Introspection and Metacognition: Mechanisms of self-knowledge

Stanislas Dehaene
Chair of Experimental Cognitive Psychology

Lecture

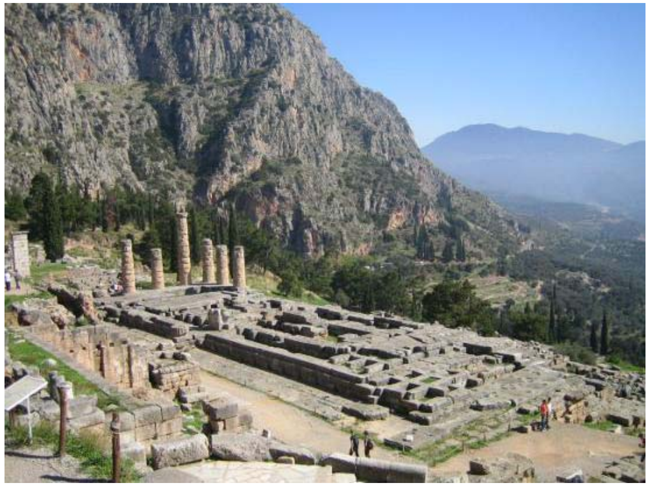
Definitions and first paradoxes

Course translated from the French version by CG Traduction et Interprétation

Can we conceptualize that we are aware of….. being aware?

γνῶθι σεαυτόν : Know thyself

Inscription in the pronaos of Apollo's Temple in Delphi



"I continually escape and I do not quite understand, when watching myself act, that the person I see acting is the same as the person watching, and who wonders and doubts he can be actor and watcher all at once.

André Gide, *The Counterfeiters*

"Being aware of being aware of being… In other words, if I not only know that I am, but also know that I know it, then I belong to the human species. All the rest follows—the glory of thought, poetry, a vision of the universe. In that respect, the gap between ape and man is immeasurably greater than the one between amoeba and ape.

Vladimir Nabokov, *Strong Opinions*

What brain architecture allows us to turn our thoughts onto themselves in this way?

Some simple examples

- Serial strategies
  - Calculate 13+28. In what order did you make your calculations?
  - Have you detected any error? Did you need to go back?
  - Strategies and action plans are often accessible to our conscious mind whereas elemental operations are not.
- « Tip-of-the-tongue » experience
  - How do you call a prodigious being, half-man, half-horse?
  - We might not remember the answer, while being aware that we know it!
- Learning monitoring and metamemory
  - How do you decide to revise before an exam?
  - Have you already danced with a famous actress?

Some elements of definition

- **Cognition:** Set of mental processes involved in (internal or external) information processing.
- **Metacognition:** Set of knowledge and beliefs concerning our own (passed, present or future) cognitive processes; processes used to manipulate them.
  - **Metamemory: our knowledge and beliefs on memorization processes and memory retrieval**
- **Introspection:** (literally: « in-sight »)

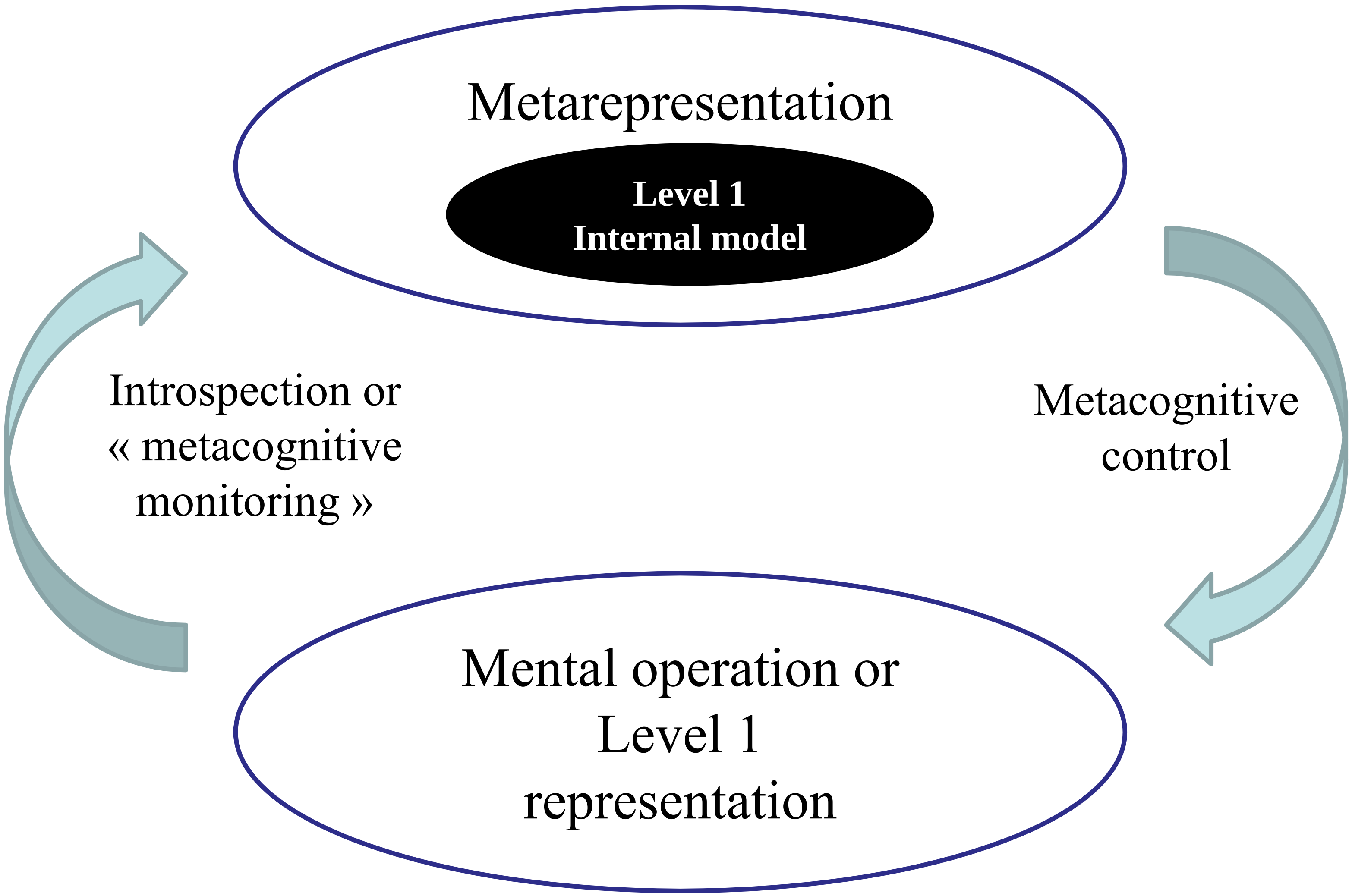
Ability to consciously access our mental operations and attribute them to ourselves or to someone else.

  - **Knowing we understood a theorem**
  - **Estimating we are close to the solution**

- **(Meta)cognitive Control:** Ability to regulate our own mental processes according to our introspection
  - **Change strategies**
  - **Be more careful**

Emergence of a theoretical framework

Nelson, T.O. & Narens, L. (1990). Metamemory: A theoretical framework and some new findings.
In G.H. Bower (Ed). *The Psychology of Learning and Motivation*, 26, 125-173. New York: Academic Press



Metarepresentation

Level 1
Internal model

Introspection or
« metacognitive
monitoring »

Metacognitive
control

Mental operation or
Level 1
representation

In 19th century psychology introspection is central

• Until the 19th century, introspection was considered as the central method for studying human mind, a form of direct observation of mental « facts ».

• Wilhelm Wundt (1832-1920, Leipzig): the very purpose of psychology is to study subjective mental experiences, which can only be studied by introspection.

Franz Brentano (1838-1917) promoted ‘descriptive psychology’ or ‘phenomenology’ (before Husserl) consisting in the study of internal perceptions, from the « first person » perspective ».

• Oswald Külpe (1862-1915), Wundt’s student and leader of the Würzburg School, developed methods of verbal description of introspection (for instance, describe what comes to your mind when you read the word « meter ») – but discovered the “imageless thoughts” : the subject is not always able to report relevant percepts.

•Edward Titchener (1827-1927, Cornell), Wundt’s student, claims that introspection is psychology’s only method.

Edwin Boring (1886-1968, Harvard): « If the subject matter is immediate experience, it is plain that the method is immediate experiencing » (*A History of Experimental Psychology*, 1929)

• In France, Théodule Ribot (1839-1916) and Alfred Binet (1857-1911) defend similar viewpoints: « introspection, one may say, is the basis of psychology, it characterizes psychology so precisely that any study by introspection deserves to be called a psychological study, and any study made by another method belongs to another science » (A. Binet, *Introduction to experimental psychology*, 1894)

•For Jérôme Sackur, introspection never disappeared from psychology methods.

Comte’s paradox

« It is tangible indeed, that by an invincible necessity, human mind can directly observe any phenomenon, except its own. Because who would make the observation? (…) The thinking individual couldn’t divide into two, one reasoning, the other watching the reasoning. The observed organ and the observing organ being identical in this case, how could the observation be made? This so-called psychological method is therefore totally worthless in its very principle. »

Auguste Comte, *Lecture on Positive Philosophy* (1830-1842), Vol. 1, pp. 31-32

« It could have come to Mr. Comte’s mind that it is possible to examine a fact through memory, not at the moment of its perception, but in the following moment: this is actually the method by which, in general, the best of our science concerning our intellectual acts was acquired. We think about what we have done once the act is over, but with the impression still vivid in our memory.
This simple fact shoots down Mr. Comte’s entire argument. »

John Stuart Mill, *Auguste Comte and Positivism* (1865), pp. 68-69.

Contemporary disproof of Comte’s paradox

Our mental processes are made up of many partially specialized processors, so it can’t be ruled out that some of them might be “observing” others.

The prefrontal cortex, in particular, is in a position to receive information from all our other mental processes :

« A good way to begin to consider the overall behavior of the cerebral cortex is to imagine that the front of the brain is ‘looking at’ the sensory systems. »

Crick & Koch, *Nature Neuroscience*, 2003

However, Comte’s observation points to two interesting and open questions :
- Does the very fact of asking the participant to perform introspection affect the initial processing of the information?
- It is probably possible to use Comte’s paradox to demonstrate that perfect and complete introspection is impossible (*Data*, the android from Star Trek, who is supposed to have a perfect memory and perfect access to the motives of all its decisions).

Behaviorist criticism

« Psychology as the behaviorist views it is a purely objective discipline of natural sciences. Its theoretical goal is to predict and control behavior. Introspection is not part of its essential methods, and the scientific value of its data does not depend on whether they can be interpreted in terms of awareness. »

John Watson (1913), *Psychology as the behaviorist views it*

Main criticism: subjectivity of observations.

« The consequence of the major postulate that something such as consciousness does exist, and that we are able to analyze it through introspection, is that there are as many analysis as there are psychologists. »

John Watson (1925), *Behaviorism*

Some answers to Watson :

- His criticism mistakes introspection as a method for accessing mental architecture for introspection as a *research subject.*
- Introspection (as well as its limitations) is a perfectly legitimate research subject which leads to empirical results that can be reproduced from one individual to another.

Metacognition and cognitive sciences program

John Flavell introduces the study of *metamemory* (1971), and the distinction between *monitoring* and *regulation* (1976).

In 1979, he suggests a first theorization for metacognition, which identifies :
-metacognitive knowledge (conscious or not, right or wrong)
-conscious experiences
-goals and tasks
-strategies and actions.

Influenced by Piaget, he underlines the importance of metacognition in child education (active research strategies and information memorization).

1960-1990: Heated debate and active research on the reliability of introspection :

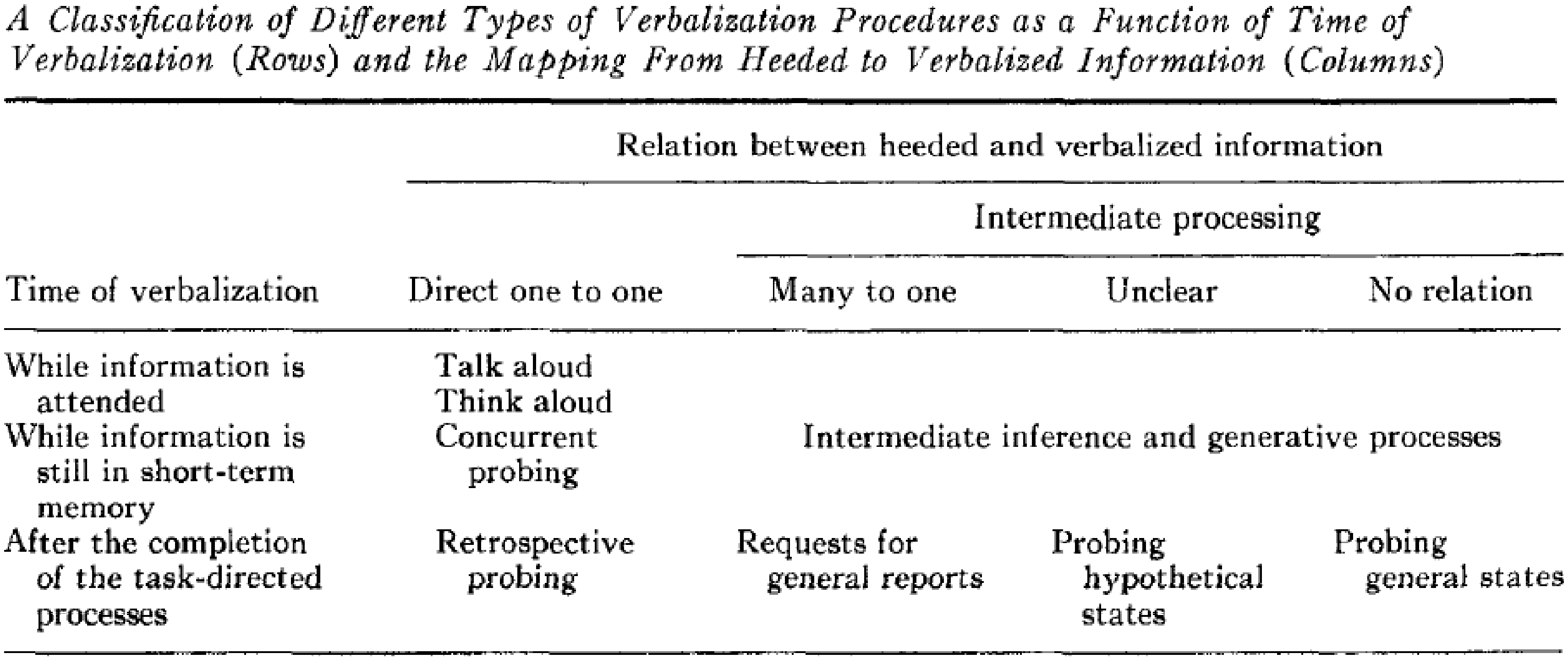
- For Nisbett and Wilson (1977), introspective judgments are very often fictitious. Example: preference for the right hand side when choosing among 4 equivalent objects

- For Ericsson and Simon (1980), verbal reporting is appropriate and useful if the reported information is still present in the short term memory

General theory on verbal reporting

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87(3), 215-251.*

Ericsson and Simon introduced a classification of introspective tasks distinguishing the *time* of verbalization (immediate or delayed), from the *type* of verbalization procedure (direct, unclear or no relation).



A Classification of Different Types of Verbalization Procedures as a Function of Time of Verbalization (Rows) and the Mapping From Heeded to Verbalized Information (Columns)

	Relation between heeded and verbalized information			
		Intermediate processing		
Time of verbalization	Direct one to one	Many to one	Unclear	No relation
While information is attended	Talk aloud Think aloud	Intermediate inference and generative processes		
While information is still in short-term memory	Concurrent probing			
After the completion of the task-directed processes	Retrospective probing	Requests for general reports	Probing hypothetical states	Probing general states

Their review of experimental data suggests that a verbal report is reliable when it is *direct* and describes the *current* content of short term memory.

Under such circumstances, what subjects say and what they do can be remarkably consistent (for instance, in one card sorting test [Dulany and O'Connell, 1963], 11 answers at variance with verbalization out of 34408 = 0.03 %).

Metacognition and cognitive sciences program

Major experimental breakthrough in 1960-2000: invention of **new experimental measurements for introspection :**

- **Metamemory tasks :**
  - *judgment of learning:* after a learning phase, the individual is asked what his performance will be in the subsequent memory test.
  - *feeling of knowing:* immediately after failing to remember an item, the individual is asked to judge prospectively whether he could recognize it among several others.

- **'Secondary' Judgments :**
  - *confidence:* digital judgment of confidence in a primary answer
  - *wagering:* wager on the accuracy of one's answer
  - Error-detection

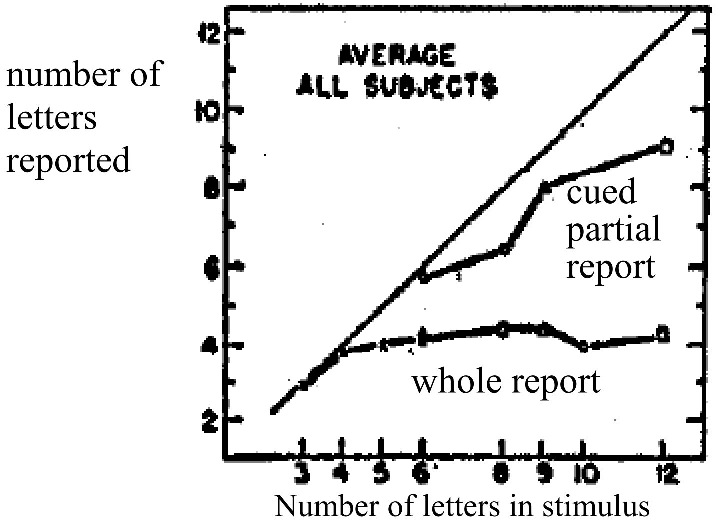
- As Jérôme Sackur points out, **psychophysics** itself**,** on a regular basis, use carefully quantified and replicated introspection (with verbal or non-verbal report).

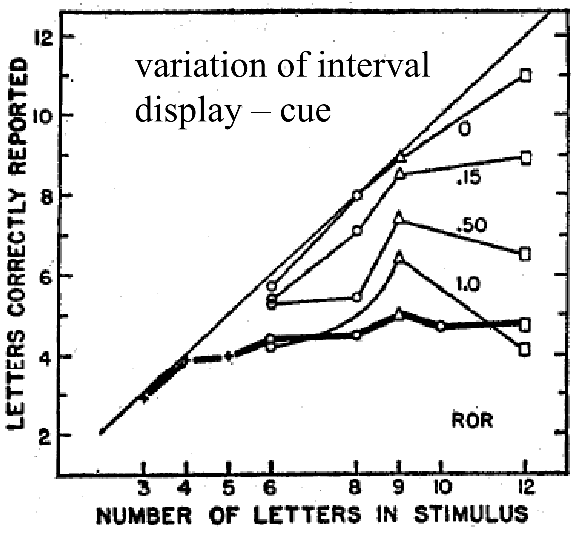
Historically important example: Sperling's experiment(1960).

Sperling’s iconic memory experiment

Sperling, G. (1960). The information available in brief visual presentation. *Psychological Monographs, 74*, 1-29.

Q	T	A	W
Z	F	B	D
V	M	I	K





•After being briefly shown an array of letters (50 ms), we are only able to report ~4.

• Nevertheless, if an auditory cue follows the offset of the display and indicates what row should be reported, we are able to report most characters

• This ability was found to decay rapidly with the time between the offset of the display and the sounding of the auditory cue.

• This experiment is important for several reasons: discovery of an **iconic memory** which decays exponentially**;** ability to **direct attention** to a representation stored in memory; last but not least, invention of a **partial report** method validating and surpassing introspection :

« when complex stimuli composed of many alphanumeric characters are displayed with a tachistoscope, subjects enigmatically insist that they saw more than they can remember in retrospect, i.e. report in retrospect »(Sperling, 1960).

« Sperling operationalized a form of introspection expressed by verbal reports of dissatisfaction or fleeting impressions. »(Sackur, 2009).

Psychophysics substantiates rough introspection, but can also qualify it (see course 2010).

Limitations of introspection

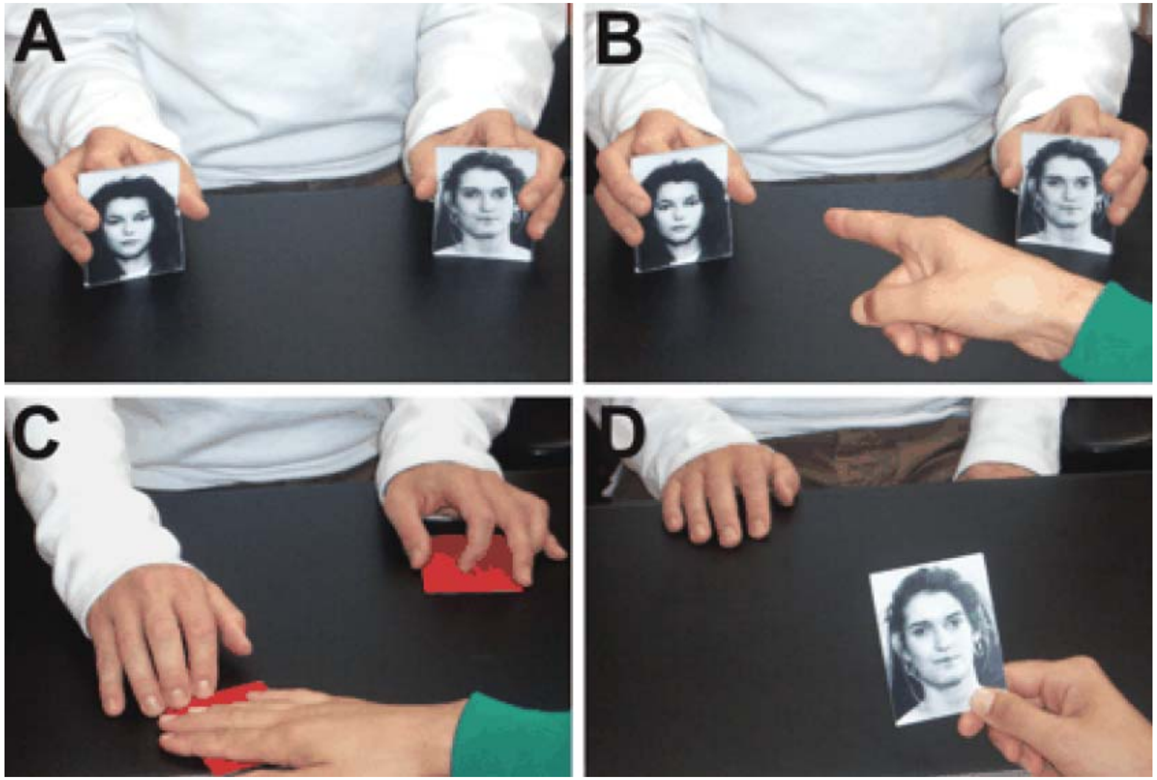
An essential idea is now well-accepted: the introspective ability should be studied for itself, without assuming it is necessarily accurate, but simply as a mental operation the mechanisms and limitations of which remain to be elucidated.

Meta-knowledge can be classified according to its degree of accuracy :

	Level 2 Meta-knowledge:	
Level 1 Knowledge:	**Present and accurate**	**Absent or inaccurate**
Present and accurate	**Knowing that I know:** - reliance on our answers - knowledge of our strategies	**Not knowing that I know:** - subliminal operations - grammar of mother tongue
Absent or inaccurate	**Knowing that I don't know:** - awareness of our errors - awareness of our oversights - judgments of learning	Not knowing that I don't know, or **thinking I know**: - False memories - Fictitious justifications of our behavior

Example of mental fiction:
The explanation of our choices

Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005).
Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310(5745), 116-119.*



A new phenomenon: *choice blindness*

Phase 1. The participant chooses the most attractive face to him (between face pairs chosen on the basis of their similar attractiveness)

Phase 2. The participant receives his card and explains the reasons for his choice.

In 20% of tests, cards are covertly manipulated.

74% of manipulations are not noticed, neither immediately or in retrospect.

The participant then starts giving “explanations” for a choice he did not make! These explanations are given with the same level of detail, the same confidence and the same emotional tonality.

Type	%		
Specific Conf.	13.3		She's radiant. I would rather have approached her at a bar than the other one. I like earrings! [M]
Detailed Conf.	17.3	She looks like an aunt of mine I think, and she seems nicer than the other one. [F]	
Emotional Conf.	9.3		Yes, well, [laughter] she looks very hot in this picture. [M]
Simple Conf.	10.8		Just a nice shape of the face, and the chin. [M]
Relational Conf.	21.3		I thought she had more personality, in a way. She was the most appealing to me. [F]
Uncertainty	11.6	Eh... I don't know. [F]	
Dynamic report	5.2		Oh, [short laughter] Why did I choose her? She looks very masculine! [M]
Original choice	11.2	Because she was smiling. [F]	

Explanations vary from pure confabulation (a feature present in the outcome but which couldn’t have been used in the choice phase) to accurate introspection (but inappropriate for the outcome presented)

Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310(5745), 116-119.*

Another example: « unskilled and unaware of it »

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol, 77(6), 1121-1134.*

In a range of very different tests (humor evaluation, logical reasoning, grammar…), participants with the poorest performance misjudge their incompetence.

Paradoxically, training improves cognitive *and* metacognitive performances, which helps subjects to recognize their incompetence.

Darwin (1871, *The Descent of Man*): « Ignorance more frequently begets confidence than does knowledge »

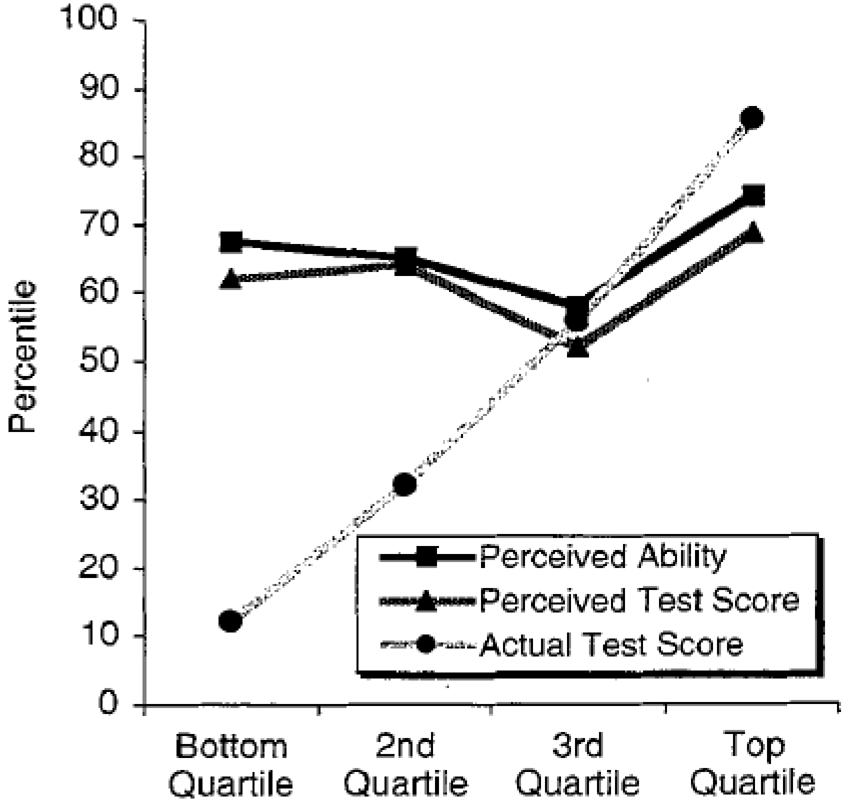
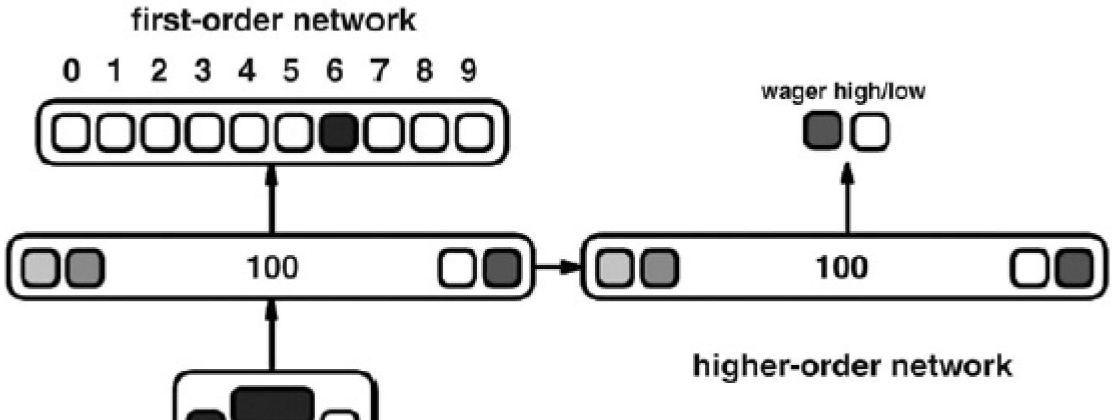


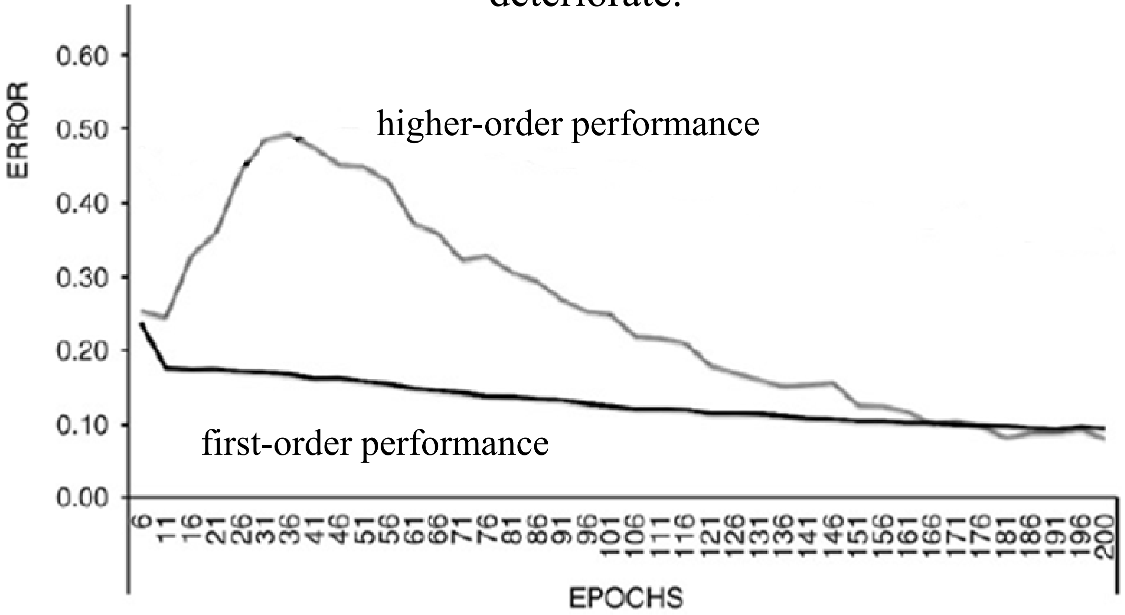
Figure 2. Perceived logical reasoning ability and test performance as a function of actual test performance (Study 2).

An attempt to metacognition modeling by neural networks

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007).
Consciousness and metarepresentation: a computational sketch. *Neural Netw, 20(9), 1032-1039.*



- A first-order network learns to categorize images.
- A higher-order network observes the states of the first-order network and learns to predict errors.
- For a transitory period, first-order performances improve while second-order performances deteriorate.



- This network and its variations (Pasquali and al, *Cognition*, 2010) can capture part of Persaud et al. data (2007) according to which metacognitive wagering comes behind first-order performance.

Central issues in the field of metacognition

- Are metacognitive judgments always true?
  - When do we have authentic introspective access to our mental state?
  - When do we build mental fictions?
  - Why are we unaware that these representations are fictitious?
- What is the format of our self-knowledge?
  - Is there such a thing as a special register of self-knowledge, or do we use the same processes to encode « one-self as another »?
  - Is metacognitive knowledge necessarily conscious?
- What mental and brain architecture underlies metacognitive judgments?
  - What cues are used for these judgments?
  - Can metacognitive decision-making be modelled as some sort of perceptive decision, but based on higher level cues?
  - What are the brain areas concerned?
  - Is the architecture of metacognition specific to human species?
- What are the practical implications of these researches?
  - in particular in the field of education (knowing what I don't know)

Syllabus

- Tuesday 4th, January. Definitions and first paradoxes
- Tuesday 11th, January. Is our introspection ability an illusion?
- Tuesday 18th, January. Links between awareness and metacognition
- Tuesday 25th, January. Links between metacognition and theory of mind
- Tuesday 1st, February. Experimental models of introspection in animals.
- Tuesday 8th, February. Brain mechanisms

Seminar:
Mental fiction psychology and neuropsychology

Shedding light on how introspection can be at variance with reality, particularly among patients suffering from brain damage.

- January 4: **Lionel Naccache** (Hôpital de la Salpêtrière, Paris) : Interpretations and Beliefs Neuropsychology
- January 11: **Olaf Blanke** (Federal Polytechnical School of Lausanne): How the brain computes the self's point of view
- January 18: **Paul Fletcher** (University of Cambridge, UK): Misperceiving and misbelieving: towards an understanding of psychosis
- January 25: **Gilles Fénelon** (Hôpital Henri Mondor, Créteil): hallucinations, illusions and presence sensing during Parkinson's disease
- February 1: **Henrik Ehrsson** (Karolinska Institutet, Stockholm): The construction of an experience of our own body
- February 8: **Predrag Petrovic** (Karolinska Institutet, Stockholm):
- Expectations, beliefs, and the origins of the placebo effect

Some books and journal articles reviewed

Books:

- **Dunlosky, J., & Metcalfe, J. (2008).** ***Metacognition.*** **Sage Publications, Inc.**
- Kahneman D, Slovic P, Tversky A (1982) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.
- Vickers, D. (1979). *Decision processes in visual perception.* London: Academic Press.
- Wegner, D. M. (2003). *The illusion of conscious will.* Cambridge: MIT Press.

Articles:

- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87(3), 215-251*.
- Harvey, N. (1997). Confidence in judgment. *Trends Cogn Sci, 1(2), 78-82.*
- Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51, 102-116.*
- Nisbett, Richard, & Wilson, Timothy. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Sackur, J. (2009), L'Introspection en psychologie expérimentale, *Revue d'Histoire des Sciences*, 62, 2, 5-28
- Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. (2008). The comparative study of metacognition: sharper paradigms, safer inferences. *Psychon Bull Rev, 15(4), 679-691*.
- Smith, J. D. (2009). The study of animal metacognition. *Trends Cogn Sci, 13(9), 389-396.*
- Terrace, H. S., & Son, L. K. (2009). Comparative metacognition. *Curr Opin Neurobiol, 19(1), 67-74.*