1

From Social Cognition to Metacognition

Guy Lories, Benoit Dardenne and Vincent Y. Yzerbyt

Metacognition is a fundamental characteristic of human cognition. Not only do we have cognitive activities but it would seem that they can apply to themselves: we have cognitions about cognition. The possibility of metacognition seems typical of the human species and may be related to our being linguistic animals. It stands as one of the important differences between animal and human cognition and the very existence of psychology is proof of our interest in our own mental processes.

Interestingly, however, metacognition has long been neglected as a valid object of scientific inquiry. This state of affairs may have something to do with the disappointments – if not the trauma – that accompanied the historical attempt to use introspection for scientific purposes. It may also have to do with the early preeminence of behaviorism. Last but not least, it may have to do with a healthy defiance regarding some of the philosophical problems involved. Whatever the actual reasons may be, what is now commonly known as metacognition has not been at the center of preoccupations until recently. Worse, metacognition was long considered a nuisance. As noted by Nelson (1993), one usually prefers to short circuit most self-reflective mechanisms through experimental control. Still, this may hurt the ecological validity of psychological research. Many interesting cognitive activities are accompanied by rich contents of consciousness. It is necessary to wonder whether these contents are epiphenomenal or whether they play a role in the organization and functioning of our cognition. The present volume was born from the idea that all social interactions fall into the category for which the self-reflective character of cognitive activity is essential.

Social psychologists very often use devices that require at least some interest and confidence in self-reflectivity (e.g. rating scales). As Banaji and Dasgupta (Chapter 9 of this volume) note, however, they also investigate a number of phenomena in which awareness or, rather, the lack of awareness plays a central role. For instance, a large portion of stereotype research attempts to explain when and why perceivers remain largely unaware of the impact of their stereotypic biases. If people were able to detect these biases spontaneously, the biases would probably not have existed in the first place. Moreover, if warning people that some unwanted influence may bias their judgment was enough to allow them to detect that influence and adjust for its effects, the whole field would disappear. So, while social psychological

approaches acknowledge the importance of self-reflective elements, they also recognize and even claim that there are limits to what metacognitive or self-reflective activity can do. There are limits to what people know about themselves as well as to what people know about their fellow human beings. Social psychological approaches even suggest that these limits may be quite similar.

The differences and similarities between metacognitive knowledge of the self and metacognitive knowledge of the other are examined in great detail by Nelson, Kruglanski, and Jost in Chapter 5. The clear message that emerges from this review is that the various sources of information made available when people attempt to assess their knowledge of themselves and others only provide the raw material, indications that require interpretation in the light of more or less implicit theories. This perspective stands in sharp contrast with the idea that we are in complete control of our behavior and cognitive activity. It also stresses the fact that what we know about others and ourselves is the result of a complex construction process. But why do we need such a complex process in the first place? What is metacognition good for?

The problems of reflectivity

One prominent idea is that metacognitive activities may monitor and control other cognitive activities. This seems to imply a distinction between levels. The processes belonging to a metacognitive level would control and monitor the activities of the processes at a cognitive level. Yet, such a formulation almost immediately evokes the specter of infinite regress: If we need one level to control our cognitive functioning, why not another to control the previous one, and so on.

The conceptual difficulty is made even worse if we consider that metacognitive processes are often thought to be conscious while many other cognitive processes certainly are not. Because the notion of a process being "conscious" or not may seem obscure – it depends on a report by the subject – other distinctions have been developed. A process can be considered, for instance, as semantically penetrable or not (see Pylyshyn, 1984, for a complete discussion). A penetrable process is a process that can be affected by specific instructions or by giving the subject some explicit information. To give a specific example, the Müller-Lyer perceptive illusion is not cognitively penetrable. This means that the illusion is not affected by the knowledge that there is an illusion and that no form of experimental instructions can alter the impression produced by the stimulus. There is a general feeling that a large number of so-called automatic processes (memory access, spreading of semantic activation, etc.), that are not cognitively penetrable are also not conscious.[1]

A distinction can then be made between unconscious and automatic processes on the one hand and conscious and controlling (and themselves

uncontrolled?) metacognitive processes on the other. In some areas of cognitive psychology, researchers even suggested that the most reasonable object of study was the "module," a non-penetrable, encapsulated, automatic psychological function. A large enough number of modules would make up the cognitive architecture and whatever process is cognitively penetrable might be left for philosophy to study. As already noted above, this leads us to investigate situations in which experimental control is easier, which is a valuable advantage. One difficulty, however, is that this may tell us only part of the story.

From a philosophical point of view, this conception of the mind very strongly resembles a picture of consciousness as a unique place where mental life "happens," the so-called "Cartesian theater" conception strongly opposed by Dennett (1991). The "modules" just build a model of the world on the stage of the "Cartesian theater" in which the mind is the audience and where conscious processes will apply to themselves, indefinitely. This would indeed be a very convenient solution to the infinite regress conundrum: Ignore it and leave it to philosophy.

A different and maybe more valuable strategy is to address the problem directly. Nelson (1996, p. 105, note 5) indicates very clearly that "it would be a mistake to suppose that there must be different physical structures for object-level cognition and for meta-level cognitions. . . . For example, we do not need special structures for looking at our eyes – just a mirror. Feedback loops could play the role of the mirror for metacognition." The basic idea here is that the same architecture must be responsible for both cognitive and metacognitive processes simultaneously. At the same time, the meta level and the object level are no longer defined as parts of the mental architecture but are seen as a distinction applied to behavior by the observer. We now have to consider the usual cognitive architecture in order to identify the "mirror" postulated by Nelson.

One should note that whatever appears in the mirror must have been put in front of it by some process that is, in a sense, more elementary. So there is a problem of levels after all, but it is a different one. It has to do with the fact that the cognitive level itself emerges from more elementary levels. For instance, Newell (1990) describes the levels of biological, cognitive, rational, and social functioning as shown in Figure 1.1.

Figure 1.1 depicts the basic scheme that underlies the computational metaphor in cognitive psychology. Each level of description is independent in the sense that it has its own laws or principles that can be described independently of the levels below it. Each is nevertheless implemented using relatively simple operations that belong to a lower one. One original aspect added by Newell (1990) is to assign a time band to each level. Loosely speaking, significant social processes such as those that involve interactions between a great many individuals, seem to require days, while a neuron operates on the millisecond scale. Starting from the bottom, the 10 milliseconds assigned to the operation of a neural circuit correspond to the time it takes for a couple of neurons to conclude a transaction of some sort (they

Duration	Action	Temporal band
month		
week		social band
day		
hour	task	
10 min	task	rational band
1 min	task	
10 sec	unit task	
1 sec	cognitive operation	cognitive band
100 msec	deliberation	
10 msec	neural circuit	
1 msec	neuron	biological band
100 μsec	organelle	

Figure 1.1 *Time bands and levels (adapted from Newell, 1990)*

need to integrate action potentials over at least a short period). This is what Newell takes as the lower limit of cognition because it would be the time to access a simple symbol. The 100 millisecond level is the level of an elementary deliberation, i.e. an elementary choice between two possible – automatic – mental operations. The typical cognitive act takes about one second and may be something like uttering the sentence "Pass the salt please!" At the next level, the unit task is described by Newell as the first level of full cognitive functioning; it is an assembly of simple cognitive operations and takes about 10 seconds. Newell cites reasons to believe that the unit task level is indeed meaningful. Newell argues that the time spent by an expert chess player to consider a move (6 to 8 seconds) is indicative of this level of complexity and that, in most verbal (think aloud) protocols, the elementary steps take about that long. Above this limit, the behavior observed with these protocols is essentially rational and obeys some explicit algorithm that the subject obviously learned or developed with experience. This is why think aloud protocols are a nice way to investigate problem solving: They give us access to the appropriate level and allow us to identify the subject's strategies in terms that enable us to determine whether they are rational or not and what their limitations are.

The methodology of think aloud protocols is based on the idea that these protocols allow access to the products of cognitive activities, that is, to the contents of a working memory (Ericsson & Simon, 1984); they can be used

to describe large scale, non-automatic, quasi rational behavior provided that the experimenter keeps in mind the possibility that the verbal report of these contents may be incomplete. The difficulty is to make sure, through proper experimental instructions, that the participants do not start introspecting but stick to the description of the contents of working memory.

The above analysis suggests that we may consider metacognition *as the processing of the contents of (working) memory by standard cognitive processes*, and forget the idea of any direct and mysterious access by people to the intricacies of their own mental functioning. In cognitive architectures like Newell and Simon's "physical information systems", Newell's SOAR or Anderson's ACT* and ACT-R, the general processing principle is the matching of productions with the content of working memory. There is nothing to prevent specialized productions from applying to these contents and implementing effective metacognitive abilities, thereby providing a measure of monitoring and control. It is exactly here, to use Newell's words, that "cognition begins to succeed" and begins to apply to its own products. We believe that this is also where metacognition emerges. Although the social band, where interactions between individuals belong, is the upper one in Newell's hierarchy, the durations we meet in social cognition suggest that cognitive social psychology is concerned with the upper part of the cognitive band and maybe also the lower part of the rational band.

Because social behavior and metacognition rely on general cognitive processes, metacognition must also rely on approximations, limited capacity processes, arguable heuristics, intrusion of naive theories, and so forth. The interesting question is whether this explains why our judgments about others and ourselves cannot be better and how they go wrong. The various chapters in this book analyze simple cognitive tasks and a number of problematic social situations. The cognitive analysis seems to converge on a small number of principles that may not be completely understood yet but that seem to agree with what has been observed in social cognition. As such, they hold great promise for future research and the possibility of an integrated approach to social cognition.

Familiarity, availability, accessibility, representational richness, etc.

The idea

The general idea to provide some ground for metacognitive processes is to use the contents of working memory. We will imagine that according to these contents and depending on the circumstances, people reach a judgment concerning their cognitive situation. Consider a person trying to remember something. What will the content of working memory be? It may be something like a name or a place evoked by a cue but it may also be something more like a feeling. It may be an impression that some memory trace has been easily accessed, that some face is familiar, that a pattern has

already appeared. In any case there will be the problem of making appropriate inferences from that evidence.

This can be done in several ways. It may involve a fairly explicit and rational line of reasoning or it may be done in a more automatic manner, some of the heuristics may be appropriate and some not, the conclusions reached may or may not be followed by an action, the whole process may come to be more or less automatized with practice and so on. Eventually, a decision will be reached that whatever is shown is familiar because it has been seen already, that information retrieved by memory search indicates that more information can and will be retrieved, etc. The trouble, of course, is that, as for any heuristic, things can go wrong. An unknown face may seem familiar because it simultaneously resembles a lot of other faces that we do know, information retrieved from memory on a given topic may actually be wrong, the impression that something studied has been mastered and is more easily accessed after a learning trial may indicate only that we have not let enough time go by, etc. In other words, the contents of memory in the broad sense indicated above may mislead because it is difficult to determine exactly why they are what they are and what produced them in the first place.

This may be considered as a classical attribution problem. A given response has been evoked in a subject, for instance a feeling of familiarity, and a causal attribution is required. If the correct attribution is made, it will identify a mental state and provide information regarding the functioning of the mind at the moment. It will actually monitor a state of mind. According to standard attribution theory, attributions are made by people about themselves, as they could be made about others if the same kind of information were available. This introduces a fundamental similarity between self and others but also agrees with the idea that there is nothing especially mysterious about metacognition (but see Jones & Nisbett, 1972).

The analysis is similar to the analysis of the source problems in cognitive psychology. In various memory tasks it will often appear that the subjects will erroneously identify a stimulus as having been presented in some context while it has been presented in another. A similar problem is to make the difference between what has been imagined and what has been actually presented (Johnson & Raye, 1981). The question here apparently bears on some external attribution (was the stimulus presented?), but the alternative is internal generation. It has recently become apparent that the mechanisms involved may be especially brittle and among the first to go when there is neurological damage to the memory system.

The problem is made complex by the fact that some memory contents may result from priming. The presentation of one stimulus in the context of another may lead to the automatic retrieval of more or less relevant information and/or to an easier retrieval of the relevant information. This may lead to an impression of greater familiarity. Most heuristics do not take into account automatic activation effects and do not make the distinction between whatever content has been evoked by the stimulus and

some aspect of the context. In other words, the source identification problem (the attribution problem) is especially difficult because of the presence of automatic (non-penetrable) activation phenomena.

The feeling-of-knowing example

The above line of thinking pretty well matches what has been discovered in feeling-of-knowing (FOK) research. The FOK is a rating made by people about the probability that they will be able to recognize an element of information they have just been unable to recall. In most experimental operationalizations, participants answer general information questions and make a FOK rating when they fail to remember. A recognition test is given at the end of the session to assess whether their predictions are valid.

The FOK has an intuitive similarity with the well-known tip-of-the-tongue situation, but the FOK does not occur spontaneously and it is not as intense. In the case of the FOK, the theory evolved from a relatively mysterious "partial access" process to a two-process theory based on cue familiarity and amount of material retrieved. As suggested by Nhouyvanisvong and Reder (Chapter 3 in this volume), it seems necessary to distinguish between a fast and a slow FOK. The first type of FOK would be automatic, would rely on cue familiarity and would be involved in pre-retrieval decisions (as, for instance, whether to search memory or not). This fast FOK requires an interpretation of a feeling (familiarity) that is open to errors. Nhouyvanisvong and Reder develop a complete theory of how attribution errors are possible in this context and suggest that, from a cognitive point of view, the problem stems from a confusion of sources. Some forms of cognitive processing, especially fast assessments like this FOK judgment, would allow for more source confusions because they allow the contributions of various cues to add up.

The slow FOK depends on the results of the retrieval activity itself (see Koriat, Chapter 2 in this volume). It is slower because it requires that at least some retrieval attempts take place before an evaluation of what has been retrieved can be made. It is essentially a rating based on the contents of working memory and subjective norms of knowledgeability but it is less clear how it could be very effective as an adaptive strategy (at least at the beginning of the search process). It may not be sensitive to the same source misattribution effects as the fast form of the FOK judgment but it is, in any case, a construction. Koriat describes how this construction takes place and introduces the concept of *accessibility* as the cornerstone of the FOK rating. The idea is that whenever a cue is effective in retrieving a lot of material from memory, this high accessibility indicates that still more material can be retrieved and, presumably, the correct answer also. The problem is that some cues may be deceptive in the sense that retrieving a lot of material when attempting recall does not guarantee that the correct item will be retrieved in due time. What will be retrieved or recognized later on may just be an error. This is why, according to Koriat, the FOK accuracy depends

on *output-bound accuracy*, the probability that an answer is correct, once it is actually given. Some items will be essentially deceptive and will produce commission errors that the accessibility heuristic cannot forecast. The absolute FOK level, however, depends on accessibility as defined above.

Although accessibility is defined by the amount of information memory search retrieves and very much resembles informational richness, it does not follow that only internal cues are used to determine the FOK rating. As Lories and Schelstraete (Chapter 4 in this volume) argue, accessibility is a good basis on which to make the FOK judgment because it summarizes, but also confounds, information from a number of different sources. Because of the general laws of human memory, accessibility is bound to correlate with numerous contextual cues like domain familiarity, the existence of related episodic traces, etc. A number of sources that may be described as "meta-informational" are potentially involved.

As a result, recognition performance, these sources, accessibility and, of course, the rating will all correlate. So the very reason why accessibility is a sound heuristic also makes it likely that accessibility will correlate (be confounded) with other cues and it is difficult using a correlational design to make sure that causality goes one way or the other.

On the other hand, whether the answer is well known or not, the information retrieved during memory search will usually decrease recognition uncertainty. For instance, it may make some distractors less plausible and guessing more likely to succeed. This will increase recognition performance for precisely those items that have led to the retrieval of large amounts of information. Whether this specific characteristic of the recognition test leads to a bias or to an appropriate assessment of FOK accuracy is debatable. In any case, it means that a specific class of items may yield higher FOK accuracy if this mechanism is made more effective for that class.

The conclusion is, as in Koriat's view, that the FOK, as a rating, will have a significant accuracy, because of the way human memory works in general and that it will correlate with accessibility for the same reasons, but Lories and Schelstraete stress the constructed aspect of the rating a little more. The analysis shows that the accuracy of the FOK will be constant only in a given experimental context, with a specific type of items and specific recognition alternatives. Things are worse for the absolute level of the rating. It will not be very stable from one item list to another because there is no principled way to set it.

Generalizing: availability and representational richness

Bless and Strack (Chapter 6 in this volume) investigate the idea that people have metacognitive theories about the memorability of objects and events. When these theories lead people to feel uncertain about the occurrence of an event, the situation is set up for the impact of social influence on memory. At this point, and in line with Festinger's social comparison theory, perceivers can decrease or increase their confidence by relying on other people.

Uncertainty is particularly high, and so is social influence, when there is no memory trace, because the absence of such a trace could indicate either that the event has not appeared or that it has occurred but could not be recollected.

A simple example may illustrate the above reasoning. Naive theories about memorability hold that some events are more memorable and thus would not have been forgotten had they happened. These theories could be right or wrong, the fact is that people hold them. Imagine that you want to hide something. You probably hold firm beliefs about what could be a highly memorable place, such as the water tank or the refrigerator for hiding your jewels. Later on, your partner tries to convince you that you probably put the jewels in the water tank. If you did not, you would ridicule your partner because you will be confident that the water tank could not be the place. In contrast, if your partner suggests a location which is much less memorable (i.e. salient), you might end up checking. In their own paradigm, this is exactly what Bless and Strack found (see Chapter 6).

This provides a conceptualization of social influences on memory: Because memory itself is a reconstructive process, there is a always the potential for manipulation after the fact. This is what the misinformation paradigm is all about and the constructive nature of memory has been long known, but the critical fact is that the process is cognitively penetrable. This is not just a matter of automatic inference and automatic – if erroneous – reconstruction. Low salience, hence *perceived* suboptimal encoding, or *perceived* suboptimal retrieval conditions are necessary for the effect to be obtained. The reconstruction is penetrable, the error is sensitive to properly informational influences.[2]

Bless and Strack's chapter deals with a case in which metacognition is inaccurate because it rests on limited information, e.g. when memory is empty. Swann and Gill (Chapter 7 in this volume) further explore the problem of accuracy and more precisely the overconfidence – high confidence with low accuracy – we often display in intimate relationships. Based on the daily observation that we come to feel we know and trust our partner well even in the absence of true accuracy gains, these authors consider several mediators for this overconfidence effect. They suggest that *representational richness* is at the heart of overconfidence. Representational richness is defined as the amount of information available – increasing with relationship length – and the degree of its integration – increasing with involvement in the relationship. Richness does not mean accuracy because, for instance, any information will increase richness but only truly diagnostic and pertinent information can increase accuracy.

Interestingly, because accessibility (probability of retrieval) is a matter of memory organization and coherence, representation richness is very much like accessibility in FOK research. Like accessibility, it is independent of the accuracy of the representation. The information may be rich and integrated but non-diagnostic and, in this case, representation richness may increase confidence without increasing accuracy. Availability as a strategy for

estimating probability has similar properties. Our capability to retrieve examples may be taken as a probability estimate but the problem is that an increase in the number of retrieved examples may result from many different causes. For instance, recency will increase the probability of retrieval but this indicates only that one case occurred recently. Swann and Gill found strong support for their conception in several correlational as well as experimental investigations that are reviewed in their chapter.

People receiving information on a given topic may or may not become aware that it is relevant and important depending on a number of cues that may be present. To take a trivial example, participants may be warned that they will be given this information by an experimenter. A less trivial example is the subjective – but purely apparent – availability of individuating information. Such cues are obviously not part of the relevant information themselves but they should be expected to increase confidence in a judgment compared to a situation in which the person is *not* made aware of the available information. Because the information provided is actually held constant in this design, this suggests that confidence increases whether information has actually been provided or not.

According to Swann and Gill, this type of overconfidence can be generated without increasing representational richness. In other words, meta-informational cues do not seem to increase confidence via representational richness. The specific mechanism involved in this case could be that meta-information simply increases the accessibility of the information, which does not affect richness but nevertheless increases confidence.

Judgment construction and correction: metacognition and naive theories

Judgment construction

Meta-informational cues as well as overconfidence are at the heart of Yzerbyt, Dardenne, and Leyens' chapter (Chapter 8 in this volume). According to the social judgeability model (e.g. Yzerbyt, Schadron, Leyens, & Rocher, 1994), (over)confidence is a function of a variety of non-diagnostic aspects of the information and of the judgmental context. The extreme situation is one in which meta-information only influences people's evaluation of confidence. As a special case, the judgeability model foresees that the link between confidence and meta-information may be indirect: Some additional richness may be derived from the meta-informational cues granting appropriate inferences.

In the eye of any social perceiver, a judgment is loaded with meaning: It involves some personal commitment. Because people are to a certain degree committed to their judgment, they want to respect a number of criteria. One is accuracy, but social norms and the framing of the information provided are also important to the perceiver. These additional criteria are the focus of

the chapter. The point is that people take social and identity concerns into account to estimate the validity of their judgments, which they proceed to do by relying on naive theories about judgment process.

Yzerbyt, Dardenne and Leyens begin with a description of the social judgeability model and then provide several studies that assess the impact of the naive theories that proceed from social norms. In one of these experiments, the subjective availability of individuating information contributes to the expression of stereotyped judgments. The authors further show that naive theories are truly affecting private beliefs and are not used simply for the purpose of impression management. In another investigation, they present evidence that implicit rules of judgment can have an impact from the beginning of the impression construction, and not only at the end of the process.

In the work reviewed by Yzerbyt and colleagues, metacognitive processes have an impact on social judgments without the perceiver being aware of that influence. Banaji and Dasgupta (Chapter 9 in this volume) present some other non-conscious ways in which beliefs, attitudes, and behaviors are influenced. Based on Johnson-Laird's view of consciousness, the authors focus on spontaneous, uncontrolled, and unconscious beliefs about social groups. They review several experiments in which people display no awareness of and are not able to consciously control the impact of their naive beliefs on the judgments. For instance, among a list of potentially familiar criminal names, black names will be (mis)identified more often than white names as perpetrators of criminal acts (which fits the stereotype linking blacks and criminality). In that case, people are confident that their judgments were based on genuine memory for criminal names. Moreover, they hold explicit egalitarian and non-racist values!

The authors then explore the concepts of responsibility and intention. They clearly dissociate intention as fair and egalitarian on the one hand from discriminatory act and prejudice on the other hand. They discuss this issue with regard to the law. The problem can be summarized in the following manner: Are people responsible for a discriminatory judgment if they are not aware of it and cannot control their reaction (Fiske, 1989)? To make things even more difficult, the very same discriminative act can have intentional as well as unintentional causes.

The need for correction

Although they may not be aware of the details of automatic activation effects, most subjects are aware that aspects of the context can have an impact on their judgment. For instance, they will correctly suppose that a stereotype can influence their perception of others, although they may not understand that the activation of the stereotype and the priming of the features consistent with it are automatic. They will make conscious attempts to determine whether a source confusion may have taken place and will take measures to adjust for these effects. The difficulty is that they will not

have much information to work with. Some external aspects of the situation may signal that an error is possible. People may have become familiar, for instance, with the general contrast or assimilation effects that occur when a stimulus is presented in a given context and they may try to discard the part of their impression that they know must be related to the context. Several theories have been proposed to describe this phenomenon. Three chapters in this book deal with people's naive theories of biases and unwanted influences and contribute to our knowledge concerning the ways people try to remove biases guided by their naive metacognitive theories.

Wilson, Gilbert and Wheatley (Chapter 10 in this volume) investigate the role of lay beliefs in protecting our minds against unwanted influences on our own beliefs and emotions – what the authors call mental contamination. People's naive theories about how and when their beliefs and emotions could change determine the specific strategy they follow in order not to be contaminated. The authors suggest that we distinguish between an implicit level of psychology which operates largely outside of awareness (people's use of schematic knowledge) and an explicit level of psychology (meta-beliefs about cognitive processes). Explicit psychology very much corresponds to metacognition.

Wilson and colleagues present an extensive version of their general model, from *exposure* control – whether the person allows the stimulus to enter their mind or not – to *behavior* control – acting as if our mind was not contaminated. Between these extremes, people can use different strategies. Whereas the best way not to be contaminated is simply not to be exposed to the stimulus, the authors speculate that people do not always prefer that strategy, at least in the case of mental contamination. According to the authors, people believe that mental (but not affective) contamination does not necessarily have detrimental effects. Moreover, people seem to think that they can control their beliefs (more easily than their emotions); in other words, according to the authors, people would feel that their beliefs are not "penetrable" by external and unwanted information. As a consequence, they do not systematically avoid exposure to mental contamination. For instance, whereas people think they can freely decide whether to accept a proposition as true, they do seem to understand and believe a proposition at once. The authors review indirect support for their model, bearing on studies conducted for other purposes. They then present recent and more direct data.

Wegener, Petty and Dunn (Chapter 11 in this volume) offer another look at the naive theories of bias. They discuss some of the alternative models of correction and present the unique features of their flexible correction model under the form of several postulates. The model is based on the idea that correction is highly flexible. Corrections are driven by highly context-dependent naive theories of how a given factor can influence judgment. This is why appropriately chosen information will produce adjustments, again making the correction process a "penetrable" one. They suggest that naive theories can be stored in long-term memory but are also likely to be

generated on-line (see also Yzerbyt, Dardenne & Leyens, Chapter 8 in this volume). The correction is also flexible in that it depends on the level of motivation and ability. Clearly, thus, correction is costly.

The authors review several initial tests of the flexible correction model and present new data. As predicted by their model, they find opposite corrections for the same target, depending on the specific naive theory people entertain concerning the potential biasing effect of the context. For instance, participants may consider that, if context involves an extremely violent person (vs. an extremely non-violent person), people would judge a target as less (vs. more) violent than if such a context was not present. In the main study, participants were confronted with an extremely violent or non-violent context and were asked either to rate the target immediately or to first correct for the context. Results show that participants obey their naive theory and correct their judgments away from the perceived bias, even when no bias has actually occurred.

In the last chapter of this trilogy (Chapter 12 in this volume), Martin and Stapel begin with a critique of theory-based models of correction. In their eyes, a critical aspect is that those models are guided by a priori and verbalizable naive theories. The authors do not dispute the fact that people have naive theories about their judgment processes. However, they argue that these conscious and naive theories are generally not the causal factor in people's judgmental correction processes. They then discuss the view that people's accuracy attempts are guided by non-conscious processes initiated by features of the general judgment setting (i.e. the implicit processing objectives activated by features of the setting).

In line with Martin's earlier set/reset model, the authors thus accord a much smaller role to naive theories than the two preceding models. Martin and Stapel's model is based instead on a production system that is not open to awareness – not penetrable. Correction, or "reset" in the model jargon, as well as assimilation, or "set," can lead people to experience conscious thoughts and feelings as outputs of the production system. In that way, people's theories come after the judgment and stand as post-hoc explanations or rationalization.

Conclusion

The theme of metacognition is inextricably related to problems of awareness, verbalization, penetrability and to the paradoxes of reflectivity. It is necessary to determine how the mind may actually deploy effective metacognitive abilities without postulating mysterious powers or generating infinite regressions. One way to do this is to conceive metacognition as ordinary cognition applied to its own products in a standard cognitive architecture. The present volume is born from the idea that cognition becomes social at the very moment when metacognition becomes possible, that both social cognition and metacognition depend on the possibility of

using the products of cognitive activity in self and others to monitor the cognitive processes themselves.

Yet, this possibility requires an inductive step: Because we go back from the products to the processes, metacognition is a reconstruction and our representation of our own mental functioning is not exact. In the language of social psychology, the inductive step can be described as an attribution process. Attributional work is necessary if we are to go back from the products of cognition to their source and to the conditions that produced them, to bridge the gap between content and process (see Nisbett & Wilson, 1977). Familiarity, accessibility of information or representational richness all play a similar role in this sense; they are used as data that require an interpretation. Metacognition (as well as consciousness?), from this point of view, is woven by a complex set of inferential processes using a variety of elements and there may be holes in the fabric. What we reconstruct may even be plainly wrong.

The main reason for this is that cognitive activity involves automatic processes that we are not aware of. This has been known since Helmholtz and clearly appears, for instance, in our short discussion of verbal protocols. These automatic, non-penetrable, processes will alter the contents of working memory without leaving any perceptible trace of their action. For instance, various contextual cues may prime a stereotype. Hence, an attribution problem occurs: The source of working memory modifications may not be identified properly and errors will follow. Of course, people are not completely naive regarding the whole matter. We usually suspect, for instance, that stereotypes can be primed by contextual elements. Although we may not be aware of the effect of automatic cognitive processes we do know, from education or experience, that influences do occur in specific conditions and may lead to judgment errors. We entertain naive theories regarding these conditions that allow us to call for corrections when "meta-informational" cues indicate that it may be appropriate. We may even decide to protect ourselves from such situations by eliminating some sources of information. Unfortunately, these theories and the corrections they prescribe may not be perfectly correct or effective. There is no guarantee that they will exactly describe the actual influences we undergo.

Naive theories are built and applied "from the outside." They apply to everybody, self and other, in the same way. They rely on the idea that context will have an effect on your judgment. The cues to that effect are in the context, so naive theories can be expected to work in the same way for self and other. Interestingly, it is also a characteristic of the standard theory of attribution that attribution works the same way in self and others. Although the contents of working memory may not always be open to verbalization, and although some of these contents may be privileged in the sense that they are accessible only to ourselves, they nevertheless require the same kind of interpretation that would be required if similar information was available regarding others. So, the information that is available for making inferences regarding self and other may be different,

but the process is similar in the sense that this information requires an interpretation, an inductive step. Hence, for most purposes, the lesson of our social approach to metacognition agrees with the poet: "Je est un autre".

Notes

1. Although the distinction between cognitively penetrable and cognitively non-penetrable processes is most useful, it should be mentioned that declarative knowledge used in a controlled and cognitively penetrable manner may be compiled during learning and eventually become automatized and inaccessible to conscious observation.

2. In other words, not only does memory not work like a recording module but data even show that committing oneself to an interpretation has a definite effect on subsequent retrieval attempts.

References

Dennett, D.C. (1991). *Consciousness explained.* Boston, MA: Little, Brown & Company.

Ericsson, K.A. & Simon, H.J. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Fiske, S.T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J.S. Uleman & J.A. Bargh (Eds), *Unintended thought: Limits of awareness, intention, and control* (pp. 253–283). New York: Guilford Press.

Johnson, M.K. & Raye, C.L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85.

Jones, E.E. & Nisbett, R. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E.E. Jones, D.E. Kanouse, H.H. Kelley, R.E. Nisbett, S. Valins, & B. Weiner (Eds), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Morristown, NJ: General Learning.

Nelson, T.O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology: General*, *122*, 269–273.

Nelson, T.O. (1996). Consciousness and metacognition. *American Psychologist*, *51*, 102–116.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Pylyshyn, Z.W. (1984). *Computation and cognition: Toward a foundation for cognitive science.* Cambridge, MA: MIT Press.

Yzerbyt, V., Schadron, G., Leyens, J.-Ph. & Rocher, S. (1994). Social judgeability: The impact of meta-informational rules on the use of stereotypes. *Journal of Personality and Social Psychology*, *66*, 48–55.