Metarepresentational Versus Control Theories of Metacognition

Santiago Arango Muñoz

Werner Reichardt Center for Integrative Neuroscience
Tuebingen Universität
Paul-Ehrlich-str. 17, 72076, Tuebingen
santiagoarangom@gmail.com

Abstract

It is still unclear what metacognition is. Two main theories about metacognition are reviewed, each of which claims to provide a better explanation of the phenomenon, while discrediting the other theory as inappropriate. My claim is that in order to do justice to the complex phenomenon of metacognition, we must distinguish two levels of this capacity. It can be shown that each of these theories has been trying to explain only one of the two levels and that, consequently, the conflict between them can be dissolved. Finally, I characterize each level and explain some of their interactions.

Introduction

Current discussions of metacognition have focused on questions like the following: What is the nature of metacognition? What is the function of this mental capacity? What is the content and epistemic status of metacognitive assessments? Which living beings are endowed with it? At the present stage of the discussion, two main theories have been proposed concerning the proper set of answers to these questions. One claims that metacognition is a metarepresentational capacity to *self-ascribe* mental states, whereas the other claims that it is mainly a capacity to *evaluate* our cognitive processes via a mental simulation of them.

My suggestion is that our metacognitive capacity can be understood as involving two different levels of complexity (as is also suggested by Koriat (2000)), each having a different structure, a different content and a different function within the cognitive architecture, and that each of the competing theories has been addressing a different level. Thus, in the end, the two theories can be shown to be compatible since they provide explanations of different levels of metacognition. Moreover, my argument also shows that any theory that aims to explain human metacognition should be able to account for both levels and their interactions.

The discussion will proceed as follows: section 1 introduces the two main theories of metacognition together with the empirical data that they cite in their favor. Section 2 develops the idea of two levels of metacognition and characterizes each level according to the *dual-process* theories of cognition. Finally, section 3 considers some interactions between the two levels and the question of how to individuate the levels.

1. Two Theories of Metacognition

1.1 Metarepresentational Theory of Metacognition

From the point of view of the metarepresentational theory, metacognition refers literally to "thinking about thinking", i.e., to the self-ascription of mental states carried out by forming a second order thought about a first order one, and more generally forming an n+1-order thought about an n-order thought. Such self-ascription depends on a more general mindreading capacity consisting in an inferential capacity to attribute mental states in order to interpret and rationalize other people's behavior. Thus metacognition in this sense is no more than "turning our mindreading capacities upon ourselves" (Carruthers 2009, 2006; Larkin 2010; Gopnik 1993; Bogdan 2001, 2005; Flavell 2004).

Mindreading requires the possession of mental concepts by the subject in order to apply them to other people and interpret their behavior. Mental concepts are concepts referring to propositional attitudes[1] such as perceptions, feelings, intentions, knowledge, beliefs and expectations, among others. Therefore, the necessary structure of metacognitive judgments is composed by: 1) a proposition (e.g. "it rains"), 2) a first-order attitude directed to that representation, such as believing or

[1] It is still a matter of hot debate whether all these mental attitudes are propositional; especially in the cases of perception, emotion and feeling. And it is also unclear if all the metarepresentational theorists hold a propositional view of attitudes. To my knowledge, within the group of metarepresentational theorists, at least Carruthers (2009c) is committed to this view. My own view is that the content of perception, emotion and feelings is nonconceptual and non-propositional, though I am not going to claim so here.

intending, denoted by a *mental concept*, and 3) a second-order attitude, namely a metacognitive judgment, directed to the first order attitude (2) and its proposition (1) (Proust 2007). In other words, the content of a second-order representation is necessarily constituted by the self-attribution of a mental concept together with a first-order representation:

[3] I believe that [2] I *KNOW (or PERCEIVE, BELIEVE, FEEL, ETC*) that [1] it rains.

It should be highlighted then that a necessary condition to form thoughts with this structure is to possess and be able to apply *mental concepts*. Thus, from an evolutionary point of view, only beings endowed with the capacity for mindreading would be able to form metacognitive judgments. However, empirical studies using the 'false belief test' have suggested that non-human animals lack the mindreading capacity since they are not able to attribute false beliefs to others (Bermúdez 2009; Povinelli 2000; Hare et al. 2000, 2001; Povinelli and Vonk 2003) and thus are also incapable of metacognition in this sense. From a developmental point of view, infants are only capable of self-ascriptions of mental attitudes after they have acquired a theory of mind that provides them with mastery of the relevant mental concepts (Wellman 1990; Baron-Cohen 1995; Gopnik and Melzoff 1997). Most of these theorists acknowledge that metacognition in this sense starts between the ages of 3 and 4, but others hold that it does not occur until the age of 5 or even later (see Bodgan 2001, 2005).

From an epistemological point of view, there should be *almost* no difference between the knowledge that a subject has about herself and her knowledge about others because both are based on the similar behavioral cues,[2] use the same conceptual resources to make inferences and are produced by the same cognitive mechanism. Both kinds of knowledge derive from mindreading and thus have the same epistemic status. For example, the following judgments: "He intends to listen to the teacher" and "I intend to listen to the teacher" are based on the observation of some bodily behavior (as turning the head in some direction) together with contextual factors and other cues, and are produced by the mindreading capacity using the same concepts and therefore have the same epistemic status.

[2] I say "almost" because Carruthers' later account of mindreading acknowledges that first person mindreading has access to some cues (such as visual, auditory and motor imagery and inner speech) that the third person mindreading does not have access to. This creates a quantitative difference but not a qualitative one: access in both cases is interpretative and thus its epistemic status remains the same.

1.2 Control Theory of Metacognition

The control view on metacognition claims that it is mainly a capacity to evaluate and control our cognitive processes and mental dispositions by means of mental simulation. In Joëlle Proust's words: "The aim is, rather, to evaluate one's present mental dispositions, endorse them, and form epistemic and conative commitments" (Proust 2009b). The main point is that this evaluative capacity does not follow from a theoretical capacity to meta-represent attitudes (mindreading), i.e., subjects do not need to form a second-order representation about their first order attitudes in order to evaluate and control them. As Pamela Hieronymi puts it: "The forming and revising of beliefs and intentions is not voluntary nor does it require the same kind of reflective distance or awareness" (Hieronymi 2009).[3] Thus, it does not require the possession of mental concepts, a theory of mind or mindreading. Control theorists speculate that such evaluative capacity derives from an off-line simulation of the cognitive process in question which permits predicting and adjusting future cognitive performance on a given task (Proust 2007, 2008, 2009a; Peacocke 2007, 2008, 2009), as happens in the case of bodily actions where the subject runs an off-line motor program to predict her future performance (Grush 2004).

Inspired by the psychological literature on human and non-human metacognition, the control theory holds that the postulated evaluative capacity can be understood as a capacity to monitor and control cognitive activities, such as remembering or perceiving, which is more primitive than mindreading. Empirical data supporting control theories come from the domains of experimental and animal psychology: 1) Different behavioral paradigms support the claim that *very often* humans *do not rely* on metarepresentations to control their cognitive activities (Koriat 2000, Reder 1996; Paynter, Reder, and Kieffaber 2009; Walsh and Anderson 2009). For example, Reder and Schunn (1996) have shown that subjects are able to assess and decide among possible cognitive strategies, e.g., whether they will be able either to remember or calculate a math problem, based on a feeling of knowing produced by the properties of the retrieval process and not by the content of the solution. And 2) some non-human animals that lack mental concepts and mindreading seem nonetheless to be endowed with the capacity to monitor and control their cognitive performance in memory and perception tasks, allowing them to have an accurate performance similar to human behavior in perception and memory (see Smith 2009 for a review). These findings seem to support a distinction between mindreading and metacognition and also suggest the idea that metacognition is a function exerted by means of a different

[3] Even if Hieronymi never speaks of metacognition as such, it seems to me that her theory of metal control can be understood as a control theory of metacognition, regardless of whether or not she uses the term and of whether or not we accept her account.

representational basis in both cases: it does not rely on conceptual representations in order to monitor and control (Proust 2009d).

2. Two Levels of Metacognition

Theorists of both sides have been accusing each other of misinterpreting the phenomenon and giving the wrong account of metacognition. While metarepresentation theorists accuse control theorists of putting too much weight on a sub-personal mechanism, a "gate-keeping mechanism" (Carruthers 2008, 2009a), control theorists accuse the former of over-intellectualizing a more basic phenomenon (Proust 2007, 2009c). My suggestion is that we can dissolve this conflict if we analyze our metacognitive capacity as comprising two different levels of complexity - each having a different structure, a different content and a different function within the cognitive architecture. Each of the competing theories has been addressing a different level. So, at the end there is no real disagreement between both theories because they are trying to explain different phenomena. Nevertheless, it is also worth highlighting that once the two levels are functioning, they interact and influence each other dynamically, though neither of the two theories has taken these interactions into account. Each of these two levels may be associated with one of the two cognitive systems proposed by *dual-process* theories of mind (Evans 2008, 2009; Thompson 2009). The high-level has a metarepresentational structure and it is associated with an analytic theory-based level belonging to system 2, whereas the low-level is a control structure, an experience-based level belonging to system 1.

2.1 High-level: Theory-Based Metacognition

I largely agree with Carruthers' (2009) characterization of the high-level. From his point of view, the subject self-ascribes some mental states, properties or capacities based on her beliefs about her cognitive capacities and some perceptual or quasi-perceptual cues (such as visual, auditory and motor imagery and inner speech). Thus, this is a self-interpretative level where the subject tries to make sense of the observation of her behavior, in other words to rationalize and justify what she does (Evans and Over 1996; Thompson 2009). In order to do so, the subject needs a) some mental concepts that permit her to self-attribute mental states, b) a language in which to formulate her judgments (Bermúdez 2003; Evans 2009), and c) a theory of mind, understood as a set of beliefs concerning the functioning of the mind and allowing her to make inferences. In short, the subject needs a mindreading capacity to self-ascribe mental states. These characteristics seem to point to the kind of processing associated with what has been called *system 2* by cognitive psychologists. This is characterized as being slow, analytic, controlled and conscious (Evans 2008; Thompson 2009). Metarepresentational theories have been trying to explain this level of metacognition.

High-level metacognition thus deals with conceptual content and its main structure is inferential. Subjects interpret their behavior and make inferences thanks to a theory of mind they possess. Metarepresentational judgments are thus "drawn upon the *content* of domain-specific beliefs and knowledge that are retrieved from memory" (Koriat 2007: 19). A striking example of the interpretative nature of this level is revealed when subjects in an experiment are given an alternative explanation of the origin of their feelings of familiarity that consequently prevents them from relying on those feelings because of their unreliability. Based on this *new theory*, subjects will be less likely to self-attribute knowledge or memories and to rely on the fluency of retrieval as a cue, as they did before they were given the *new theory* (Schwarz and Vaughn 2002; Sanna and Schwarz 2003). In this respect, I move away from Carruthers' view (2009c): metacognitive beliefs and theories are not just "faux-thoughts", they do play an important role in the production of behaviour as these experiments show (see section 3.1).

Given that the main cognitive function of high-level metacognition is interpretive (since it developed in order to interpret others' behavior), it follows that high-level metacognition should have co-evolved with mindreading (Bodgan 2001; Jacob 2005) and subjects may often be wrong in their self-interpretative judgments about their own propositional attitudes and cognitive capacities: "People will (falsely) confabulate attributions of judgments and decisions to themselves in a wide range of circumstances, while being under the impression that they are introspecting" (Carruthers 2009a). On the one hand, people can be easily deluded concerning the content of their memory since "recognition or direct questioning can have 'contaminating' effects on memory" (Loftus 1989). On the other hand, people seem to hold false theories about their memory or their perception. For example, they often think that their visual field is like a TV screen or that their memory is like a hard disk.

What is striking, however, is that subjects do not rely on such theories or confabulations to control their cognitive behaviour. In other words, what they believe they do and what they actually do are not consistent. They do not behave as if their memory was perfect, they often make little mnemonic notes and consult their notebooks when they feel uncertain. They do not behave as if their perceptual field was a TV screen, they constantly scan the visual scenes in order to grasp all the relevant details. Koriat and Ackerman's (2010) recent study provides a interesting example of this inconsistency: subjects were presented with a learning task and they had to make judgements of learning (JOLs) concerning how well they had learned the items. Their overall behaviour was based on the *implicit* heuristic that the more study time they

invest in an item, the less likely they judge that they are to recall it later. However, when they were *explicitly* asked about the basis of their judgment, they reported different, inconsistent stories. These behavioural facts give rise to the idea that normal behaviour is not driven by high-level metacognition and metarepresentational beliefs (though it can sometimes be driven by it) but by low-level metacognition and epistemic feelings.[4] Along the same lines, many experiments on reasoning have pointed out that many of the reasoning biases, such as the belief bias or the myside bias, are actually caused by the subjects' propensity to accept uncritically (i.e., without an analysis or revision by S2) a heuristic response (Thompson 2009).

2.2 Low-level: Experience-Based Metacognition

Control theories focus on actual behavior rather than on the judgments that the subject forms about her mental events, dispositions and capacities. Proponents of this view have remarked that our cognitive behavior is often caused not by reflexive thinking but by emotional states (de Sousa 2008; Berkowitz 2000; Caver and Scheier 1990). For example, a subject confronted with a cognitive problem such as a multiplication task has to select a cognitive strategy to solve it. In the case of a familiar problem she has to choose either to remember the answer or to calculate. This decision does not seem to be based on a reflexive process considering all the possible alternatives and the pros and cons of each one (a maximizing and metarepresentational process which is computationally very demanding), but on a feeling that affords or makes salient one of the possible strategies (Walsh and Anderson 2009; Paynter, Reder and Kieffaber 2009; Kahneman 2003). Feelings are one kind of output of what cognitive psychologists have called *System 1* and which has been characterized as being fast, based on heuristics, mostly automatic and unconscious (Evans 2008).

But how is *system 1* supposed to generate such feelings? As mentioned earlier, control theorists speculate that such an evaluative capacity derives from an *off-line simulation* of the cognitive process in question that elicits feelings and emotions (Proust 2007, 2009a;[5] Peacocke 2007, 2009). I disagree with simulation theorists that the subject needs to simulate (on- or off-line) her mental processes in order to control them and that such a simulation is the origin of metacognitive feelings and emotions. This idea has been extracted from the simulation theory of understanding others' mental dispositions (Goldman 1993, 2006), which in turn has been coupled with the motor theory of action and mirror neuron theory (Meltzoff and Decety 2003; Gallese, 2003; Wolpert et al., 2003; Metzinger and Gallese, 2003), and then transferred *ad hoc* to explain metacognition. Though a thorough criticism of this strategy would require a whole paper, I will suggest some of the reasons why I doubt its adequacy. First, the theory from which the concept was extracted, simulation theory (Goldman 2006), has not yet provided definitive arguments for accepting simulation as the key to the acquisition and application of mental concepts (Jacob 2002; Jacob and Jeannerod 2005). Second, the idea of mental simulation, understood as running an off-line cognitive program in order to control one's own cognitive processes, seems less clear and plausible in the case of mental actions than in the case of bodily actions (Carruthers 2009b). Third, if the assessments delivered by metacognition were a product of a mental simulation, it seems that metacognition should not have to rely on perceptual cues, such as the frequency of the stimulus presentation (Reder 1996) or its perceptual fluency (Wittlesea 2001), as seems in fact to happen in metacognitive assessments (Koriat, 2000).

A modest solution is to postulate that metacognition is a "criticism system" or set of criticism systems, rather than a simulation system, provided with a list of heuristics in the form of conditionals prescribing the production of a given epistemic feeling for a given situation: if P-event then Q-feeling, if R-event then S-feeling, and so on (Minsky 2006). For example, in the case of memory, the criticism system would diagnose a good performance by a feeling of familiarity if the stimulus is perceptually fluent (Wittlesea 1993; Wittlesea and Williams, 2001). The feeling itself is metacognitive in the sense of being directed towards a mental disposition (knowledge, uncertainty, ignorance, etc.), but the content of the epistemic feeling that determines decision-making is non-conceptual and thus not metarepresentational. The feeling points or is directed to a mental property which is not necessarily within the gaze of the subject. For instance, a feeling of uncertainty points to a lack of knowledge or indicates that something is wrong with our perceptual or mnemonic activity, allowing a subsequent correction or improvement, without the need for an introspective effort by the subject. In a nutshell: low-level metacognition is the capacity of a being a) to entertain epistemic feelings that *nonconceptually* point to mental dispositions and b) to be able to exploit such feelings in order to control its cognitive activities (Proust 2009a).

[4] This might seem to commit me to Evans' (2008) and Carruthers' (2009c) thesis that behaviour is *always* driven by system 1. However, my claim is that it is so driven when system 1 can cope with the situation or problem; otherwise system 2 is activated, as Thompson claims (2009).

[5] "Controlled thinking should similarly proceed by triggering self-simulations based on prior performance", "Part of this activity is performed unconsciously, just as the preparation of a bodily action (which also involves simulation and evaluation) is shown to be performed outside awareness. It is hypothesized by scientists, rather than experienced by subjects, that a set of *comparators* allows one to *anticipate* how things normally develop for such and such a type of mental action (say: a directed remembering, or a planning)" (Proust 2009a).

3. Interactions, Mechanisms and Advantages of the Two-Level Account

3.1 Interactions Between the Two Levels

If my analysis is right, then the two competing theories of metacognition are indeed trying to explain two different levels of this metacognitive capacity instead of a single phenomenon. The high-level addressed by the metarepresentational theory is a rationalizing level where the subject uses concepts and theories to interpret her own behavior. The low-level is by default a controlling level where feelings induce the subject to adjust her cognitive activities in different ways without the need to engaging in second order thought. This obviously does not mean that these mechanisms cannot affect each other (as Carruthers suggests (2009c)). Let us consider three possible interactions:

1) On the one hand, one may have something like bottom-up causation: a feeling of error or uncertainty, for instance, elicited by the low-level may trigger an inferential process of verification carried out by the high-level (Thompson 2009). What is interesting in these cases is that feelings seem to emerge without the need for any higher-order belief concerning the cognitive processes and come to indicate a way of acting. The tip-of-the-tongue phenomenon is a good example. You unsuccessfully try to recall some piece of information together with the unpleasant feeling that you are in possession of such information. Notice that the higher-order belief that the unsuccessful retrieval should elicit is a negative one ("I don't know the information"), while the feeling points in the opposite direction. Then you might persist trying to recall it using your best known mnemonic strategies, or give up your attempt based on the metacognitive belief that when this happens it is better to distract yourself while waiting for the information to come to mind spontaneously.

2) On the other hand, one may have something like top-down causation in the following cases: making some concept salient, such as the concept of forgetting, may facilitate the production of some experiences in the subject, such as the feeling of forgetting something that would not be elicited otherwise (Koriat et al., 2004). This phenomenon suggests that epistemic feelings are not cognitively impenetrable, as perceptual experience seems to be (e.g., in the classic case of the Müller-Lyer illusion), that having some beliefs may trigger some particular experience concerning your cognitive activity. For example, thinking about the fallibility and unreliability of your memory may trigger a strong feeling of uncertainty that might even interfere with your normal cognitive performance (Pieschl 2010).

3) Another possibility is a top-down effect causing an inhibition of the low-level: when, e.g., the second-order belief that you are going to make a mistake activates the high-level as the controlling level and then inhibits low-level responses (Thompson 2009). The same case occurs in the example cited above, when subjects inhibit their propensity to rely on the feeling of familiarity after they are given a *theory* concerning the unreliability of the feeling (Schwarz and Vaughn 2002; Sanna and Schwarz 2003).

3.2 One or Two Mechanisms?

These considerations suggest a dichotomy: Either (A) these two levels are parts of the same mechanism or (B) they constitute two (or several) different mechanisms? To answer this question, we should establish an individuation criterion for cognitive mechanisms. Normally, the criterion used in cognitive science is to individuate a mental mechanism according to the mental operations that it carries out and its cognitive function (Bechtel 2008). The first possibility (A) would imply that high-level metacognition (mindreading) is grounded on low-level metacognition, as the simulation theorists hold (Peacocke 2007, 2009; Goldman 1993, 2006). And then we should expect a parallelism between judgments concerning self and the others.[6] This seems unlikely given the differences in function, content and structure that have been analyzed so far. Moreover, recent research in cognitive psychology has provided evidence of differences in the bases of metacognitive judgments about self and others (Koriat and Ackerman 2010). Even if we grant, for the sake of the argument, the simulationist claim that mindreading is specifically subserved by premotor cortex, this would be of not help since neuroscientific studies have pointed to the ventral-medial prefrontal cortex as the mechanism responsible for the monitoring and control of cognitive tasks (Pannu and Kaszniak 2005; Shimamura 1996, 2000; Simons and Spiers 2003).

The second possibility is that both levels are distinct mechanisms that have evolved in virtue of different evolutionary pressures to carry out different cognitive functions (Nichols and Stich 2003). This seems a more plausible option given the deep differences in structure, content and cognitive function that have just been described. However, even if I agree with Nichols and Stich (2003) in their claim that there are two or several different mechanisms, I deeply disagree with them in the main function they attribute to low-level metacognition. Whereas for them its main function is to provide introspective self-knowledge of our propositional attitudes, for me its main function is rather to control the cognitive activities without the need of meta-representation. Arguably this is not a kind of knowledge since the classic

[6] Carruthers' metarepresentational view of metacognition in terms of mindreading would belong to option (A) since it presuppose the existence of only one mechanism, but it cannot be considered to be in competition with the other accounts since it begins by rejecting the existence of low-level metacognition.

concept of knowledge involves true-justified propositions; at best it could be conceived as a practical one, a know-how rather than a know-that about the self. This know-how may be understood in terms of Sosa's animal knowledge, whereas the know-that would be a reflective one:

> "One has *animal knowledge* about one's environment, one's past, and one's own experiences [including epistemic feelings] if one's judgments and beliefs about these are direct responses to their impact – e.g. through perception or memory – with little or no benefit of reflection or understanding. One has *reflective knowledge* if one's judgment or belief manifests not only such direct response to the fact known but also understanding of its place in a wider whole that includes one's belief and knowledge of it and how these come about" (Sosa 1991: 240; note added).[7]

3.3 Some Advantages of the Two-Level Account

Some of the advantages of conceiving metacognition as involving two levels are the following:

a) By accepting and explaining low-level metacognition in terms of experiences and heuristics, we do justice to the empirical findings on animal and infant metacognition that demonstrate the presence of this capacity in non-linguistic or mindreading beings. Thus, we can avoid the counterintuitive consequence of the metarepresentational view that we have to endow animals with a theory of mind. This also allows us to understand that there is a low level of mental self-control, a *primitive mental agency*, that is non-conceptual, non-reflexive in terms of high-order thought and that does not require an inferential capacity. This nicely draws a line of continuity between human and non-human cognition: it is not the case that we humans have a *sui generis* form of cognition, but some of the capacities already present in animals became much more complex in the human mind.

b) By accepting and explaining high-level metacognition in conceptual and metarepresentational terms, we do justice to the empirical findings that show that only beings that posses and are able to apply mental concepts are able to self-attribute mental states. In other words, high-level metacognition is consistent with the developmental symmetry (Wellman 1990) of self/other mental ascriptions. It also allows us to explain the confabulation data.

c) A key question for my account is why we should consider these as two levels of the same capacity and not just two different and independent phenomena. My answer to this question is that actually they have evolved as different and independent mechanisms (or sets of mechanism) to carry out different cognitive functions, but from the moment in which they start to interact and influence each other, as I have sketched in section 3.1, we have to consider both as forming one complex capacity to evaluate and self-ascribe mental properties (from the low-to-high level) and self-ascribe and control (from the high-to-low level), that would not have emerged in the absence of either of the two levels.

[7] However, the claim is not that animal and human metacognition will remain identical if one subtracts one of them, but it acknowledges that once reflection is at play even low-level is modified by top-down causation (as shown 3.1), and therefore even human low-level metacognition has some particularities that animal metacognition lacks.

4. Conclusion

If my argument is right, the two competing theories are trying to explain two different levels of metacognition instead of a single phenomenon, and therefore the conflict is resolved. The high-level is a rationalizing level where the subject uses concepts and theories to interpret her own behavior. The low-level is a controlling level where the subject exploits epistemic feelings to adjust her cognitive activities. As I have tried to show, each has a different structure, a different content and a different function in the cognitive architecture. Moreover, my argument also shows that any theory that aims to explain human metacognition should be able to account for both levels and their interactions in order to produce a satisfactory account of this capacity.

Acknowledgments

This work was supported by the Center for Integrative Neuroscience (CIN), Tübingen, Germany. I received valuable comments from David Papineau, Albert Newen, Reinaldo Bernal and David Fajardo. I am particularly indebted to Fabián Bernache, Jérôme Dokic, Kirk Michaelian, Joëlle Proust and Tobias Schlicht for their thorough comments and corrections to this paper.

References

Baron-Cohen, S. 1995. *Mindblindness. An Essay on Autism and Theory of Mind*. Cambridge, Mass.: MIT Press.

Berkowitz, L. 2000. *Causes and Consequences of Feelings*. Cambridge, England: Cambridge University Press.

Bermúdez, J. L. 2003, *Thinking without words.* Oxford: Oxford University Press.

Bermúdez, J. L. 2009. Mindreading in the animal kingdom. In Lurz, R. ed. 2009. *The Philosophy of Animal Minds.* Cambridge: Cambridge University Press.

Bogdan, R. J. 2001. "Developing Mental Abilities by Representing Intentionality". *Synthese.* 129 (2): 233-258.

Bogdan, R. J. 2005. Why self-ascriptions are difficult and develop late? In Malle B. F. and Hodges S. D. eds. 2005.

Other minds. How humans bridge the divide between self and the others. New York: The Guilford Press.
Carruthers, P. 2009a. How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32: 1-18.
Carruthers, P. 2009b. Action-Awareness and the Active Mind. *Philosophical Papers* 38: 133-156
Carruthers, P. 2009c. Invertebrate concepts confront the generality constraint (and win). In Lurz R. ed.), 2009. *The Philosophy of Animal Minds.* Cambridge: Cambridge University Press.
Carruthers, P. 2008. Meta-cognition in animals: a skeptical look. *Mind and language* 23: 58-89.
Caver, C. S., and Scheier, M. F. 1998. *On the self-regulation of behavior.* New York: Cambridge University Press.
Dennett, D. 1991. *Consciousness explained.* Boston: Little Brown and Co.
de Sousa, R, 2008. Epistemic feelings. In: Brun, G., Doğuoğlu, U., and Kuenzle, D. eds. *Epistemology and emotions.* Hampshire: Ashgate Publishing Limited.
Evans, J. St. B. T. 2008. Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology* 59: 255–278.
Evans, J. St. B. T. 2009. How many dual-process theories do we need: One, two or many?. In Evans J. St. B. T., and Frankish, K. *In two minds: Dual processes and beyond*, Oxford, Oxford University Press.
Evans, J. St. B. T. and Over, D. E. 1996. *Rationality and reasoning.* Hove: Psychology Press.
Flavel, J. H. 2004. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly* 50: 274–290.
Gallese, V. 2003, The manifold nature of interpersonal relations: the quest for a common mechanism. In Frith, C., and Wolpert, D. eds. *The Neuroscience of Social Interaction* Oxford: Oxford University Press: 159-182.
Goldman, A. 2006. *Simulating minds: the philosophy, psychology and neuroscience of mind-reading.* Oxford: Oxford University Press.
Goldman, A. 1993. The psychology of folk psychology. *Behavioral and Brain Sciences* 16: 15–28.
Gopnik, A. 1993. How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16: 1-15, 90–101.
Gopnik, A., and Melzoff, A. N. 1997. *Words, thoughts, and theories*. MIT Press.
Grush, R. 2004. The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences* 27: 377-442.
Hare, B., Call, J., Agnetta, B. and Tomasello, M. 2000. Chimpanzees know what conspecifics do and do not see. *Animal Behavior* 59: 771–785.
Hare, B., Call, J., Agnetta, B., and Tomasello, M. 2001. Do chimpanzees know what conspecifics know? *Animal Behavior* 61: 139–51.
Hieronymi, P. 2009. Two kinds of agency. In O'Brien, L., and Soteriou, M. eds. 2009. *Mental actions and agency.* Oxford: Oxford University Press.
Jacob, P. 2002. Scopes and limits of mental simulation. In Dokic, J. and Proust, J. eds. 2002. *Simulation and knowledge of action.* Amsterdam: John Benjamins.
Jacob, P. 2005. First-person and third-person mindreading. In Giampieri-Deutsch, P. ed. *Psychoanalysis as an empirical, interdisciplinary science*. Vienna: Austrian Academy of Sciences Press.
Jacob, P. and Jeannerod, M. 2005. The motor theory of social cognition: a critic. *TRENDS in Cognitive Science* 9(1): 21-25.
Kahneman, D. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist* 58: 697-720.
Koriat, A. 2000. The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition* 9: 149-171.
Koriat, A., and Ackerman, R. 2010. Metacognition and mindreading: judgments of learning for self and other during self-paced study. *Consciousness and cognition*. doi:10.1016/j.concog.2009.12.010
Koriat, A., Bjork, R. A., Sheffer, L., and Bar, S. K. 2004. Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General 133:* 643-656.
Koriat, A., and Goldsmith, M. 1996. Monitoring and Control Processes in the Strategic Regulation of Memory Accuracy. *Psychological Review* 103 (3): 490-517.
Larkin, S. 2010. *Metacognition in young children.* New York: Routledge.
Loftus, E. F., Coan, J.A. and Pickrell, J. E. 1996. Manufacturing false memories using bits of reality. In: Reder, L. 1996. *Implicit Memory and Metacognition.* New Jersey: LEA.
Meltzoff, A. N., and Decety, J. 2003. What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive science. In Frith, C., and Wolpert, D. eds. *The Neuroscience of Social Interaction.* Oxford: Oxford University Press: 109-130.
Metzinger, T. and Gallese, V. 2003. The emergence of a shared ontology: building blocks for a theory. *Consciousness and Cognition* 12: 549-571.
Minsky, M. 2006. *The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind.* NY: Simon and Schuster.
Nichols, S., and Stich, S. P. 2003. *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford: Oxford University Press.
Pannu, J. K., and Kaszniak, A. W. 2005, Metamemory experiments in neurological populations: A review. *Neuropsychology Review* 15(3), 105–130.
Paynter, C. A., Reder, L., and Kieffaber P. D. 2009, Knowing we know before we know: ERP correlates of

initial feeling-of-knowing. *Neuropsychologia*: doi:10.1016/j.neuropsychologia.2008.12.009
Peacocke, C. 2007. Mental Action and Self-Awareness (I). In McLaughlin, B., and Cohen, J. *Contemporary Debates in the Philosophy of Mind*. Oxford: Blackwell.
Peacocke, C. 2008. *Truly Understood*. Oxford: Oxford University Press.
Peacocke, C. 2009. Mental Action and Self-Awareness (II): Epistemology. In O'Brien, L., and Soteriou, M. eds. 2009. *Mental actions and agency.* Oxford: Oxford University Press.
Pieschl, S. 2010. Epistemological beliefs and learning. Talk at the APIC Seminar, Institut Jean-Nicod, ENS, EHESS, Paris 22-01-2010.
Povinelli, D. J. 2000. *Folk physics for apes: The chimpanzee's theory of how the world works.* Oxford: Oxford University Press.
Povinelli, D. J., and Vonk, J. 2003. Chimpanzee minds: Suspiciously human? *Trends in Cognitive Science 7*: 157-160.
Proust, J. 2007. Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese* 159: 271-295.
Proust, J. 2008. Epistemic Agency and Metacognition: A Externalistic View. *Proceedings of the Aristotelian Society* 108(3): 241-268.
Proust, J. 2009a. It there a sense of agency of thought? In O'Brien, L., and Soteriou, M. eds. 2009. *Mental actions and agency.* Oxford: Oxford University Press.
Proust, J. 2009b. What is Metacognition. *Philosophical compass.*
Proust, J. 2009c. Overlooking the metacognitive experience. *Behavioral and Brain Sciences* 32: 38-39.
Proust, J. 2009d. The representational basis of brute metacognition: a proposal. In O'Brien, L., and Soteriou, M. eds. 2009. *Mental actions and agency.* Oxford: Oxford University Press.
Reder, L. 1996. *Implicit Memory and Metacognition.* New Jersey: LEA.
Sanna, L. J., and Schwarz, N. 2003. Debiasing the hindsight bias: The role of accessibility experinces and (mis)attributions. *Journal of experimental social psychology* 39: 287-295.
Shimamura, A. P. 2000. Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition* 9: 313–323.
Simons, J. S. and Spiers, H. J. 2003. Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews: Neuroscience*,4: 637–648.
Smith, J. D. 2009, "The Study of animal metacognition". *Trends in Cognitive Sciences* 13(9): 389-396.
Sosa, E. 1991. *Knowledge in Perspective.* Cambridge: Cambridge University Press.
Schwarz, N., and Vaughn, L. A. 2002, The availability heuristic revisited: ease to recall as distinct sources of information. In Gilovich, T., Griffin, D. and Kahneman, D. eds. *Heuristics and biases: the psychology of intuitive judgment.* NY: Cambridge University Press.
Thompson, V. A. 2009. Dual-process theories: a metacognitive perspective. In Evans, J. and Frankish, K. eds. *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.
Walsh, M. M. and Anderson, J. R. 2009. The strategic nature of changing your mind. *Cognitive psychology* 58: 416-440.
Wellman, H. 1990. The child's theory of mind. MIT Press.
Wittlesea, B. 1993. Illusion of familiarity. *Journal of experimental psychology* 19(6): 1235-1253.
Wittlesea, B. and Williams, L. 2001. Source of the feeling of familiarity: the discrepancy-attribution hypothesis. *Journal of experimental psychology* 26(3): 547-565.
Wolpert, D., et al., 2003, A unifying computational framework for motor control and social interaction. In Frith, C., and Wolpert, D. eds. *The Neuroscience of Social Interaction.* Oxford: Oxford University Press: 305-322.