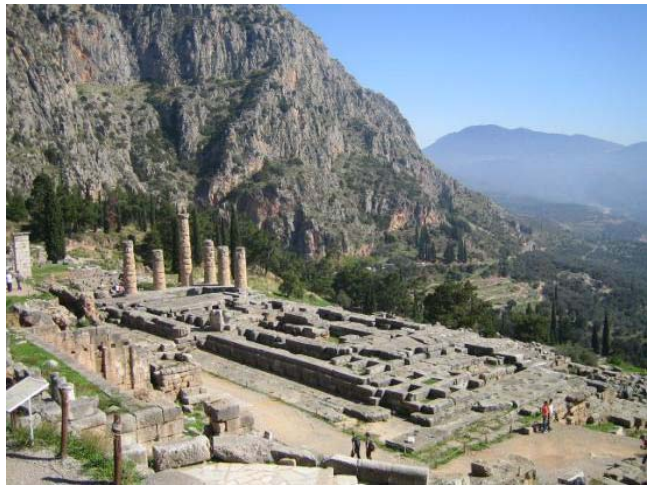
Introspection et métacognition :
Les mécanismes de la connaissance de soi

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

Cours

Définitions et premiers paradoxes

Comment concevoir que nous soyons conscients… d'être conscients?



γνῶθι σεαυτόν : Connais-toi toi-même

Inscription dans le pronaos du Temple d'Apollon à Delphes

"Je m'échappe sans cesse et ne comprends pas bien, lorsque je me regarde agir, que celui que je vois agir soit le même que celui qui regarde, et qui s'étonne, et doute qu'il puisse être acteur et contemplateur à la fois.

André Gide, *Les faux-monnayeurs*

"Etre conscient d'être conscient d'être… Si je sais non seulement que je suis, mais également que je sais que je le sais, alors j'appartiens à l'espèce humaine. Tout le reste en découle – le fleuron de la pensée, la poésie, une vision de l'univers. A cet égard, l'écart entre le singe et l'homme est incommensurablement plus vaste que celui qui sépare l'amibe du singe.

Vladimir Nabokov, *Strong Opinions*

Quelle est l'architecture cérébrale qui nous permet de tourner ainsi nos pensées en direction d'elles-mêmes?

Quelques exemples simples

- Les stratégies sérielles
  - Calculez 13+28. Dans quel ordre avez-vous fait les calculs?
  - Avez-vous détecté une erreur? Avez-vous eu besoin de revenir en arrière?
  - Les stratégies et les plans d'actions sont souvent disponibles à notre conscience, tandis que les opérations élémentaires ne le sont pas.
- Le mot sur le bout de la langue
  - Comment appelle-t-on un être fabuleux, mi-homme, mi-cheval?
  - Nous pouvons ne pas nous souvenir de la réponse, mais néanmoins savoir que nous la connaissons!
- Le contrôle de l'apprentissage et la méta-mémoire
  - Comment décidez-vous de réviser avant un examen?
  - Avez-vous déjà dansé avec une actrice célèbre?

Quelques éléments de définition

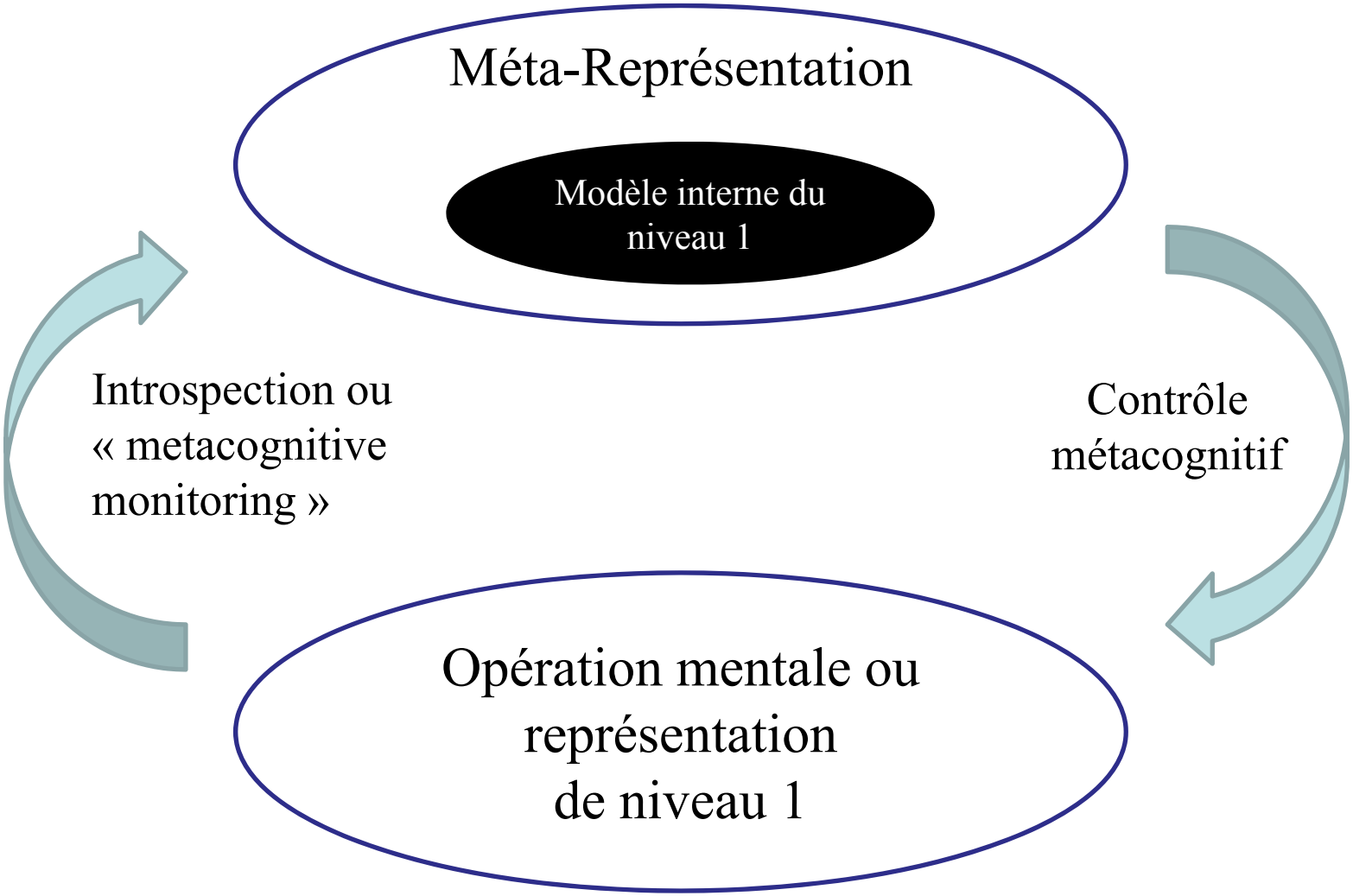
- **Cognition:** l'ensemble des processus mentaux qui nous permettent de traiter des informations (internes ou externes).
- **Métacognition:** L'ensemble des connaissances et des croyances que nous possédons sur nos propres processus cognitifs (passés, présents ou futurs); Les processus qui permettent de les manipuler.
  - **Méta-mémoire: nos connaissances et nos croyances sur nos processus de mémorisation et de récupération en mémoire**
- **Introspection:** (littéralement le regard intérieur)

Capacité d'accéder consciemment à nos opérations mentales, et de les rapporter à nous-mêmes ou à autrui.

  - **Savoir que nous avons compris un théorème**
  - **Juger qu'on est proche de la solution**
- **Contrôle (méta)cognitif:** Capacité de réguler nos propres processus mentaux en fonction de notre introspection
  - **Changer de stratégie**
  - **Faire plus attention**

Un début de cadre théorique

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



La place centrale de l'introspection dans la psychologie du dix-neuvième siècle

• Jusqu'au 19ème siècle, l'introspection était considérée comme la méthode centrale pour l'étude de l'esprit humain, une forme d'observation directe des « faits » mentaux.
• Wilhelm Wundt (1832-1920, Leipzig): l'objet même de la psychologie est l'étude de l'expérience mentale subjective, dont l'introspection est la seule méthode d'étude.
Franz Brentano (1838-1917) promeut une 'psychologie descriptive' ou 'phénoménologie' (avant Husserl) qui consiste en l'étude des phénomènes de la perception intérieure, d'un point de vue « à la première personne ».
• Oswald Külpe (1862-1915), élève de Wundt, chef de file de l'école de Würzburg, développe des méthodes de description verbale de l'introspection (par exemple, décrivez ce qui vous vient à l'esprit quand vous lisez le mot « mètre ») – mais découvre la « pensée sans images »: le sujet ne peut pas toujours rapporter des percepts pertinents.
•Edward Titchener (1827-1927, Cornell), élève de Wundt, prétend que l'introspection est la seule méthode de la psychologie.
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)
• En France, Théodule Ribot (1839-1916) et Alfred Binet (1857-1911) défendent des points de vue similaires: « l'introspection, peut-on dire, est la base de la psychologie, elle caractérise la psychologie d'une manière si précise que toute étude qui se fait par l'introspection mérite de s'appeler psychologique, et que toute étude qui se fait par une autre méthode relève d'une autre science » (A. Binet, *Introduction à la psychologie expérimentale*, 1894)
•Pour Jérôme Sackur, l'introspection n'a jamais disparu des méthodes de la psychologie.

Le paradoxe de Comte

« Il est sensible, en effet, que, par une nécessité invincible, l'esprit humain peut observer directement tous les phénomènes, excepté les siens propres. Car, par qui serait faite l'observation? (…) L'individu pensant ne saurait se partager en deux, dont l'un raisonnerait, tandis que l'autre regarderait raisonner. L'organe observé et l'organe observateur étant, dans ce cas, identiques, comment l'observation pourrait-elle avoir lieu? Cette prétendue méthode psychologique est donc radicalement nulle dans son principe. »

Auguste Comte, *Cours de Philosophie Positive* (1830-1842), Vol. 1, pp. 31-32

« Il aurait pu venir à l'esprit de M. Comte qu'il est possible d'étudier un fait par l'intermédiaire de la mémoire, non pas à l'instant même où nous le percevons, mais dans le moment d'après : et c'est là, en réalité, le mode suivant lequel s'acquiert généralement le meilleur de notre science touchant nos actes intellectuels. Nous réfléchissons sur ce que nous avons fait quand l'acte est passé, mais quand l'impression en est encore fraîche dans la mémoire. (…)
Ce simple fait détruit l'argument entier de M. Comte. »

John Stuart Mill, *Auguste Comte et le positivisme* (1865), pp. 68-69.

Réfutations contemporaines du paradoxe de Comte

Nos processus mentaux sont constitué de multiples processeurs partiellement spécialisés, il n'est donc pas exclu que certains en « observent » d'autres.

Le cortex préfrontal, en particulier, est en situation de recevoir des informations de tous nos autres processus mentaux :

« 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

Cependant, l'observation de Comte pointe vers deux questions intéressantes et ouvertes:

- Le fait même de demander au participant une introspection affecte-t-il le traitement primaire de l'information?
- Il est probablement possible d'utiliser le paradoxe de Comte pour démontrer l'impossibilité d'une introspection parfaite et complète (l'androïde *Data* de Star Trek, qui aurait une mémoire parfaite et un accès parfait aux raisons de toutes ses décisions).

La critique behavioriste

« La psychologie telle que le béhavioriste la voit est une branche purement objective des sciences naturelles. Son but théorique est la prédiction et le contrôle du comportement. L'introspection ne fait pas partie de ses méthodes essentielles, et la valeur scientifique de ses données ne dépend pas de la façon dont elles se prêtent à une interprétation en termes de conscience. »

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

Critique principale: la subjectivité des observations.
« La conséquence du postulat majeur qu'il existe une chose telle que la conscience, et que nous pouvons l'analyser par l'introspection, c'est qu'on trouve autant d'analyses qu'il existe de psychologues. »

John Watson (1925), *Behaviorism*

Quelques éléments de réponse à Watson:
- Sa critique confond l'introspection en tant que *méthode* pour accéder à l'architecture mentale, et l'introspection en tant qu'*objet d'étude.*
- En tant qu'objet d'étude, l'introspection (ainsi que ses limites) est un sujet de recherche parfaitement légitime, et qui conduit à des résultats empiriques reproductibles d'un individu à l'autre.

La métacognition et le programme des sciences cognitives

John Flavell introduit l'étude de la *métamémoire* (1971), et la distinction entre *monitoring* et *regulation* (1976).

En 1979 il propose une première théorisation de la métacognition, qui distingue:
-connaissances métacognitives (conscientes ou non, justes ou fausses)
-expériences conscientes
-buts et tâches
-stratégies et actions.

Influencé par Piaget, il souligne l'importance de la métacognition dans l'éducation chez l'enfant (stratégies de recherche active et de mémorisation des informations).

1960-1990: Vifs débats et recherches actives sur la véracité de l'introspection:

- Pour Nisbett et Wilson (1977), les jugements introspectifs sont très souvent fictifs. Exemple de la préférence pour la droite dans le choix de 4 objets équivalents

- Pour Ericsson et Simon (1980), les rapports verbaux sont adéquats et utiles si l'information rapportée est présente en mémoire à court terme

Une théorie générale du rapport verbal

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

Ericsson et Simon introduisent une classification des tâches introspectives qui distingue le *moment* du rapport verbal (immédiat ou différé), et le *type* de rapport (direct, avec recodage, ou sans 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

Leur revue des données expérimentales suggère que le rapport verbal est fidèle lorsqu'il est *direct* et qu'il décrit le contenu *actuel* de la mémoire à court terme.
Dans ces conditions, la correspondance entre ce que les sujets disent et ce qu'ils font peut être absolument remarquable (par exemple dans un test de tri de cartes [Dulany et O'Connell, 1963], seulement 11 réponses en désaccord avec la verbalisation sur 34408 = 0.03%).

La métacognition et le programme des sciences cognitives

Avancées expérimentales majeures des années 1960-2000: l'invention de **nouvelles mesures expérimentales de l'introspection:**

- **Tâches de méta-mémoire:**
  - *judgment of learning:* Après une phase d'apprentissage, on demande à une personne d'estimer quelle seront ses performances dans un test ultérieur de mémoire.
  - *feeling of knowing:* Juste après n'être pas parvenu à se souvenir d'un item, on demande à la personne d'estimer, prospectivement, si elle saurait le reconnaître parmi plusieurs.

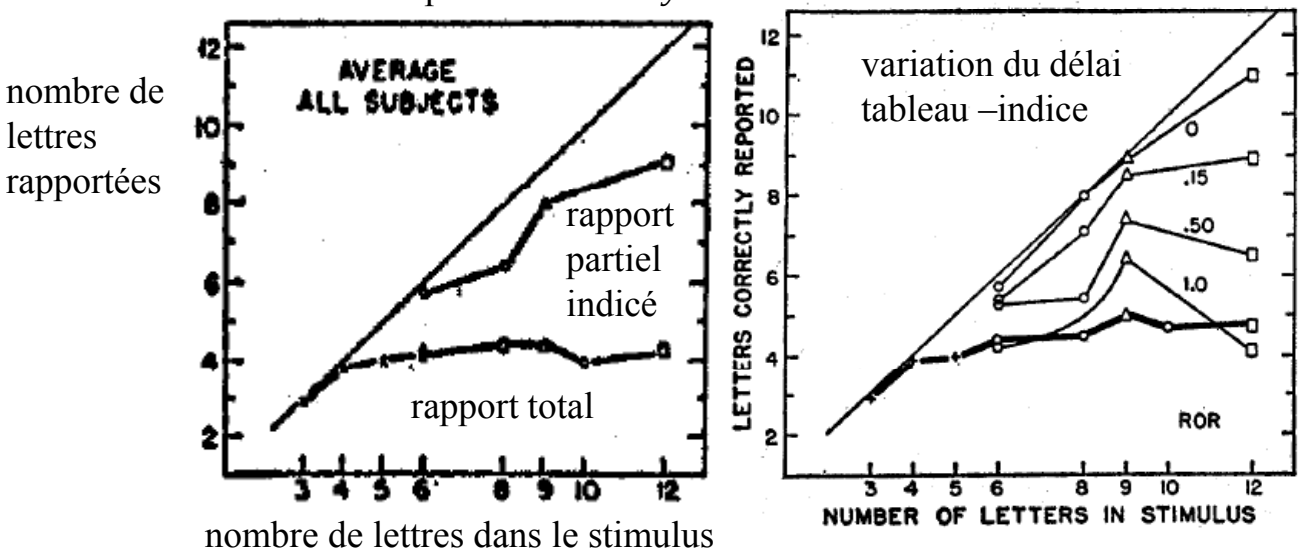
- **Jugements 'de second ordre':**
  - *confidence*: jugement numérique de confiance dans une réponse de premier ordre
  - *wagering*: pari sur la véracité de sa réponse
  - Détection d'erreurs

- Comme le note Jérôme Sackur, la **psychophysique** elle-même fait régulièrement appel à l'introspection (avec rapport verbal ou non-verbal), soigneusement quantifiée et répliquée.
Exemple historiquement important: l'expérience de Sperling (1960).

L'expérience de mémoire iconique de Sperling

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



•Lorsque l'on flashe un tableau de lettres (50 ms), nous ne sommes capables d'en rapporter que ~4.

• Pourtant si un indice auditif suit le tableau et indique quelle ligne rapporter, nous pouvons rapporter la plupart des lettres

• Cette capacité décroît rapidement avec le délai qui sépare le tableau de l'indice auditif.

• L'expérience est importante à plusieurs titres: découverte d'une **mémoire iconique** qui décroît exponentiellement; de la capacité d'**orienter l'attention** vers une représentation en mémoire; mais surtout, invention d'une méthode de **rapport partiel** qui valide et dépasse l'introspection:

« quand des stimuli complexes faits de nombreuses lettres sont présentés au tachistoscope, les sujets soutiennent de manière énigmatique qu'ils ont vu plus que ce dont ils peuvent se souvenir après coup, c'est-à-dire rapporter après coup » (Sperling, 1960).

« Sperling a opérationnalisé une forme d'introspection qui s'exprimait par des rapports verbaux d'insatisfaction ou d'impressions fugitives. » (Sackur, 2009).

La psychophysique corrobore l'introspection brute, mais peut également la nuancer (cf. cours 2010).

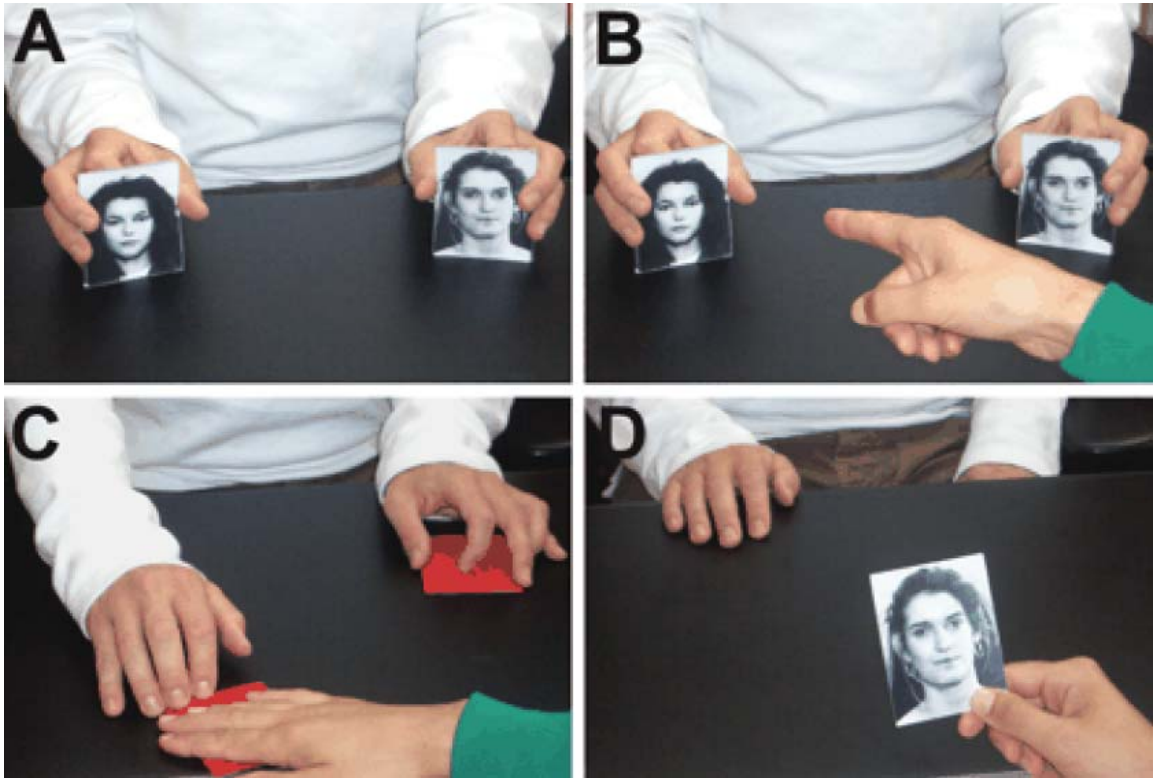
Les limites de l'introspection

Aujourd'hui, une idée essentielle est bien acceptée: il faut étudier la capacité d'introspection pour elle-même, sans supposer qu'elle soit nécessairement juste, mais simplement comme une opération mentale dont les mécanismes et les limites restent à élucider.
Les méta-connaissances peuvent donc être classées selon leur valeur de vérité:

Connaissance de niveau 1:	**Méta-connaissance de niveau 2 :** **Présente et véridique**	**Absente ou erronée**
Présente et véridique	**Savoir que je sais:** - confiance dans nos réponses - connaissance de nos stratégies	**Ne pas savoir que je sais:** - opérations subliminales - grammaire de la langue maternelle
Absente ou erronée	**Savoir que je ne sais pas:** - conscience de nos erreur -conscience de nos oublis -jugements d'apprentissage	Ne pas savoir que je ne sais pas, ou **croire savoir**: -Faux souvenirs - Justifications fictives de nos comportements

Un exemple de fiction mentale: l'explication de nos choix

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.*



Un phénomène nouveau: *choice blindness,* la cécité au choix

Phase 1. La personne choisit le visage qu'elle juge le plus attrayant (parmi deux visages choisis pour être de beauté similaire)

Phase 2. La personne reçoit la carte et explique les raisons de son choix.

Dans 20% des essais, les cartes sont échangées subrepticement.

74% de ces échanges ne sont pas détectés, ni immédiatement, ni rétrospectivement.

La personne se met alors à donner des « explications » d'un choix qu'elle n'a pas fait! Ces explications sont données avec le même niveau de détail, la même confiance, la même tonalité émotionnelle.

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]	

Les explications varient depuis la confabulation pure et simple (un trait qui est présent dans l'image, mais n'a pas pu être utilisé lors de la phase de choix) jusqu'à l'introspection correcte (mais qui n'est plus appropriée à l'image présente)

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.*

Un autre exemple: « inexperimenté et inconscient de l'être »

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.*

Dans une série de test très divers (évaluation de plaisanteries, problèmes logique, grammaire…), les participants les moins habiles méjugent leur incompétence.

Paradoxalement, l'entraînement améliore les performances cognitives *et* métacognitives, ce qui rend les sujets plus conscients de leur incompétence.

Darwin (1871, *La filiation de l'homme*): « L'ignorance suscite la confiance plus souvent que la connaissance elle-même »

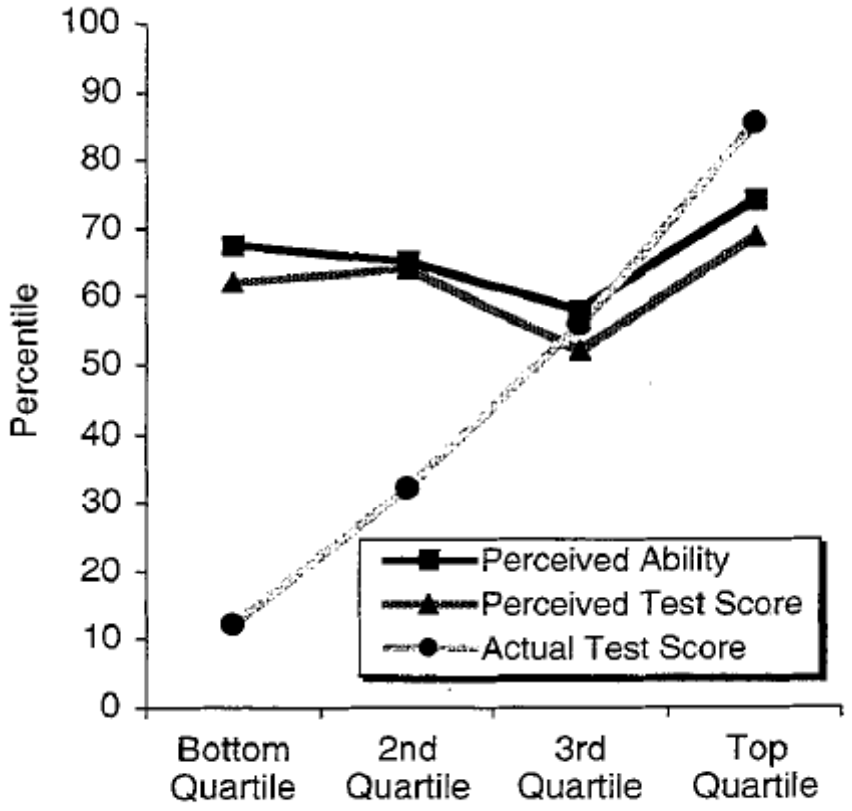
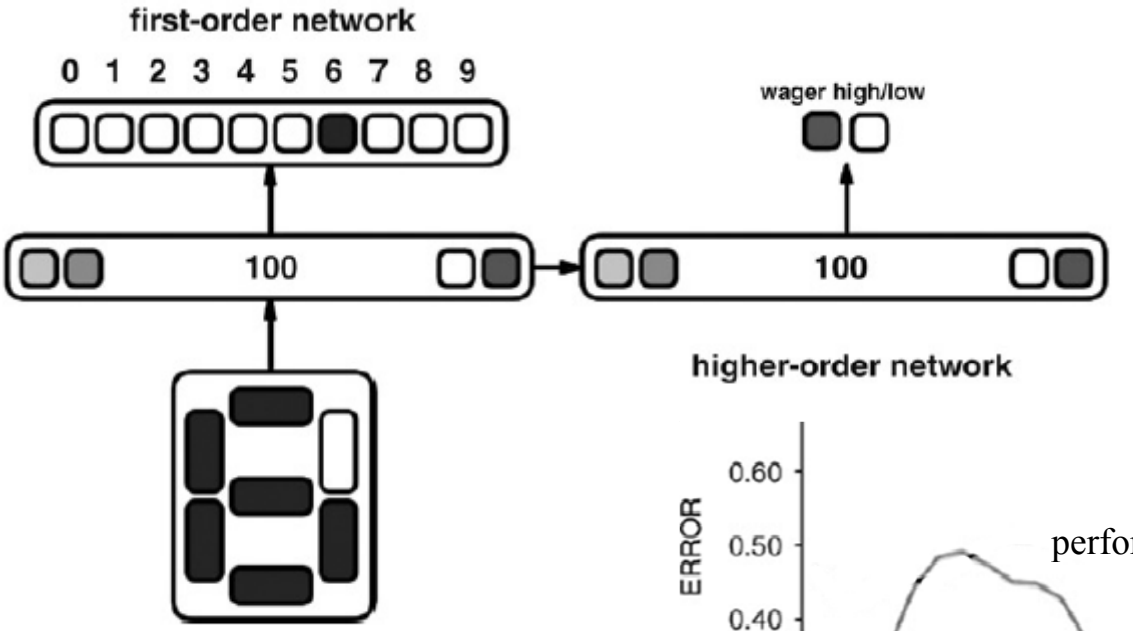


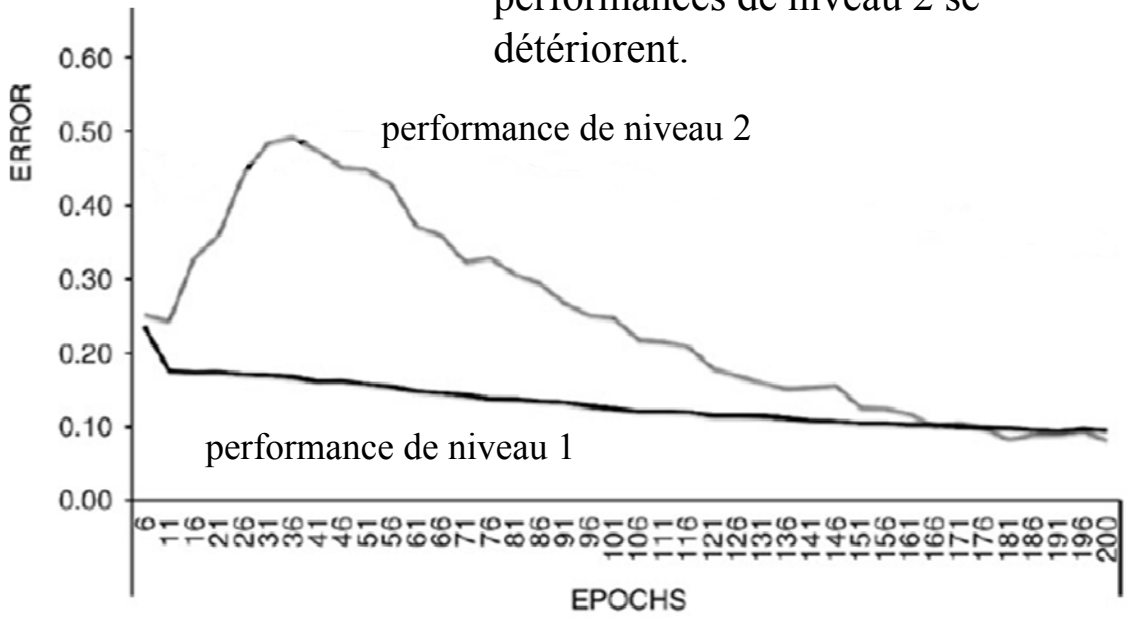
Figure 2. Perceived logical reasoning ability and test performance as a function of actual test performance (Study 2).

Un début de modélisation de la métacognition par un réseau de neurones

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



• Un réseau de niveau 1 apprend à étiqueter des images.
• Un réseau de niveau 2 observe les états du réseau de niveau 1, et apprend à prédire les erreurs.
• Pendant une période transitoire, les performances de niveau 1 s'améliorent tandis que les performances de niveau 2 se détériorent.



• Ce réseau et ses variantes (Pasquali et al, *Cognition*, 2010) peuvent capturer une partie des données de Persaud et al. (2007) selon lesquelles le pari métacognitif est en retard sur la performance de premier ordre.

Questions centrales dans le domaine de la métacognition

- Les jugements métacognitifs sont-ils toujours véridiques?
  - Quand avons-nous un authentique accès introspectif à notre état mental?
  - Quand construisons-nous des fictions mentales?
  - Pourquoi n'avons-nous pas conscience que ces représentations sont fictives?
- Quel est le format de notre connaissance de nous-mêmes?
  - Existe-t-il un registre spécial de la connaissance de soi, ou bien utilisons-nous les mêmes processus pour coder « soi-même comme un autre »?
  - Les connaissances métacognitives sont-elles nécessairement conscientes?
- Quelle architecture mentale et cérébrale sous-tend les jugements métacognitifs?
  - Quels sont les indices utilisés dans ces jugements?
  - Peut-on modéliser la prise de décision métacognitive comme une sorte de décision perceptive, mais fondée sur des indices de plus haut niveau?
  - Quelles sont les aires cérébrales concernées?
  - L'architecture de la métacognition est-elle propre à l'espèce humaine?
- Quelles sont les conséquences pratiques de ces recherches?
  - notamment dans le domaine de l'éducation (savoir ce que je ne sais pas)

Plan du cours

- Mardi 4 Janvier. Définitions et premiers paradoxes
- Mardi 11 Janvier. Notre capacité d'introspection est-elle illusoire ?
- Mardi 18 Janvier. Liens entre conscience et métacognition
- Mardi 25 Janvier. Liens entre métacognition et théorie de l'esprit
- Mardi 1er Février. Modèles expérimentaux de l'introspection chez l'animal
- Mardi 8 Février. Mécanismes cérébraux

Séminaire:
Psychologie et neuropsychologie des fictions mentales

Eclairer la manière dont l'introspection peut s'écarter de la réalité, particulièrement chez les patients atteints de lésions cérébrales.

- 4 Janvier: **Lionel Naccache** (Hôpital de la Salpêtrière, Paris): Neuropsychologie des interprétations et des croyances
- 11 Janvier: **Olaf Blanke** (Ecole Polytechnique Fédérale de Lausanne): How the brain computes the self's point of view
- 18 Janvier: **Paul Fletcher** (University of Cambridge, UK): Misperceiving and misbelieving: towards an understanding of psychosis
- 25 Janvier: **Gilles Fénelon** (Hôpital Henri Mondor, Créteil): hallucinations, illusions et sensations de présence au cours de la maladie de Parkinson
- 1er Février: **Henrik Ehrsson** (Karolinska Institutet, Stockholm): The construction of an experience of our own body
- 8 Février: **Predrag Petrovic** (Karolinska Institutet, Stockholm): Expectations, beliefs, and the origins of the placebo effect

Quelques ouvrages et articles de revue consultés

Livres:

- **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.*