaio 2012), the feeling of error (Arango-Muñoz *et al.* in preparation), the feeling of forgetting (Henceforth FOF; Halamish *et al.* 2011), and the tip-of-the tongue experience (Henceforth TOT; Schwartz 2002).[1] These feelings tell the subject something about her own mind and motivate her to perform certain mental actions, such as retrieving information from her memory when she has a feeling of knowing, or endorsing retrieved information when she feels certain about it, but they also motivate self-reflection and/or introspective self-ascriptions.

The main aim of this paper will be to clarify the relation between E-feelings, mental actions, and self-ascriptions. The first section presents Dokic's argument for the insufficiency of the "ascent routine" to account for non-transparent cases of self-ascription, and his account of E-feelings. Then, the second section presents a two-level model of metacognition that builds on Dokic's account of E-feelings and my own view of the issue. The two-level model of metacognition links E-feelings to a mindreading capacity to account for non-transparent self-ascriptions. Finally, the third section develops a deeper characterization of the relation among E-feelings, mental action, and self-ascription of mental states based on epistemic rules. The main idea is to determine how epistemic or metacognitive feelings motivate mental action and/or thought about one's own mental states.

2. *Dokic on the Ascent Routine and E-Feelings*

When it comes to accounting for our capacity to self-ascribe metal states and properties, some philosophers seem inclined for versions of the so-called "ascent routine" (e.g., Gordon 1995, 2007; Moran 2001; Byrne 2005, 2011a, 2011b, 2011c). The main idea of the ascent routine was originally proposed by Gareth Evans (1982); according to his account, one can derive a self-ascription about an internal mental state from a judgment about the external world: "I get myself in position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p" (Evans 1982: 225). The classic example proposed by Evans concerned the way one finds out whether one believes that there will be a third world war; in order to determine this, one only needs to consider the external factors that would determine a third world war. Richard Moran (2001) synthe-

[1] For a more comprehensive list, see "Epistemic feelings, epistemic emotions: Review and introduction to the focus section", in this issue.

sizes the idea in the following way: "In ordinary circumstances a claim concerning one's attitudes counts as a claim about their *objects*, about the world one's attitudes are directed on" (Moran 2001: 92). In other words, in order to determine whether one *believes* that p one just needs to judge whether p.

In a recent paper, Jérôme Dokic (2012) presents a simple but convincing argument demonstrating the insufficiency of the "ascent routine" to account for cases of non-transparent self-ascription; i.e., when a subject self-ascribes a mental state without having access to its intentional content. This happens mainly when the mental state in question is a disposition, a non-occurrent belief, or a mental capacity. Dokic's strategy is to show that one can only resort to the "ascent routine" to obtain a self-ascription from a judgment about the external world if one focus on yes-no or polar questions like (Q1) "Is Lima the capital of Peru?" or (Q2) "Do you believe that Lima is the capital of Peru?". Although Q1 concerns the external world and Q2 concerns an internal belief of the addressee, the answer to Q2 can be derived from the answer to Q1. This is the core of the ascent routine: one derives a self-ascription from a judgment about the world. In contrast, the same does not follow if one focus on non-polar questions like (Q3) "What's the capital of Peru?" or (Q4) "Do you know what the capital of Peru is?" where there is not a complete proposition that can be judged true or false. As Dokic points out, "the addressee can answer Q4 (by saying 'yes [I know the answer]') without being in a position to answer Q3 (by saying 'Lima'). In fact, she can answer Q4 without having any city in mind" (Dokic 2012: 304; bracketed phrase added). That is, the subject is able to answer a question about her mind without being in a position to answer the corresponding question about the world. In cases like this, one can answer a question about one's mind without being able to answer a question about the external world. In what follows, I will call such cases of self-ascription "non-transparent", because the subject is able to self-ascribe a mental state without having access to its intentional content, such as when a subject self-ascribes the capacity to solve a reasoning problem before solving it, or when she self-ascribes the capacity to recall some information from her memory before actually retrieving it:

> One can say that success in doing a cognitive task hangs on possessing beliefs or pieces of information that are not immediately transparent in the subject's situation. For instance, solving the bat-and-ball puzzle is a cognitive task because it requires that one work out the correct answer (even at the implicit level), which is not immediately given in the puzzle itself (Dokic 2012: 316).

So, how does one perform this trick? In other words, how can a subject make a self-ascription about her own mental capacities and mental states if

she has no access to their intentional content? Dokic's answer appeals to E-feelings. One can answer Q4 (without having an answer to Q3) by relying in one's FOK, i.e., the affective experience that points to one's capacity of answering the question before one is actually able to do it. The classic example of this phenomenon is the TOT. In the TOT state, a subject is confronted with a question to which she does not have an immediate answer, although she *feels* that she knows the answer. The subject has not retrieved the answer to the question, yet she self-ascribes the proposition "I know the answer" based on a metacognitive feeling (Schwartz 2002; Schwartz & Metcalfe 2010). Asher Koriat presents this case in an insightful way:

> Although clearly the TOT represents a state of awareness, the awareness is about something that the person does not (yet) know (…). In a sense, the TOT phenomenon illustrates *a dissociation between subjective and objective indexes of knowing*—between the subjective conviction that one "knows" the sought-after name, and the actual inability to produce it (Koriat 2000: 151; italics added).

As Koriat remarks, to acknowledge the existence of E-feelings among the constituents of the mind stresses the diversity of ways one comes to know one's own mental states. One may realize that one knows the answer to a question or the solution to a problem by retrieving it (as happens in the ascent routine), but one can also get to know this via a subjective E-feeling. The content of one's mental states can be opaque (as in the TOT case), or it can simply be absent (as when one feels that one is forgetting something); in these cases it is by the way one feels that one gets to know about it. Moreover, the content of a mental state itself does not tell one anything about its correctness or wrongness. It is by the way one feels about that content that one knows how to *epistemically* stand towards it (Mangan 1993, 2000; see also Dokic this volume).

In the same paper, Dokic (2012) proposes an embodied account of E-feelings that he calls "the water diviner model". According to his view, these feelings are "first and foremost bodily experiences" (Dokic 2012: 307), i.e., experiences about bodily states. They are "diffuse affective states registering internal physiological conditions and events". But in the same way that water diviner's bodily sensations reliably co-vary with physical conditions, namely the presence of underground water, E-feelings reliably co-vary with mental conditions. For example, the feeling of knowing – which is mainly a bodily feeling according to this view – reliably co-varies with the fact that a given piece of information can be retrieved from the subject's memory. This is why, according to Dokic, self-ascribing mental states based on such E-feelings leads to self-knowledge.

It is worth highlighting that, although Dokic stresses the fact that E-feelings are "first and foremost bodily experiences", i.e., they have *intrinsic* bodily content, he also acknowledges that E-feelings have *derived* intentional content that goes beyond the body. He calls his own account of the derived content of E-feelings "the competence view". According to this view, the derived content of E-feelings is about one's own cognitive competence at a given cognitive task and has the form "I can [or cannot] do this" (or "this can [or cannot] be done") (Dokic 2012: 316; bracketed text added), where "this" refers to the cognitive task at hand.

Dokic's criticism of the ascent routine and his account of E-feelings provides a novel perspective on the way one comes to know mental capacities, mental processes, and dispositions. However, his account falls short when it comes to explaining the relation between E-feelings and self-ascriptions. According to Dokic's account, learned heuristics mediate the transition from an E-feeling to a self-ascription or second-order judgment so that we are able to "move spontaneously from our feelings to judgments concerning the task at hand [… and/or], her own mental states" (Dokic 2012: 308, 309, bracketed text added). However, given the paucity of the intentional content of E-feelings (which, on Dokic's view, points only to the competence of the subject), it remains unexplained how a subject can derive a self-ascription about a particular mental state from a mere feeling. In other words, given that E-feelings are bodily sensations and given that bodily sensations take place at many times and in many contexts, there are many possible ways of interpreting those bodily feelings; so we need to describe the mental mechanism that does the job and specify more fully the "learned heuristics" or the "epistemic rules" (see §3) that it uses to infer a self-ascription from an E-feeling. In a nutshell, there is an explanatory gap between the E-feeling one has and the judgment that one forms about one's mind. The following sections will propose a model that aims to close this gap.

3. *A Two-Level Model of Non-Transparent Self-Ascriptions*

Dokic's considerations on self-ascription and E-feelings direct our attention to an important, yet often neglected, question that any model of self-ascriptions and self-knowledge should answer: how can a subject make a non-transparent self-ascription about her own mental capacities, mental processes and dispositions based only on a mere E-feeling? As we saw in previous section, this is not an easy task. To this end, I will propose a two-level model of metacognition (Koriat 2000; Thompson 2009; Koriat & Ackerman 2010; Arango-Muñoz 2011) that links E-feelings with the mindreading capacity.

The two-level model of metacognition that I want to defend claims that, in cases of non-transparent self-ascriptions, two elements are at play:

> Low-Level Metacognition: E-feelings are elicited by metacognitive monitoring mechanisms according to heuristics and the relevant cognitive task.
>
> High-Level Metacognition: The mindreading mechanism interprets the E-feelings according to their valence and the context, and then produces self-ascriptions of mental states.

The idea of two levels of metacognition is similar to Nichols and Stich's model of self-knowledge (2003) which also includes two independent – but interactive – mechanisms: a monitoring mechanism and a mindreading mechanism. The main difference between their account and mine concerns the way we understand the monitoring mechanism. Whereas, for them, the monitoring mechanism is a subpersonal mechanism (or a set of mental mechanisms) that scans mental states in a "Belief Box" and whose main function is to detect mental states and their contents, for me (following the psychological tradition of metacognition research; e.g., Reder 1987, 1996; Koriat 1993, 2000), the main function of low-level metacognition is to elicit E-feelings and control mental action based on cues and heuristics (see "Epistemic feelings, epistemic emotions: Review and introduction to the focus section"). In other words, the monitoring mechanism does not actually scan mental states.

3.1. Low-Level Metacognition: E-Feelings as Signals of Absent Objects

The two-level model of metacognition claims that, in cases where the object of one's mental attitude is not transparent, as in answering Q4 by saying 'yes, I know the answer' when no answer is yet available to the subject, two elements are responsible for the formation of self-ascriptions: E-feelings elicited by a monitoring mechanism according to certain heuristics (and the cognitive task that the subject confronts), and the mindreading mechanism. In this section, I will compare Dokic's view of E-feelings with my own view.

Following Dokic's account, I understand E-feelings as involving two ingredients: a bodily component, and an intentional content that points towards mental conditions (see also Goldie 2002). However, there are two important differences between our views. First, according to Dokic, the intentional content of E-feelings has the form "I can [or cannot] do this" (or "this can [or cannot] be done") (Dokic 2012: 316; bracketed text added). Though I agree with the general idea, I consider that it is more accurate to describe their content as representing "value by means of positive or negative valence" (Arango-Muñoz

2013a: 4). That is, by feeling a positive or negative affect, the subject becomes aware of whether she can or cannot do it, i.e., the implicit metacognitive evaluations of the cognitive task.[2] In this respect, Dokic and I both take the content of the feeling to refer to the competence of the subject, but it is subjectively experienced as a positive or negative affect directed towards a mental state, process, or disposition.

The second difference between our views is that I accept – but Dokic does not – that E-feelings have richer intentional content; that is, that they can also refer to the specific piece of information that the subject is aiming for, though this intentional content is non-transparent[3] (see also Mangan 1993, 2000, 2001; Norman, Price and Duff 2010). For example, in the case of the TOT, the content is (A) a positive affect that points to the possibility of retrieving whatever the subject wants to retrieve (in Dokic terms "I can retrieve it" or "this can be retrieved"), and (B) the specific intentional content that the subject is looking for. The reason for conceiving the content in this twofold way is that subjects are often able to discriminate the target of their feeling among different possible objects. In many cases (e.g., FOK, TOT, and FOF), the intentional content of the feeling (B) is somehow absent – because it has not yet been retrieved, in the FOK and TOT, or it has been lost, in the case of FOF. As Koriat remarks, E-feelings are often caused by contentless cues and heuristics[4] (Koriat 2000), but the feelings themselves can also (and often do) condense implicit knowledge or information, as Mangan (1993, 2000, 2001) and Norman *et al.* (2010) have proposed.

This becomes even clearer when one consider cases of E-feelings where the subject is aware of their intentional content. For example, the feeling of rightness, the feeling of certainty, the feeling of uncertainty, and the feeling of error, among others, point towards an explicit content ("x is bigger than y", "Lima is the capital of Peru", etc.) and evaluate the content as correct or incorrect. Given these cases, it is clear that the subject feels a positive or negative affect towards a given content, not only about her mental competence.

In a nutshell, in my view, E-feelings are implicit assessments of value – positive or negative – concerning a given cognitive or mental task. They indicate

[2] This does not mean that she is aware of the metacognitive evaluations as such. She is only aware of the positive or negative valence concerning the cognitive task she is confronting.

[3] To say that it is non-transparent is another way of saying that it is unconscious but present. I claim that the content is present because subjects' behavior exhibits intelligence that would not be possible without intentional content (see Arango-Muñoz 2013a).

[4] For example, E-feelings are triggered by sensory cues such as the frequency of encounter with a stimulus (Reder 1996; Paynter, Reder & Kieffaber 2009), its perceptual fluency (Whittlesea 1993; Whittlesea & Williams 2001), or the fluency of the processing of a stimulus (Koriat 2000).

or point towards a non-transparent and/or transparent object and, at the same time, motivate certain types of bodily and/or mental behavior (Arango-Muñoz 2013a, 2013b).

3.2. High-Level Metacognition: Mindreading mechanism interprets E-feelings

According to some theorists, the mindreading mechanism is an inferential mechanism provided with (A) a theory of mind (TOM) and (B) a set of psychological concepts that, given a behavioral or perceptual input, generates a self-ascription as output (Carruthers 2009, 2011; Gopnik 1993; Wegner 2002; Bogdan 2010; Flavell 2000). On the one hand, (A) the theory of mind can be understood as a broad sketch of how the mind works and how it causes behavior. Subjects rely on such a theory to understand others' behavior. On the other hand, (B) psychological concepts are concepts referring to propositional attitudes such as perceptions, feelings, intentions, knowledge, beliefs and expectations, among others. Subjects use and combine this type of concept to ascribe mental states to other people and to themselves. The mindreading mechanism has evolved to interpret others' behaviors, but it can also be turned upon oneself (Carruthers 2009, 2011).[5]

So, the idea is that in cases of non-transparent self-ascription, the subject relies on E-feelings in order to self-ascribe mental properties using her mindreading capacity. The reason why the subject has to resort to her mindreading capacity is that, in non-transparent cases of self-ascription (e.g., when the mental state in question is a disposition, a non-occurrent belief or emotion, or a mental capacity), she is in a epistemic position with respect to her own mind *similar* to the position that she is in when she evaluates and interprets others' minds. That is, she only has access to cues and indirect evidence about her own mental states. In these cases, the content is opaque or absent, and therefore she self-ascribes a mental state based only on her E-feelings without having any access to the intentional content of the ascribed mental state.

Thus, in non-transparent self-ascription, the mindreading capacity takes as input the E-feeling elicited by low-level metacognitive monitoring, contextual factors and knowledge (such as the kind of cognitive task the subject is confronting and the kind of possible mental states related to that task), and then generates a positive or negative self-ascription according to the valence of the feeling. A positive E-feeling motivates a positive self-ascription concerning a

[5] I do not want to commit myself to the claim that subjects *only* understand others' minds by means of a theory. There may be cases where they resort to mental simulations. However, it seems unlikely that they use mental simulations to understand and get self-knowledge of their own mind, as is acknowledged even by simulation theorists (e.g., Goldman 2006).

mental capacity, whereas a negative E-feeling motivates a negative self-ascription. So, if a subject is confronted with a question to which she does not have an immediate answer, she may rely on her intuitive E-feeling to generate a self-ascription concerning her knowledge or ignorance of the answer (see Figure 1). For instance, if a subject had a negative feeling when confronted with a memory task, the mindreading mechanism might self-ascribe the concept of uncertainty or forgetting (see below §3).

In this way, the puzzle of how a subject self-ascribes a non-transparent mental property, one to which she does not have yet conscious access, is solved.

4. *E-Feelings, Mental Action and Self-Ascriptions*

The previous considerations broadly describe how the mindreading mechanism forms non-transparent self-ascriptions based on E-feelings. However, describing the mechanism is not sufficient; one should also describe the "learned heuristics" (as Dokic [2012] called them) or "epistemic rules" (as I call them, following Byrne [2005, 2011a]) that govern the functioning of the mindreading mechanism.[6] This section will provide a deeper characterization of the relation among E-feelings, mental action and self-ascriptions of mental states based on epistemic rules.

4.1. Rule-following considerations

Some philosophers have found appealing the idea that rule-following considerations can explain the formation of self-ascriptions (Byrne 2005, 2011a, 2011b, 2011c; Peacocke 2008: 206; Goldie 2002). According to this strategy, a subject follows a more or less implicit rule of the sort "if P, then believe that you believe P" (Byrne 2005, 2011) whenever she makes a self-ascription of a mental state. Although I broadly follow Byrne's idea of epistemic rules for self-ascriptions, the rules that I posit here are quite different in character: they are neither transparent nor neutral, in contrast to the rules proposed by him (see §3.3). Moreover, in my view they are not only for reasoning about one's own mental states, but also for directing mental action.

In the following, I will talk *as if* subject followed rules for self-ascribing mental states, but this should be taken just as an approximate metaphor that

[6] Although I'm adopting Byrne's concept, there are important difference between his view and my own view. The main disagreement is that he claims that epistemic rules are applied and followed by subjects themselves, whereas I claim that they are applied or followed by the mindreading mechanism – something Byrne wouldn't accept. But this disagreement is not really relevant here, since we are trying to explain different kinds of self-ascription.

describes the cognitive processes that guide subjects' behavior and self-ascriptions. In other words, even if the subject does not actually follow a rule (after all, she does not have a clear understanding of the rule, and she is not even able to articulate it), her behavior can be characterized as embodying or displaying some epistemic rules or heuristics. That is, although these rules and heuristics can be inferred by observing subjects' behavior and are useful for predicting it, more research is needed before we can commit to their actual existence.

To begin with, let us assume that a subject employs two different epistemic rules when confronted with E-feelings; she does not need to employ both rules on every occasion. A rule can be understood as a reference that one (or the cognitive system – the mindreading mechanism in this particular case) invokes each time that one needs to determine something. In simpler terms, when confronted with an E-feeling, a subject needs to know what to do, and epistemic rules tell her roughly what to do; as we will see, two rules are the links among E-feelings, mental action and self-ascriptions. On the one hand, *epistemic rules for action* (R1) are those one invokes for determining which mental action to perform given an E-feeling. On the other hand, *epistemic rules for self-ascription* (R2) are those we invoke for determining what to believe about one's self, one's mental states, and one's dispositions given an E-feeling.

Let me briefly introduce both rules. The next three sections will analyze and explain them in detail.

> Epistemic rule for mental action (R1): If E-feeling Y arises, then do mental action X.
> Example: If FOK, then "try to remember".
> Epistemic rule for self-ascriptions (R2): If E-feeling Y arises, then you are entitled to form a second-order belief about your mental activities, dispositions or capacities according to the valence of your E-feeling: a positive judgment if a positive E-feeling and vice versa.
> Example: If FOK, then form the belief "I can remember".

4.2. Epistemic rule for mental action (R1)

At first glance, R1 seems to be a *practical* rather than an epistemic rule since the term "epistemic" is often exclusively associated with belief formation, as in R2. R1 is about what to do instead of what to believe. However, R1 should be considered epistemic since any mental action, which can be roughly defined as a directed change in the mind (see Proust 2001, 2009a), necessarily entails

epistemic changes, e.g., perceptual or informational improvements and/or relapses.

Some theorists may criticize this account by suggesting that the action in the former clause implies the second-order belief suggested by the latter: "to do X mental action" or "to try to remember" implies the second-order belief that "you believe that you can do X (or remember)", since you cannot try to remember without believing that you can do it. But, as I will show, R1 does not imply R2, and applying R1 does not require applying R2 (see §3.4). The first argument for the dissociation between R1 and R2 is the phenomenological observation that *prima facie* one performs many mental actions such as remembering, calculating, or reasoning just by attending to those actions, and without any need to make second order judgments about one's self (Vierkant 2012; Proust 2007, 2012). Furthermore, the fact that some non-human animals and infants may grasp R1 but do not reach R2 also supports the distinction. These agents are able to engage in information-acquiring acts although they are unable to form metarepresentations, *stricto sensu* (Carruthers 2008, 2009, 2011; Bermúdez 2009), i.e., second-order thoughts that involves the deployment of psychological concepts and a self-concept. For example, rhesus monkey and young children are able to monitor and control their cognitive performance in memory and perception tasks, allowing them to attain an accurate performance similar to human behavior in perception and memory (Hampton 2001; Balcomb & Gerken 2008; see Smith 2009 for a review). A plausible explanation of their ability to monitor and control their mental capacities without resorting to metarepresentation is that their behavior is guided by metacognitive feelings (Proust 2009b; Arango-Muñoz 2011; Dokic 2012). Thus, R1 concerns the relationship between E-feelings and mental action, and can be possessed and applied even by beings lacking the possibility of forming second-order beliefs about themselves, i.e., introspective self-ascriptions.

Thus, according to R1, E-feelings can motivate and guide action *directly* (i.e., without passing through the reflective process of self-ascription), and the subject can cite them to account for a given action: "I have done X because I had Y feeling". Similarly, we can cite these kinds of mental states to interpret the behavior of others: "The subject acted in a given way because she had that E-feeling".

4.3. Epistemic rule for self-ascriptions (R2)

R2 determines the relation between feelings and introspective self-ascriptions or second-order beliefs; in particular, it determines when to form a second-order belief about oneself. In other words, it dictates the formation

of self-ascriptions based on E-feelings. Based on the particular experience of easiness or difficulty of carrying out a cognitive task, the subject constructs a semantically articulated self-ascription that involves the deployment of psychological concepts (KNOWING, INTENDING, etc.) and the concept of the self. Thus, an E-feeling gives prima facie reason to a subject to form the self-ascription that she has or lacks a piece of information (Peacocke 2008; Proust 2009a).

Notice that, in contrast to ascent routine (discussed in §1), the content self-ascribed by following R2 is not already present in the antecedent part of the conditional rule, i.e., it is not transparent. The conditional rule used in the ascent routine says "If P, believe that you believe P" (Byrne 2005, 2011a). So, the self-ascribed content is typically already present in the reasoning process; this is why it can be considered to be a transparent procedure. R2, in contrast,

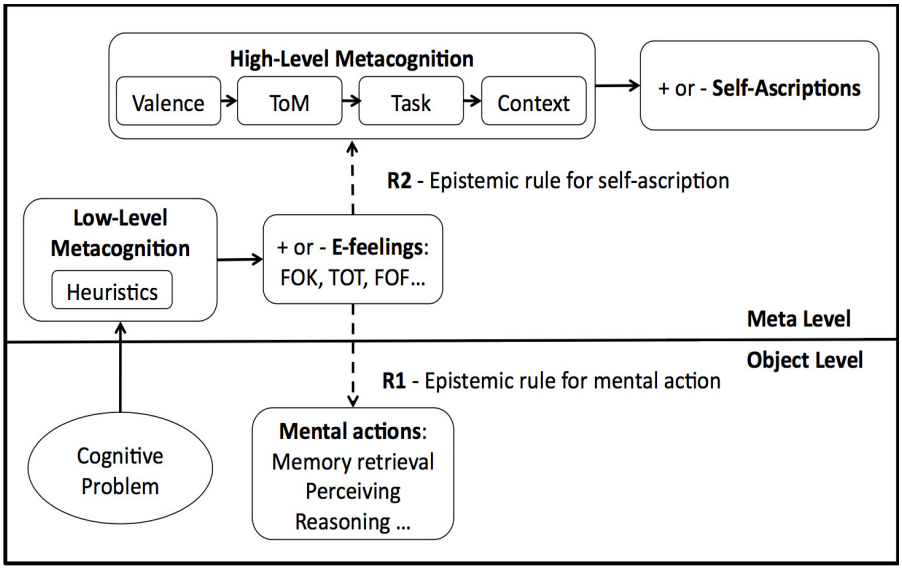

Figure 1: Schema of the two-level model of metacognition. Each time a subject is confronted with a cognitive problem, Low-Level Metacognition evaluates it based on some heuristics and elicits positive or negative E-feelings (see 2.1). There are two possible outcomes of E-feelings: Mentally acting (object level – below horizontal line) and thinking about one's mental states (meta-level – above horizontal line). The dotted line pointing downwards depicts the way E-feelings *directly* modulate mental actions (see §3.2); this can happen following R1, which is an *epistemic rule for mental action*. The dotted line pointing upwards depicts the way E-feelings influence high-level metacognition (see §3.3): self-ascriptions based on E-feelings are produced by high-level metacognition following R2, which is an *epistemic rule for self-ascription*.

derives the content of the self-ascription from the valence of the E-feeling, the context, and related information about the task. Thus, the procedure of forming self-ascriptions based on R2 has the advantage of providing the subject with *new* information that was not already contained in subject's representational state, as in the ascent routine.

Another feature of R2 is that it is not "neutral", as are the rules proposed by Byrne (2005, 2011a). According to Byrne's (2005, 2011a) account of self-knowledge, an epistemic rule is neutral if "the antecedent condition C of an epistemic rule R is not specified in terms of the rule follower's mental states" (Byrne 2005: 94; 2011a). In contrast, according to R2, the antecedent part of the conditional is specified in terms of rule follower's mental states, mainly by E-feelings and beliefs about the context and the task.[7] Because of these features, R2 has a "relative disadvantage" with respect to Byrne's model of self-ascriptions: self-ascriptions following R2 are not self-verifying, in contrast to those provided by Byrne's version of the ascent routine. Because of the interpretative character of R2, it leaves open the possibility of misattribution and introspective error. This would be a disadvantage if one wanted to hold that introspection is infallible. But, in my view, this feature of R2 is in fact an advantage because it helps to explain misattribution of mental states and confabulations, two common phenomena in introspective reports (see Carruthers 2009, 2011).

4.4. Differences between R1 and R2

A key difference between the two kinds of epistemic rules is the way we determine their quality: R1 is a good rule if it leads to successful mental actions, whereas R2 is a good rule if it leads to the formation of reliable beliefs about one's own mental dispositions. Moreover, the first is an imperative to act mentally, to do something, whereas the second is merely a suggestion to form a second-order belief, which can be dissociated from action. Thus, R2 permits a reflective distance that is not allowed by R1; the subject can contemplate the content of her self-ascription without engaging in action. This suggests that the content of R1 is imperative, whereas the content of R2 is merely descriptive or propositional (see Boghossian 2008).

What counts as possession and application of R1 is successfully dealing with mental uncertainty, that is, the ability to accurately predict and retrospectively

[7] Dokic proposes (personal communication) that we could preserve the neutrality of rules by granting that E-feelings are mainly bodily experiences: "If I am in bodily state B, then believe that…". Although I agree that E-feelings have a bodily component, I don't accept the identification of the former with the latter; E-feelings have properties that bodily experiences lack.

correct cognitive outputs. This is a recurrent activity that involves a subject's sensitivity to her mental activity and her capacity to control it: an ability to exploit E-feelings (also called "mental affordances" by Proust 2009b, 2007). In contrast, what counts as the possession and application of R2 is the explicit formation of conceptual self-ascriptions that may be true or false. This is an occasional act of intellect and conceptual understanding that may facilitate the correction of behavior in terms of social rules or theories of how the mind works. A subject would not be able to check whether she is mentally acting according to a given social rule unless she is able to make a self-ascription that allows her to contrast what she is doing with what the rule suggests (Flavell 1998).

E-feelings need not cause both action and second-order belief every time they occur. The subject may act without forming the second-order belief, and she may form a second-order belief without acting. The first case is illustrated by our daily behavior: we perform many mental actions, we remember, calculate, reason, and we execute all of these actions in a very precise and controlled way guided by E-feelings without making second-order judgments about each action (Vierkant 2012; Proust 2007, 2012; Arango-Muñoz 2013a). That may be one of the reasons why we are often unaware of how we carry out all the mental processes we routinely carry out. The second case is illustrated by cases in which an E-feeling motivates a second-order belief but we fail to act on it. A FOF makes a subject believe that she is forgetting something, but the subject, due to hurry, stress, or simply neglect, may be unable to do anything about it: she fails to check what she is forgetting.

5. *Concluding remarks*

As I said at the outset, we often mentally act guided by our E-feelings, such as when we check what we just saw based on a feeling of visual uncertainty; but we can also start thinking about our own perceptual states and capacities based on the same E-feeling. E-feelings are phenomenal experiences that point towards mental capacities, processes and dispositions of the subject such as knowledge, ignorance, or uncertainty. The main aim of this paper was to clarify the relation between E-feelings, mental actions and self-ascriptions. As I have shown, mentally acting and/or thinking about one's mental states are two possible outcomes of E-feelings, but they need not always go together: sometimes we mentally act without thinking about, and sometimes we think about our cognitive processes without acting upon them. The take-home idea, then, is that, in non-transparent cases, E-feelings guide mental action and/or self-ascriptions. In the context of

self-knowledge, these remarks suggest the existence of other means, different form the ascent routine, of forming self-ascriptions.

Acknowledgements

I would like to thank Tobias Schlicht, Eric Schwitzgebel, Jérôme Dokic, Kirk Michaelian, and Kevin Reuter for their comments and suggestions on previous drafts. I would especially like to thank Kateryna Samoilova for her attentive reading of the paper, her comments and criticisms, and especially for her encouragement to improve it.

References

Arango-Muñoz, Santiago, 2011, "Two levels of Metacognition", in *Philosophia: Philosophical Quarterly of Israel*, 39.1: 71-82; doi: 10.1007/s11406-010-9279-0.

Arango-Muñoz, Santiago, 2013a, "The nature of Epistemic Feelings", in *Philosophical Psychology*. 39.1: 1-19); doi: 10.1080/09515089.2012.732002.

Arango-Muñoz, Santiago, 2013b, "Scaffolded Memory and Metacognitive Feelings", in *Review of Philosophy and Psychology*, 4.1: 135-152); doi: 10.1007/s13164-012-0124-1.

Arango-Muñoz, Santiago, Ana Lucia, Fernández-Cruz, and Kirsten, Volz, (forthcoming), *Monitoring One's Own Errors in the Number Bisection Task. Center for Integrative Neuroscience*, Tübingen Universität, Germany.

Balcomb, Frances K., & Gerken, LouAnn, 2008, "Three-year-old Children Can Access their own Memory to Guide Responses on a Visual Matching Task", in *Developmental Science*, 11, 5: 750-760.

Bermúdez, Jorge Luis, 2009, "Mindreading in the Animal Kingdom", in Lurz, R. (ed.), *The Philosophy of Animal Minds*, Cambridge University Press, Cambridge.

Bogdan, Radu, 2010, *Our Own Mind, Sociocultural basis for Self-Consciousness*, The MIT Press, Cambridge, MA.

Boghossian, Paul, 2008, "Epistemic Rules", in *The Journal of Philosophy*, 105, 9: 472-500; doi: 0022-362X/08/0000/001-029.

Brewer, William F., and Cristina, Sampaio, 2012, "The Metamemory Approach to Confidence: A Test Using Semantic Memory", in *Journal of Memory and Language*, 67: 59-77.

Byrne, Alex, 2005, "Introspection", in *Philosophical Topics*, 33, 1: 79-104. Byrne, A., 2011a, "Knowing that I'm Thinking", in Hatzimoysis (ed.), *Self knowledge*, Oxford University Press, Oxford.

Byrne, Alex, 2011b, "Knowing What I Want", in Liu, J. and perry, J. (eds.), *Consciousness and the Self: new Essays*, Cambridge University Press, Cambridge.

Byrne, Alex, 2011c, "Transparency, Belief, Intention", in *Proceedings of the Aristotelian Society*, Supplementary volume, 85: 201-221.

Carruthers, Peter, 2008, "Meta-Cognition in Animals: a Sceptical Look", in *Mind and Language*, 23: 58-89.

Carruthers, Peter, 2009, "How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition", in *Behavioral and Brain Sciences*, 32: 1-18.

Carruthers, Peter, 2011, *The Opacity of Mind, an Integrative Theory of Self-Knowledge*, Oxford University Press, Oxford.

de Sousa, Ronald, 2008, "Epistemic feelings", in Georg Brun, Ulvi Doğuoğlu, and Dominque Kuenzle (eds.), *Epistemology and Emotions*, Ashgate, Hampshire: 185-204.

Dokic, Jérôme, 2012, "Seeds of Self-Knowledge: Noetic Feelings and Metacognition", in Michael J. Beran, Johannes L. Brandl, Josef Perner, and Joëlle Proust (eds.), *Foundations of Metacognition*, Oxford University Press, Oxford: 302-321.

Evans, Gareth, 1982, *The Varieties of Reference*, Oxford University Press, Oxford.

Flavell, John H., 1998, "The Mind Has a Mind of Its Own: Developing Knowledge About Mental Uncontrollability", in *Cognitive Development*, 13: 127-138.

Flavell, John H., 2000, "Development of Children's Knowledge About the Mental World", in *International Journal of Behavioral Development*, 24: 15-23.

Goldie, Peter, 2002, "Emotion, Feelings and Intentionality", in *Phenomenology and the Cognitive Sciences*, 1: 235-254.

Goldman, Alvin, 2006, *Simulating Minds, The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford University Press, Oxford.

Gopnik, Alison, 1993, "How We Know Our Minds: the Illusion of First-Person Knowledge of Intentionality", in *Behavioral and Brain Sciences*, 16(1-15), 90-101.

Gordon, Robert M., 1995, "Simulation Without Introspection or Inference from Me to You", in Davies, M. & Stone, T. (eds.), *Mental simulation: Evaluations and applications*, Blackwell, Oxford.

Gordon, Robert M., 2007, "Ascent Routines for Propositional Attitudes", in *Synthese*, 159: 151-165; doi: 10.1007/s11229-007-9202-9.

Halamish, Vered, Shannon, McGillivray, and Alan D., Castel, 2011, "Monitoring One's Own Forgetting in Younger and Older Adults", in *Psychology and Aging*, 26: 631-635.

Koriat, Asher, 1993, "How do we Know that we Know? The Accessibility Model of the Feeling of Knowing", in *Psychological Review*, 100: 609-639.

Koriat, Asher, 2000, "The Feeling of Knowing: Some Metatheoretical Implications for Consciousness and Control", in *Consciousness and Cognition*, 9: 149-171.

Koriat, Asher, 2008, "Subjective Confidence in One's Answers: The Consensuality Principle", in *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34.4: 945-959.

Koriat, Asher, & Ackerman, Rakefet, 2010, "Metacognition and Mindreading: Judgments of Learning for Self and Other During Self-Paced Study", in *Consciousness and Cognition*, 19: 251-264; doi: 10.1016/j.concog.2009.12.010.

Mangan, Bruce, 1993, "Taking Phenomenology Seriously: The 'Fringe' and its Implications for Cognitive Research", in *Consciousness and Cognition*, 2: 89-108.

Mangan, Bruce, 2000, "What Feeling Is the 'Feeling of Knowing?'", in *Consciousness and Cognition*, 9: 538-544; doi:10.1006/ccog.2000.0488.

Mangan, Bruce, 2001, "Sensation's Ghost. The non-Sensory 'Fringe' of Consciousness", in *PSYCHE*, 7.18): http://psyche.cs.monash.edu.au/v7/psyche-7-18-mangan.html.

Moran, Richard, 2001, *Authority and Estrangement, An Essay On Self-Knowledge*, Princeton University Press, NJ.

Nichols, Shaun, & Stich, Stephen P., 2003, *Mindreading: an Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford University Press, Oxford.

Norman, Elisabeth, Mark C., Price, and Simon C., Duff, 2010, "Fringe consciousness: A useful Framework for Clarifying the Nature of Experience-Based Metacognitive Feelings", in Efklides, A. and Misailidi, p. (eds.), *Trends and Prospects in Metacognition Research*, Springer, New York: 63-80.

Paynter, Christopher A., Reder, Lynne M., and Kieffaber, Paul D., 2009, "Knowing We Know Before We Know: Erp Correlates of Initial Feeling-of-Knowing", in *Neuropsychologia*, 47: 796-803.

Reder, Lynne M., 1987, "Strategy Selection in Question Answering", in *Cognitive Psychol-ogy*, 19, 1: 90-138.

Reder, Lynne M., 1996, *Implicit Memory and Metacognition*, Lawrence Erlbaum Associates, Mahwah.

Peacocke, Christopher, 2008, *Truly Understood*, Oxford University Press, Oxford.

Proust, Joëlle, 2001, "A Plea for Mental Acts.", in *Synthese,* 129.1: 105-128.

Proust, Joëlle, 2007, "Metacognition and Metarepresentation: is a Self-Directed Theory of Mind a precondition for Metacognition?", in *Synthese*, 159.2: 271-295.

Proust, Joëlle, 2009a, "It There a Sense of Agency of Thought?", in O'Brien, L. and Soteriou, M. (eds.), *Mental actions and agency*, Oxford University Press, Oxford: 253-279.

Proust, Joëlle, 2009b, "The Representational Basis of Brute Metacognition: A Proposal", in lurz, r. (ed.), *The philosophy of animal minds*, Cambridge University Press, 165-183.

Proust, Joëlle, 2012, "Metacognition and Mindreading, one or two Functions", in Beran, M., Brandl, J., Perner, J. and Proust, J. (eds.), *Foundations of Metacognition*, Oxford University Press, Oxford: 234-251.

Proust, Joëlle, (forthcoming), *Philosophy of Metacognition, Mental Agency and Mental Awareness*, Oxford University Press, Oxford.

Schwartz, Bennett, 2002, *Tip-of-the-Tongue States, Phenomenology, Mechanisms and Lexical Retrieval*, Lawrence Erlbaum Associates, Publishers, London.

Thompson, Valerie A., 2009, "Dual-Process Theories: a Metacognitive Perspective", in Evans, J., and Frankish, K. (eds.), *In Two Minds: Dual Processes and Beyond*, Oxford University Press, Oxford.

Vierkant, Tillmann, 2012, "What Metarepresentation is For", in Beran, M., Brandl, J., Perner, J. and Proust, J. (eds.), *Foundations of Metacognition*, Oxford University Press, Oxford: 279-288.

Whittlesea, Bruce W. A., 1993, "Illusion of Familiarity", in *Journal of Experimental Psychology*, 19, 6: 1235-1253.

Whittlesea, Bruce W.A., and Lisa, Williams, 2001, "Source of the Feeling of Familiarity: The Discrepancy-Attribution Hypothesis", in *Journal of Experimental Psychology*, 26, 3: 547-565.