Psychologie cognitive expérimentale

M. Stanislas DEHAENE, professeur

COURS : INTROSPECTION ET MÉTACOGNITION : LES MÉCANISMES DE LA CONNAISSANCE DE SOI

Connais-toi toi-même, γνῶθι σεαυτόν. La maxime inscrite dans le pronaos du temple d'Apollon à Delphes attire notre attention sur le subtil paradoxe qui entoure le problème de la conscience. Non seulement notre cerveau nous fait prendre conscience de certains aspects du monde extérieur – thème du cours de l'année précédente –, mais il nous permet également d'orienter le faisceau de la conscience en nous-mêmes. *Homo Sapiens sapiens,* nous sommes conscients d'être conscients. Talentueux peintre de l'introspection, Vladimir Nabokov résume en quelques mots, dans *Strong Opinions*, cet étrange effet miroir :

> Être 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.

L'objectif du cours 2011 était de faire le point sur les mécanismes psychologiques et cérébraux qui nous permettent d'orienter ainsi le projecteur de la conscience vers nous-mêmes. Quelles en sont les limites, et que savons-nous réellement de nous-mêmes ? Quels processus cognitifs sont accessibles à un jugement introspectif d'ordre supérieur, « métacognitif » ?

Le vocabulaire de la métacognition

Si la *cognition* peut se définir, schématiquement, comme l'ensemble des processus mentaux qui nous permettent de traiter des informations (internes ou externes), alors la *métacognition* pourrait se définir comme l'ensemble des connaissances et des croyances que nous possédons sur nos propres processus cognitifs (passés, présents ou futurs), ainsi que les processus qui permettent de les manipuler. La *méta-mémoire*, par exemple, concerne l'ensemble de nos connaissances et de nos croyances sur nos propres processus de mémorisation et de récupération en mémoire. Lorsqu'un étudiant, doutant de ses connaissances, décide de réviser avant un

examen, il porte un jugement d'ordre métacognitif sur les faiblesses de sa mémoire. Calculez 23 + 18. Sans doute pouvez-vous rapporter l'ordre et la nature des opérations que vous effectuez : elles sont disponibles à un niveau métacognitif. Par contre, vous ne disposez d'aucune introspection sur la manière dont vous réalisez les calculs élémentaires tels que 2 + 1.

Pour chaque opération mentale de niveau *n*, la métacognition suppose l'existence d'une représentation mentale de niveau $n + 1$ ou « *méta-représentation* » des opérations mentales exécutées au niveau inférieur. Selon ce schéma hiérarchique, l'*introspection* (ou *cognitive monitoring* dans la littérature anglophone) s'exerce de bas en haut : elle consiste à mettre au jour la méta-représentation sur la base d'un accès aux informations du niveau inférieur (accès qui peut être partiel ou illusoire). Ainsi, nous exerçons notre introspection lorsque nous détectons une erreur dans notre calcul. Inversement, le *contrôle métacognitif* s'exerce de haut en bas : il consiste à exploiter la méta-représentation afin de modifier la stratégie et les opérations mentales exécutées au niveau inférieur. Nous l'exerçons par exemple lorsque, prenant conscience de nos erreurs, nous décidons de ralentir et de faire plus attention au prochain essai.

Un bref historique

Jusqu'à la fin du XIX^e^ siècle, l'introspection est considérée comme la méthode centrale d'étude de l'esprit humain. De nombreux psychologues, qui ne prennent guère la mesure de l'étendue des opérations non-conscientes, considèrent possible l'observation directe des faits mentaux. Pour Wilhelm Wundt (1832-1920), l'objet même de la psychologie est l'étude de l'expérience mentale subjective, et l'introspection en constitue la seule méthode. 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 subjectif, « à la première personne ». Oswald Külpe (1862-1915), élève de Wundt et chef de file de l'école de Würzburg, développe des méthodes de description verbale de l'introspection, quoiqu'il découvre alors une première limite de l'introspection : lors de la « pensée sans images », le sujet ne peut pas toujours rapporter ses percepts.

Edward Titchener (1827-1927) à Cornell, Edwin Boring (1886-1968), Théodule Ribot (1839-1916) et même Alfred Binet (1857-1911) en France 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).

Cependant, cette prétendue spécificité de la psychologie fait d'emblée débat. Auguste Comte lui oppose un argument connu aujourd'hui sous le nom de *paradoxe de Comte*. Selon lui, « 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 ? » (Auguste Comte, *Cours de philosophie positive* [1830-1842], Vol. 1, pp. 31-32).

Une réponse vigoureuse à cette critique sera fournie par John Stuart Mill : « 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). Les neurosciences cognitives contemporaines pourraient ajouter qu'il n'y a rien d'impossible à ce qu'un circuit cérébral, situé par exemple dans le cortex préfrontal, reçoive et régule les informations issues d'autres circuits hiérarchiquement inférieurs.

Le paradoxe de Comte n'en donc pas un, mais la suspicion est née : l'introspection serait une méthode scientifique inadéquate. Dans une violente critique demeurée célèbre, le chef de file du behaviorisme John Watson l'énonce avec force : « 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. » *Exit* l'introspection : la subjectivité de ses observations rendrait impossible toute construction scientifique.

Rétrospectivement, il nous semble toutefois que la critique de Watson confonde l'introspection en tant que méthode, et l'introspection en tant qu'objet d'étude. L'introspection n'est certainement pas une méthode infaillible pour accéder à l'architecture mentale : même un psychologue parfaitement entraîné ne saurait rapporter fidèlement ses processus mentaux, sinon la psychologie expérimentale serait une tâche bien aisée ! Cependant, les performances et les limites de l'introspection constituent un grand sujet de recherches, parfaitement légitime, dont nous verrons qu'il conduit à des résultats empiriques reproductibles d'un individu à l'autre.

Dès 1971, à l'aube de la révolution cognitive, John Flavell introduit l'étude de la méta-mémoire. En 1979 il propose une première théorisation de la métacognition, qui distingue les connaissances (conscientes ou non, justes ou fausses), les expériences, les buts, les tâches, les stratégies et les actions métacognitives. Influencé par Piaget, il souligne déjà l'importance de la métacognition dans l'éducation chez l'enfant. En effet, sur la base de ce qu'il comprend de lui-même, l'enfant est amené à concevoir, à tort ou à raison, des stratégies d'apprentissage et de recherche en mémoire qui influencent ses performances.

Les années 1960-1990 voient naître de vifs débats sur la fidélité de l'introspection. Pour Nisbett et Wilson (1977), les jugements introspectifs sont très souvent fictifs, donc inutiles. Pour Ericsson et Simon (1980), par contre, les rapports verbaux sont souvent adéquats dès lors que l'information rapportée est présente en mémoire à court terme. 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 avec l'expérience initiale). Leur revue des données expérimentales suggère que le rapport verbal peut être extrêmement fidèle lorsqu'il est direct et qu'il décrit le contenu présent de la mémoire à court terme. Dans ces conditions, il existe une correspondance étroite entre ce que les sujets disent et ce qu'ils font : l'introspection est crédible et utile.

Les méthodes de la métacognition

Depuis les années 1970, la validité des études de l'introspection s'est encore renforcée avec l'avènement de paradigmes expérimentaux rigoureux de mesure de la métacognition. Dans le domaine de la méta-mémoire, le jugement d'apprentissage (*judgment of learning*) demande au participant, après une phase d'apprentissage, d'estimer quelle seront ses performances dans un test ultérieur de mémoire. La mesure du sentiment de savoir (*feeling of knowing*), elle, requiert d'estimer, juste après qu'un participant ait échoué à se souvenir d'un item, s'il saurait le reconnaître parmi plusieurs. Dans les deux cas, la prévision introspective peut être comparée à la réalité objective mesurée quelques minutes plus tard.

Plus généralement, les *jugements de second ordre* requièrent d'estimer son degré de confiance dans une réponse antérieure (dite de premier ordre), de parier sur la véracité de sa réponse (*wagering*), ou de détecter ses propres erreurs. Comme le note Jérôme Sackur, la psychophysique elle-même fait régulièrement appel à l'introspection sous la forme d'un rapport subjectif, verbal ou non-verbal, qui peut être soigneusement quantifié, répliqué, et comparé quantitativement à la réalité objective.

Notre capacité d'introspection est-elle illusoire ?

Bien que la métacognition soit devenue un élément essentiel de la psychologie expérimentale contemporaine, celle-ci ne l'accepte qu'en tant qu'objet d'étude. On s'accorde à penser qu'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. L'introspection pouvant être fausse, connaissances et méta-connaissances peuvent donc être classées selon leur valeur de vérité : je peux « savoir que je sais » (confiance dans mes réponses, connaissance de mes stratégies) et « savoir que je ne sais pas » (conscience de mes erreurs et de mes oublis), mais également « ne pas savoir que je sais » (ignorance des opérations subliminales ou préconscientes) et même « ne pas savoir que je ne sais pas », autrement dit « croire savoir » (faux souvenirs, justifications fictives de mes comportements).

De nombreux exemples de telles fictions mentales ont été cités dans le cours. Mentionnons l'expérience de Johansson et coll. sur la cécité au choix (*choice blindness*) (Johansson, Hall, Sikstrom & Olsson, 2005), dans laquelle le sujet de l'expérience est amené à décrire avec force détails les raisons pour laquelle il a choisi l'une de deux photographies de jeunes femmes… alors que, par un tour de passe-passe, c'est l'image qu'il n'a *pas* choisie qui lui a été donnée ! La personne se met ainsi à donner, avec le même niveau de détail, la même confiance, la même tonalité émotionnelle, des explications d'un choix qu'elle n'a pas fait. Une autre expérience classique montre que nous pouvons être à la fois inexpérimentés et inconscients de l'être (*unskilled and unaware*) (Kruger & Dunning, 1999). Dans une série de test très divers (évaluation de plaisanteries, problèmes de logique, de grammaire…), ce sont les participants les *moins* habiles qui surestiment le *plus* leur niveau de réussite, méjugeant ainsi leur incompétence. Paradoxalement, l'entraînement, qui améliore les performances objectives, rend aussi les sujets mieux conscients de leur incompétence et peut ainsi diminuer l'estimation subjective de la performance.

Une troisième expérience classique démonte le sentiment que nous avons parfois d'être proches de la solution d'un problème (Metcalfe, 1986). Au cours de la résolution d'énigmes logico-mathématiques, toutes les 10 secondes, le sujet note sur une échelle de 0 à 10 son sentiment d'être plus ou moins « chaud » ou « froid ». Non seulement ce « réchauffement » subjectif n'est pas une bonne indication que la solution est proche, mais c'est l'inverse : l'impression d'être « chaud » est plus élevée avant une réponse erronée qu'avant une réponse correcte !

L'étude de la méta-mémoire indique que, sans être très précis, les jugements métacognitifs ont souvent une faible corrélation avec les performances de premier ordre. Par exemple, lorsque nous avons l'introspection d'avoir un « mot sur le bout de langue » (*tip of the tongue state* ou TOT), la reconnaissance ultérieure du mot est effectivement meilleure que lorsque nous n'avons pas une telle introspection. De même le « sentiment de savoir » (*feeling of knowing* ou FOK) n'est-il pas toujours faux : la performance objective, mesurée par exemple dans un test ultérieur de reconnaissance d'une chaîne de caractères, varie de façon monotone avec le « sentiment de savoir » (Koriat, 1993). Cependant, ce dernier n'est pas toujours correctement calibré. On observe généralement une surestimation systématique de nos compétences réelles, doublée parfois d'une sous-estimation de notre intuition des items les plus mal maîtrisés. Il en résulte un « effet difficile-facile » (*hard-easy effect*) : la mémoire des items faciles est surestimée, tandis que celle des items difficiles est sous-estimée. L'introduction d'un délai de mémoire peut faciliter le jugement métacognitif. En effet, immédiatement après avoir étudié, nous sommes dominés par un sentiment de savoir qui est très souvent erroné.

Comment fonctionne le sentiment de savoir ? Il n'est en aucun cas fondé sur un accès direct à l'opération de notre mémoire à long terme – nous en ignorons les mécanismes, et nous nous trompons systématiquement sur l'influence de certains facteurs, tels que l'importance d'alterner des séances d'apprentissage et de test (Karpicke & Roediger, 2008). Seule l'expérimentation sur nous-mêmes nous permet d'évaluer nos connaissances. Une théorie plausible (Koriat, 1993) propose que nous développions progressivement un ensemble d'heuristiques qui fournissent des indices partiels, mais pas nécessairement optimaux, sur le fonctionnement de notre mémoire. Ainsi, le sentiment de savoir serait issu d'au moins deux indices : la familiarité du problème et l'accès conscient à des connaissances partielles pertinentes. Le caractère plus ou moins approprié de ces indices peut expliquer, au moins en partie, pourquoi le jugement métacognitif apparaît si souvent mal calibré (Gigerenzer, Hoffrage & Kleinbolting, 1991).

La théorie présentée par Ericsson et Simon (1980), comme celle de l'espace de travail neuronal global que j'ai développée avec Jean-Pierre Changeux (Dehaene & Changeux, 2011), suggère qu'il doit exister au moins un cas où notre capacité d'introspection devrait être précise. Selon ces théories, le contenu présent ou très récent de la mémoire de travail devrait être directement accessible à l'introspection et au rapport verbal. Une série d'expériences menées en collaboration avec Jérôme Sackur et Mariano Sigman le confirme : les participants à une expérience de temps de réaction disposent d'une excellente introspection de la durée des étapes de traitement conscient (Corallo, Sackur, Dehaene & Sigman, 2008 ; Marti, Sackur, Sigman & Dehaene, 2010). Cependant, de très nombreuses informations échappent à cet espace de travail, d'une part parce qu'il est lent et sériel, d'autre part parce que, par définition, il n'a pas accès aux traitements non-conscients, qui constituent la

majorité de nos opérations mentales (informations non-attendues, transitoires, codées par des processeurs spécialisés ou par leurs connexions, etc. ; voir le cours de 2009).

En conclusion : sans être totalement illusoire, notre introspection est souvent fausse, car elle ne dispose pas d'un accès direct aux mécanismes qu'elle prétend pourtant connaître. Notre introspection semble restreinte au contenu de notre mémoire de travail. La plupart de nos jugements métacognitifs s'appuient sur une reconstruction, fondée sur des indices partiels issus de notre expérience passée, et donc sujette à caution.

Conscience et métacognition

Jusqu'ici, nous nous sommes intéressés aux opérations métacognitives conscientes qui s'appuient sur un rapport verbal. Cependant, la métacognition requiert-elle nécessairement une analyse consciente ? Ou bien certaines opérations métacognitives pourraient-elles se dérouler inconsciemment ? La question paraît étrange, tant il nous semble, par définition, qu'à chaque fois que nous « plongeons en nous-mêmes », il s'agit d'un acte d'introspection consciente. Le modèle de l'espace de travail global implique que toute représentation consciente est rendue disponible à l'ensemble des traitements réflexifs. Cependant, l'inverse est-il nécessairement vrai ? Oui, selon la théorie de la « pensée d'ordre supérieur » (*higher-order thought* ou HOT) proposée par David Rosenthal (Lau & Rosenthal, 2011). Celle-ci implique en effet que toute représentation X, dès lors qu'elle est intégrée à une pensée métacognitive d'ordre supérieur du type « je suis en train de penser que X », est nécessairement consciente.

Cependant, cette théorie semble contestable. Dans la définition des processus métacognitifs, rien ne semble interdire que des processus relativement élémentaires de surveillance et de supervision des autres opérations mentales puissent être automatiques et non conscients. De fait, il est possible de formaliser mathématiquement, dans le cadre de la théorie de la détection du signal, certaines opérations métacognitives élémentaires telles que le jugement de confiance ou la détection des erreurs (Galvin, Podd, Drga & Whitmore, 2003). La théorie montre que tout jugement de premier ordre s'accompagne de signaux exploitables pour un jugement de second ordre. Dans la mesure où un stimulus subliminal peut conduire à des performances de premier ordre supérieures au niveau du hasard, il n'y a donc aucune raison théorique pour laquelle ce ne serait pas également possible pour les performances de second ordre.

Qu'en est-il sur le plan empirique ? Plusieurs expériences montrent qu'il existe des stimuli subliminaux qui dissocient la performance de premier ordre, meilleure que le hasard, du jugement de confiance qui reste nul (Kunimoto, Miller & Pashler, 2001 ; Persaud, McLeod & Cowey, 2007). Dans ces expériences, le jugement introspectif est en corrélation avec la conscience du stimulus, ce qui n'est guère surprenant – rapport verbal et conscience sont étroitement attachés. Peut-on pour autant en conclure, avec ces auteurs, que le jugement de confiance *mesure* la conscience ? Il nous semble que non, car le point crucial n'est jamais testé : la confiance subjective n'est-elle *jamais* meilleure que le hasard lors des essais où les sujets rapportent n'avoir rien vu ?

Trois expériences récentes répondent à cette question en démontrant l'existence de processus métacognitifs non-conscients. Kanai *et al.* (2010) sont les premiers à dissocier trois paramètres : réponse de type I (objective), réponse de type II

(métacognitive), et visibilité subjective (conscience). Leurs résultats montrent que, dans une tâche primaire à choix forcé, le jugement de confiance dans sa réponse peut être meilleur que le hasard, alors même que le sujet déclare n'avoir pas vu le stimulus. Dans des conditions d'invisibilité causées par un détournement de l'attention, les participants modulent correctement leur confiance en fonction de leur performance, alors même qu'ils n'ont pas perçu le stimulus. Ainsi, le jugement de confiance n'est pas identique au jugement de visibilité. Dans sa thèse en cours au laboratoire, Lucie Charles obtient des résultats similaires : les participants à une expérience de comparaison de nombres sont capables de juger, avec une performance meilleure que le hasard, s'ils ont fait une erreur ou pas, même dans les essais où le chiffre est masqué et demeure subjectivement invisible. Toutefois, la détection d'erreurs s'améliore brutalement lorsque le seuil de conscience est franchi, et une réponse cérébrale caractéristique, la négativité à l'erreur, n'est présente qu'en réponse aux stimuli visibles. Ainsi, il est probable que plusieurs mécanismes métacognitifs, certains conscients et d'autres non-conscients, sous-tendent la détection d'erreurs.

Logan et Crump (2010) parviennent à la même conclusion. Selon eux, lorsque nous tapons à la machine, deux mécanismes de contrôle métacognitif coexistent : la « boucle externe », consciente, de haut niveau, sélectionne les mots et vérifie que le bon message a été écrit, tandis que la « boucle interne », non-consciente, vérifie les détails de la performance motrice. Effectivement, lorsqu'un programme informatique corrige automatiquement les fautes de frappe ou, au contraire, insère des erreurs, le participant qui n'en est pas informé s'attribue ces réussites et ces échecs subjectifs (boucle externe), sans prendre conscience qu'il n'y est pour rien. Cependant, le temps de réaction continue de se ralentir après une erreur *objective*, même si celle-ci n'est pas détectée consciemment – ce qui valide l'hypothèse d'un mécanisme métacognitif non-conscient.

En résumé, outre le système métacognitif dont nous avons conscience, et qui nous permet de produire des introspections verbales (fréquemment fictives), notre cerveau contient des mécanismes non-conscients de supervision de nos processus mentaux. Une estimation élémentaire de l'incertitude semble accompagner chaque jugement perceptif, même inconscient. Chaque aire cérébrale pourrait coder non seulement une estimation de nos perceptions ou de nos actions, mais également l'incertitude associée à cette estimation, et peut-être même toute la distribution de probabilité associée (Ma, Beck, Latham & Pouget, 2006). Notre cerveau comprend également des systèmes automatisés de détection des erreurs, et il se pourrait qu'un signal d'erreur de prédiction soit présent dans chaque aire cérébrale (Friston, 2005). Ces mécanismes évaluent et ajustent sans cesse nos comportements, sans qu'il soit nécessaire que nous en prenions conscience.

Métacognition et théorie de l'esprit

La métacognition consciente implique de se représenter son propre esprit en train de représenter une information (« *je crois* avoir oublié mes clés »). Le format de ces méta-représentations semble très similaire à celui que l'on suppose sous-tendre la représentation des pensées d'autrui (« *il croit* que j'ai oublié mes clés »). Dans les deux cas, la représentation mentale doit spécifier l'agent (moi ou un autre), l'attitude mentale (croire, savoir…), et la proposition examinée. Se pourrait-il donc que nous

utilisions le même format de représentation mentale et les mêmes aires cérébrales pour représenter notre esprit et celui des autres ? La réflexion métacognitive consciente et la « théorie de l'esprit » (*theory of mind*) feraient-elles appel, au moins en partie, aux mêmes mécanismes ?

Plusieurs arguments empiriques suggèrent effectivement que la connaissance de soi et la connaissance de l'autre sont étroitement liées. Tout d'abord, elles se développent simultanément chez l'enfant : c'est au même âge que les enfants commencent à comprendre l'esprit des autres et à disposer d'une représentation métacognitive de leur propre compétence (Gopnik & Astington, 1988). Des résultats très récents suggèrent que c'est un âge très précoce, vers 7 mois, que se met en place la théorie de l'esprit des autres. Dès cet âge, l'enfant représente ses propres connaissances et celles des autres dans le même format, en sorte qu'elles interfèrent (Kovacs, Teglas & Endress, 2010).

Il existe également des preuves d'une généralisation de la connaissance de soi à la connaissance de l'autre chez l'enfant de 12 mois (Meltzoff & Brooks, 2008). À cet âge, les enfants suivent du regard un adulte lorsque celui-ci tourne la tête avec les yeux ouverts, mais pas avec les yeux fermés – ce qui suggère qu'ils comprennent peut-être ce que veut dire « voir ». Cependant, ils suivent également du regard une personne qui porte un bandeau sur les yeux. Peut-être cela signifie-t-il qu'ils ont besoin d'une expérience personnelle pour comprendre ce qu'éprouvent les autres. En effet, ils ont déjà fait l'expérience de fermer les yeux alors que, n'ayant jamais fait l'expérience d'avoir les yeux bandés, ils ne comprennent pas que porter un bandeau empêche de voir. Meltzoff et Brooks (2008) confirment cette hypothèse en entraînant des enfants, soit avec un bandeau opaque, soit avec un bandeau transparent ou doté d'une ouverture. Cette expérience personnelle de l'enfant modifie sa compréhension de l'esprit des autres – ce qui implique que la représentation de soi et la théorie de l'esprit des autres partagent des représentations communes.

Un autre argument provient de l'analyse des réseaux cérébraux impliqués. Un réseau comprenant le cortex préfrontal antéro-mésial, le précuneus, la jonction temporo-pariétale (particulièrement à droite) et la partie antérieure du lobe temporal est impliqué dans la théorie de l'esprit et notamment dans les tâches de fausse croyance. Or la réflexion métacognitive sur soi-même active également une fraction de ce réseau (Jenkins, Macrae & Mitchell, 2008 ; Ochsner *et al.*, 2004 ; Vogeley *et al*., 2004), particulièrement le cortex préfrontal frontopolaire et ventromésial. La représentation des erreurs des autres évoque une « négativité à l'erreur », associée au cortex cingulaire antérieur, similaire à celle évoquée plus classiquement par nos propres erreurs (van Schie, Mars, Coles & Bekkering, 2004). Les autistes, qui ont une représentation déficiente des pensées des autres, souffrent également d'anomalies de la représentation du soi dans le cortex préfrontal mésial (Lombardo *et al*., 2010).

Ainsi, métacognition et représentation des autres partagent certains mécanismes cérébraux. Cependant, l'interprétation de ce recouvrement reste ambigüe. Soit nous disposons d'une représentation détaillée de nous-mêmes et nous utilisons ce « réseau du soi » pour simuler l'esprit des autres et tenter de le comprendre ; ou bien, nous ne disposons pas d'un système spécifique d'introspection, mais notre connaissance est fondée sur l'auto-observation répétée : nous représentons notre comportement et inférons notre état d'esprit comme nous le ferions de celui d'une autre personne, mais nous disposons simplement d'un peu plus de données sur nous-mêmes que sur les autres.

Quoi qu'il en soit, la capacité de prendre conscience de nos capacités et de nos limites joue un rôle essentiel dans le dialogue avec les autres. Une expérience

récente montre que deux personnes qui observent le même stimulus peuvent parvenir, par le dialogue, à une décision psychophysique optimale, meilleure que celle que prendrait un seul des deux protagonistes (Bahrami *et al.*, 2010). Cette réussite optimale dépend d'une représentation explicite, par chacun des protagonistes, de leur propre performance et du degré de confiance qu'ils peuvent lui accorder. L'échange verbal de ces informations subjectives implique qu'elles soient mises dans un format de représentation commun à soi et à l'autre. Il est donc permis de spéculer que l'introspection (fictive ou pas) et la théorie de l'esprit soient deux facettes d'un même système de représentation mentale qui joue un rôle essentiel dans le dialogue social propre à l'espèce humaine.

Introspection et métacognition chez l'animal

La conclusion précédente laisse penser que la métacognition consciente et interpersonnelle pourrait être le propre de l'homme. Cependant, qu'en est-il des compétences métacognitives élémentaires et éventuellement non conscientes, telles que le jugement de confiance ou la détection d'erreurs? Des méthodes non-verbales d'analyse de la métacognition seraient-elles applicables à la cognition animale ? Et montreraient-elles que certaines espèces animales disposent d'une introspection ?

Kornell, Son et Terrace (2007) proposent au moins deux manières de tester la métacognition sans langage. D'une part l'introspection peut être évaluée en examinant si l'animal parvient à accorder un degré de confiance à ses propres réponses. D'autre part, le contrôle métacognitif peut être examiné en démontrant que l'animal « sait qu'il ne sait pas » parce qu'il va activement rechercher des informations supplémentaires.

La première approche, fondée sur le jugement de confiance subjective, a fait l'objet d'une série de travaux expérimentaux de J. David Smith à l'université de New York. Le principe en est simple : On fait exécuter aux animaux une décision psychophysique simple à deux choix (par exemple juger si un son est aigu ou grave). On entraîne ensuite les animaux à utiliser une troisième réponse qui leur permet de recevoir, quoi qu'il arrive, un petit renforcement fixe. Résultat : de nombreuses espèces animales apprennent à utiliser cette « échappatoire » à bon escient. Un dauphin, par exemple (Smith *et al*., 1995) ne l'utilise que pour refuser spécifiquement les essais difficiles. L'animal présente également des réponses d'hésitation (il nage lentement et secoue même la tête !), précisément en réponse aux stimuli pour lesquels ses performances sont les plus faibles.

Une interprétation possible des expériences de Smith est donc que l'animal « sait » quels essais sont difficiles pour lui, et comprend que, dans ce cas, il vaut mieux choisir la touche d'échappatoire. Malheureusement, cette interprétation n'est pas la seule : le comportement de l'animal reste compatible avec une simple maximisation de la récompense totale (Terrace & Son, 2009). Chez le singe, une forme plus convaincante d'introspection peut être mise en évidence (Hampton, 2001) : dans un test de mémoire, l'animal sait *prévoir*, avant même de répondre, s'il saura répondre correctement ou non, et utilise la touche d'échappatoire sélectivement lors des essais qu'il juge difficile. Il généralise également ce comportement à des situations nouvelles, par exemple lorsque le délai s'allonge, ce qui suggère qu'il peut anticiper qu'il ne saura pas répondre correctement. De même, Kornell et coll. (2007) démontrent une généralisation immédiate du jugement métacognitif des

singes macaques : entraînés avec le paradigme de Smith sur une tâche psychophysique de jugement de taille, ils généralisent immédiatement à une tâche très différente de mémoire à court terme.

Kornell et coll. (2007) introduisent également, pour la première fois, un second critère de contrôle métacognitif : un animal est-il capable de rechercher activement des informations supplémentaires lorsqu'il se rend compte qu'il n'en sait pas assez ? Dans une tâche de mémoire sérielle sur écran tactile, où l'animal doit découvrir par essai et erreur l'ordre qui lui est demandé, les singes macaques apprennent effectivement à utiliser à bon escient une touche « indice » qui indique quel est le prochain élément de la séquence, mais conduit également à une dévaluation de la récompense. À mesure que chaque liste d'images est mémorisée, la proportion d'appuis sur la touche « indice » diminue proportionnellement. À ce stade, l'introduction d'une liste nouvelle augmente considérablement le nombre d'indices demandés. Cette intéressante expérience suggère donc qu'un animal peut « savoir qu'il ne sait pas » et déployer des capacités de contrôle métacognitif afin d'obtenir plus d'informations. Cependant, elle reste également ouverte à une interprétation plus behavioriste, en termes de maximisation de la récompense. Il est clair que nous n'en sommes, dans ce domaine de la métacognition animale, qu'au tout début d'une expérimentation comportementale qui manque encore de force démonstrative.

Mécanismes cérébraux de la métacognition

L'existence de modèles animaux de la métacognition ouvre la porte à un examen de ses mécanismes neuronaux. De fait, les protocoles métacognitifs qui évaluent le degré de confiance de l'animal en ses propres réponses sont devenus suffisamment simples pour conduire à une expérimentation neurophysiologique. Kiani et Shadlen (2009) ont utilisé le paradigme de la « touche échappatoire » de J. David Smith au cours d'une tâche de prise de décision statistique. L'animal doit prendre une décision binaire sur la direction du mouvement. À certains essais, avant même cette décision, une troisième possibilité de réponse apparaît (réponse « échappatoire »). Le choix de cette cible conduit à une récompense fixe mais moindre. Ce choix peut donc être interprété comme un refus de répondre à la tâche principale, ce qui pourrait indiquer que l'animal est incertain et n'a pas confiance en lui. Effectivement, les observations comportementales indiquent que la performance objective est meilleure quand l'animal dispose de l'option de refuser de répondre que quand il n'en dispose pas. Cela signifie qu'à stimulus identique, il a correctement écarté les essais où il se jugeait incapable de répondre.

Quels signaux permettent à l'animal de calculer son niveau de confiance en ses réponses ? Les décharges neuronales dans l'aire latérale intra-pariétale (LIP) reflètent la décision à venir, mais également la confiance que l'animal accorde à cette réponse, et ce, avant même que l'animal sache si l'option échappatoire sera ou non proposée. De plus, les fluctuations d'essai en essai des décharges neuronales, mesurées juste avant la présentation de l'option de refuser de répondre, prédisent le choix de cette option : l'animal choisit sélectivement d'éviter de répondre, précisément lors des essais où ses neurones pariétaux ne discriminent pas suffisamment les deux réponses de la tâche primaire. La vitesse d'augmentation des décharges pendant la présentation du stimulus contribue également de façon indépendante au choix de l'option de refuser de répondre. L'ensemble de ces

résultats est bien décrit par un modèle mathématique simple qui ramène la métacognition au rang d'une simple décision de niveau supérieur.

Ainsi le cortex pariétal comprend-il non seulement des signaux neuronaux de décision, mais également ce que l'on pourrait appeler des « méta-signaux » qui codent pour la confiance que l'on peut avoir dans la décision à venir. Il paraît toutefois probable que l'aire LIP ne soit qu'un élément d'un réseau plus vaste, impliquant vraisemblablement le cortex préfrontal, et dont les autres nœuds n'ont guère fait, pour l'instant, l'objet d'enregistrements cellulaires. De fait, seules des recherches effectuées chez le rat, par enregistrement de neurones du cortex orbitofrontal (Kepecs, Uchida, Zariwala & Mainen, 2008) convergent vers des résultats similaires. Elles confirment que, parmi les nombreuses facultés métacognitives, le jugement de confiance fait partie des opérations suffisamment simples pour être aujourd'hui étudiées chez l'animal et modélisées au niveau neuronal.

Chez l'homme, l'imagerie cérébrale permet d'explorer plus largement les réseaux d'aires cérébrales associés à la métacognition, sans toutefois descendre au niveau du neurone unique. Un faisceau convergent de recherches associe les capacités de réflexion métacognitive aux régions les plus antérieures du cortex préfrontal, et notamment l'aire 10 de Brodmann. Dès 2002, une étude en IRM fonctionnelle du sentiment de savoir (*feeling of knowing*) montre que plusieurs régions, toutes préfrontales, augmentent leur activité en proportion directe de l'impression subjective de savoir (Kikyo, Ohki & Miyashita, 2002). La lésion du cortex préfrontal mésial peut sélectivement diminuer le « sentiment d'effort » associé à une tâche cognitive difficile (Naccache *et al*., 2005). De même, le sentiment subjectif d'avoir vu un stimulus masqué est-il sélectivement détérioré par une lésion préfrontale antérieure (Del Cul, Dehaene, Reyes, Bravo & Slachevsky, 2009). Chez le volontaire sain, la stimulation magnétique transcrânienne du cortex préfrontal dorsolatéral affecte le jugement métacognitif de second ordre sur la confiance dans sa réponse, sans altérer les performances psychophysiques de premier ordre (Rounis, Maniscalco, Rothwell, Passingham & Lau, 2010).

Plus étonnant peut-être, les variations d'une personne à l'autre dans ces capacités d'introspection peuvent être mises en relation avec des variations d'organisation cérébrale (Fleming, Weil, Nagy, Dolan & Rees, 2010). En effet, les personnes les plus efficaces dans leurs jugements de second ordre sont celles qui possèdent le plus de matière grise dans le cortex frontopolaire droit. La compétence métacognitive est également corrélée avec la densité de matière blanche des participants dans la région rostrale du corps calleux, qui interconnecte les régions préfrontales droite et gauche. Enfin, il existe une corrélation entre l'activation cérébrale elle-même et les capacités métacognitives : lorsque les participants à un examen d'IRMf sont classifiés en fonction de la corrélation entre leur introspection et leur réussite objective, ce score est fortement corrélé avec l'activité du cortex frontopolaire droit durant l'exécution de la tâche (Yokoyama *et al.*, 2010).

Le cortex frontopolaire, qui recouvre l'aire 10 de Brodmann, jouerait donc un rôle très particulier dans les décisions métacognitives de second ordre, peut-être en réalisant une accumulation d'évidence sur les signaux issus des mécanismes de décision du premier niveau (Kiani & Shadlen, 2009 ; Rolls, Grabenhorst & Deco, 2010). Il est intéressant de constater que cette aire est d'évolution récente, et que son expansion différencie tout particulièrement l'espèce humaine des autres espèces de primates (Semendeferi, Armstrong, Schleicher, Zilles & Van Hoesen, 2001).

Conclusion

Au terme de ce cours, il apparaît clairement que l'introspection et, plus généralement, les capacités métacognitives sont devenues des domaines de recherche respectables, au statut psychologique incontestable, et qui font l'objet d'investigations approfondies en neurosciences cognitives. La recherche actuelle démontre la complexité des fonctions métacognitives et la diversité des systèmes cérébraux concernés. Dès les aires sensorielles, une forme de métacognition apparaît probablement dans la mesure où chaque aire cérébrale représente non seulement une série de décisions sensorielles élémentaires, mais également l'incertitude qui leur est attachée. Divers réseaux du cortex préfrontal et pariétal interviennent ensuite pour recoder, à un niveau plus élevé, la décision consciente et réfléchie, la confiance dans cette décision, l'anticipation de la récompense, l'auto-détection des erreurs, et la reprise de contrôle métacognitif. Ces réseaux sont présents chez diverses espèces animales et leur permettent de jauger la confiance associée à chaque action. Enfin, le cortex frontopolaire et la jonction pariéto-temporale semblent former un réseau particulièrement développé dans l'espèce humaine qui permet l'auto-évaluation de nos performances et la mise en relation de cette estimation personnelle avec la représentation de la pensée des autres.

Bibliographie succincte

Le cours s'est appuyé sur plusieurs ouvrages et articles essentiels.

Livres

Dunlosky J. & Metcalfe J., *Metacognition,* Sage Publications, Inc, 2008.

Kahneman D., Slovic P., Tversky A., *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, 1982.

Vickers D., *Decision processes in visual perception,* Londres, Academic Press, 1979.

Wegner D.M., *The illusion of conscious will,* Cambridge, MIT Press, 2003.

Articles

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

Harvey N., « Confidence in judgment », *Trends Cogn. Sci.,* 1(2), 1997, 78-82.

Nelson T.O., « Consciousness and metacognition », *American Psychologist,* 51, 1996, 102-116.

Nisbett R. & Wilson T., « Telling more than we can know: Verbal reports on mental processes », *Psychological Review*, 84, 1977, 231-259.

Sackur J., « L'Introspection en psychologie expérimentale », *Revue d'histoire des sciences*, 62(2) 2009, 5-28.

Smith J.D., Beran M.J., Couchman J.J. & Coutinho M.V., « The comparative study of metacognition: sharper paradigms, safer inferences », *Psychon. Bull. Rev.,* 15(4), 2008, 679-691.

Smith J.D., « The study of animal metacognition. *Trends Cogn. Sci.,* 13(9), 2009, 389-396.

Terrace H.S. & Son L.K., « Comparative metacognition », *Curr. Opin. Neurobiol.,* 19(1), 2009, 67-74.

Principaux articles cités

Bahrami B., Olsen K., Latham P.E., Roepstorff A., Rees G. & Frith C.D. « Optimally interacting minds », *Science,* 329(5995), 2010, 1081-1085.

Corallo G., Sackur J., Dehaene S. & Sigman M., « Limits on introspection: distorted subjective time during the dual-task bottleneck », *Psychol. Sci.,* 19(11), 2008, 1110-1117.

Dehaene S. & Changeux J.P., « Experimental and theoretical approaches to conscious processing », *Neuron,* 70, 2011, 200-227.

Del Cul A., Dehaene S., Reyes P., Bravo E. & Slachevsky A., « Causal role of prefrontal cortex in the threshold for access to consciousness », *Brain,* 132, 2009, 2531-2540.

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

Fleming S.M., Weil R.S., Nagy Z., Dolan R.J. & Rees G. « Relating introspective accuracy to individual differences in brain structure », *Science,* 329(5998), 2010, 1541-1543.

Friston K., « A theory of cortical responses », *Philos. Trans. R. Soc. Lond. B. Biol. Sci.,* 360(1456), 2005, 815-836.

Galvin S.J., Podd J.V., Drga V. & Whitmore J. « Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions », *Psychon. Bull. Rev.,* 10(4), 2003, 843-876.

Gigerenzer G., Hoffrage U. & Kleinbolting H., « Probabilistic mental models: a Brunswikian theory of confidence », *Psychol. Rev.,* 98(4), 1991, 506-528.

Gopnik A. & Astington J.W., « Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction », *Child Dev.,* 59(1), 1988, 26-37.

Hampton R.R. « Rhesus monkeys know when they remember », *Proc. Natl. Acad. Sci. USA,* 98(9), 2001, 5359-5362.

Jenkins A.C., Macrae C.N. & Mitchell J.P. « Repetition suppression of ventromedial prefrontal activity during judgments of self and others », *Proc. Natl. Acad. Sci. USA,* 105(11), 2008, 4507-4512.

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

Kanai R., Walsh V. & Tseng C.H., « Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness », *Conscious Cogn*., 2010.

Karpicke J.D. & Roediger H.L. 3rd, « The critical importance of retrieval for learning », *Science,* 319(5865), 2008, 966-968.

Kepecs A., Uchida N., Zariwala H.A. & Mainen Z.F., « Neural correlates, computation and behavioural impact of decision confidence », *Nature,* 455(7210), 2008, 227-231.

Kiani R. & Shadlen M.N., « Representation of confidence associated with a decision by neurons in the parietal cortex », *Science,* 324(5928), 2009, 759-764.

Kikyo H., Ohki K. & Miyashita Y., « Neural correlates for feeling-of-knowing: an fMRI parametric analysis », *Neuron,* 36(1), 2002, 177-186.

Koriat A., « How do we know that we know? The accessibility model of the feeling of knowing », *Psychol. Rev.,* 100(4), 1993, 609-639.

Kornell N., Son L.K. & Terrace H.S., « Transfer of metacognitive skills and hint seeking in monkeys », *Psychol. Sci.,* 18(1), 2007, 64-71.

Kovacs A.M., Teglas E. & Endress A.D., « The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults », *Science,* 330, 2010, 1830-1834.

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

Kunimoto C., Miller J. & Pashler H., « Confidence and Accuracy of Near-Threshold Discrimination Responses », *Conscious Cogn.,* 10, 2001, 294-340.

Lau H. & Rosenthal D., « Empirical support for higher-order theories of conscious awareness », *Trends Cogn. Sci.,* 15(8), 2011, 365-373.

Logan G.D. & Crump M.J., « Cognitive illusions of authorship reveal hierarchical error detection in skilled typists », *Science,* 330(6004), 2010, 683-686.

Lombardo M.V., Chakrabarti B., Bullmore E.T., Sadek S.A., Pasco G., Wheelwright S.J. *et al.* « Atypical neural self-representation in autism », *Brain,* 133(Pt 2), 2010, 611-624.

Ma W.J., Beck J.M., Latham P.E. & Pouget A., « Bayesian inference with probabilistic population codes », *Nat. Neurosci.,* 9(11), 2006, 1432-1438.

Marti S., Sackur J., Sigman M. & Dehaene S., « Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection », *Cognition,* 115(2), 2010, 303-313.

Meltzoff A.N. & Brooks R., « Self-experience as a mechanism for learning about others: a training study in social cognition », *Dev. Psychol.,* 44(5), 2008, 1257-1265.

Metcalfe J., « Premonitions of insight predict impending error », *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 12(4), 1986, 623-634.

Naccache L., Dehaene S., Cohen L., Habert M.O., Guichart-Gomez E., Galanaud D. *et al.* « Effortless control: executive attention and conscious feeling of mental effort are dissociable », *Neuropsychologia,* 43(9), 2005, 1318-1328.

Ochsner K.N., Knierim K., Ludlow D.H., Hanelin J., Ramachandran T., Glover G. *et al.* « Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other », *J. Cogn. Neurosci.,* 16(10), 2004, 1746-1772.

Persaud N., McLeod P. & Cowey A., « Post-decision wagering objectively measures awareness », *Nat. Neurosci.,* 10(2), 2007, 257-261.

Rolls E.T., Grabenhorst F. & Deco G., « Choice, difficulty, and confidence in the brain », *Neuroimage,* 53(2), 2010, 694-706.

Rounis E., Maniscalco B., Rothwell J.C., Passingham R. & Lau H. « Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness », *Cognitive Neuroscience*, 2010.

Semendeferi K., Armstrong E., Schleicher A., Zilles K. & Van Hoesen G.W. « Prefrontal cortex in humans and apes: a comparative study of area 10 », *Am. J. Phys. Anthropol.,* 114(3), 2001, 224-241.

Smith J.D., Schull J., Strote J., McGee K., Egnor R. & Erb L. « The uncertain response in the bottlenosed dolphin (Tursiops truncatus) », *J. Exp. Psychol. Gen.,* 124(4), 1995, 391-408.

Terrace H.S. & Son L.K., « Comparative metacognition », *Curr. Opin. Neurobiol.,* 19(1), 2009, 67-74.

Van Schie H.T., Mars R.B., Coles M.G. & Bekkering H., « Modulation of activity in medial frontal and motor cortices during error observation », *Nat. Neurosci.,* 7(5), 2004, 549-554.

Vogeley K., May M., Ritzl A., Falkai P., Zilles K. & Fink G.R., « Neural correlates of first-person perspective as one constituent of human self-consciousness », *J. Cogn. Neurosci.,* 16(5), 2004, 817-827.

Yokoyama O., Miura N., Watanabe J., Takemoto A., Uchida S., Sugiura M. *et al.,* « Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance », *Neurosci. Res.,* 68(3), 2010, 199-206.

SÉMINAIRE : PSYCHOLOGIE ET NEUROPSYCHOLOGIE DES FICTIONS MENTALES

En complément du cours, le séminaire faisait intervenir des personnalités, chercheurs ou médecins, dont les recherches éclairent la manière dont l'introspection peut s'écarter de la réalité. En effet, selon le terme proposé par Lionel Naccache, notre cerveau construit sans cesse des fictions conscientes. De même qu'un roman peut tantôt s'écarter radicalement de la réalité, et tantôt l'épouser étroitement, sans que le lecteur puisse faire la part des choses, de même nos récits autobiographiques à la première personne sont tantôt fidèles à notre vie psychologique réelle, et tantôt bien différents. La représentation du soi n'est-elle qu'une fiction ? D'où provient notre sentiment d'un point de vue personnel sur le monde extérieur ? Peut-il se déplacer ou se détériorer à la suite d'une lésion cérébrale ou lorsque le cerveau est soumis à des stimulations anormales ? Ces questions ont été abordées à travers les exposés de spécialistes de la neuropsychologie du soi et des hallucinations :

– Lionel Naccache (Hôpital de la Salpêtrière, Paris) : Neuropsychologie des interprétations et des croyances ;

– Olaf Blanke (École polytechnique fédérale de Lausanne) : How the brain computes the self's point of view ;

– Paul Fletcher (University of Cambridge, UK) : Misperceiving and misbelieving: towards an understanding of psychosis ;

– Gilles Fénelon (Hôpital Henri Mondor, Créteil) : Hallucinations, illusions et sensations de présence au cours de la maladie de Parkinson ;

– Henrik Ehrsson (Karolinska Institutet, Stockholm) : The construction of an experience of our own body ;

– Predrag Petrovic (Karolinska Institutet, Stockholm) : Expectations, beliefs, and the origins of the placebo effect.

Trois autres cours et séminaires ont été donnés à l'Institut d'études avancées Peter Wall de Vancouver (Canada), dans le cadre du programme d'échanges avec le Collège de France.

Enfin, en lien avec le thème du séminaire, le professeur **Brian Boyd**, *University Distinguished Professor* au département de langue anglaise de l'université d'Auckland, et spécialiste mondialement réputé de l'œuvre de Nabokov, a donné, sur invitation de l'Assemblée des Professeurs conjointement proposée par Stanislas Dehaene et Antoine Compagnon, deux conférences : *The evolution of stories* et *Nabokov as psychologist.*

ACTIVITÉS DE RECHERCHE DU LABORATOIRE

Nous mettons ici en valeur deux résultats récents qui nous paraissent importants. Vient ensuite une liste complète des publications du laboratoire.

Impact de l'apprentissage de la lecture sur le cerveau

Comme je l'ai souligné dans *Les neurones de la lecture*, l'écriture est une invention trop récente pour avoir influencé l'évolution génétique humaine. Son apprentissage ne peut donc reposer que sur un « recyclage » de régions cérébrales préexistantes, initialement dédiées à d'autres fonctions mais suffisamment plastiques

pour se réorienter vers l'identification des signes écrits et leur mise en liaison avec le langage parlé. Pour comprendre l'impact cérébral de ce processus d'alphabétisation, dans le cadre d'un programme ANR codirigé avec Laurent Cohen, et en collaboration avec des équipes brésiliennes (Lucia Braga, hôpital Sarah, Brasilia), portugaises (Paulo Ventura, université de Lisbonne) et belges (Régine Kolinsky et José Morais), nous avons initié un vaste programme d'étude de l'analphabétisme.

Dans notre article de *Science* (Dehaene *et al.*, 2010), nous avons mesuré, par IRM fonctionnelle, l'activité cérébrale d'adultes volontaires diversement alphabétisés en réponse à toute une batterie de stimuli : phrases parlées et écrites, mots et pseudo-mots parlés, visages, maisons, objets, damiers… 63 adultes ont participé à l'étude : 10 personnes analphabètes, 22 personnes non scolarisées dans l'enfance mais alphabétisées à l'âge adulte, et 31 personnes scolarisées depuis l'enfance. La recherche a été menée en parallèle au Portugal et au Brésil, avec des imageurs IRM à 3 Tesla de NeuroSpin (CEA Saclay) pour les volontaires portugais et au centre de recherches en neurosciences de l'hôpital Sarah Lago Norte à Brasilia pour les volontaires brésiliens. Au Brésil, voici quelques dizaines d'années, il était encore relativement fréquent que des enfants ne puissent pas aller à l'école uniquement en raison de leur environnement social. Tous les volontaires étaient bien intégrés socialement, en bonne santé, et la plupart avaient un emploi.

Les résultats apportent des éléments de réponse à plusieurs questions. Tout d'abord, comment les aires cérébrales impliquées dans la lecture se transforment-elles sous l'influence de l'éducation ? En comparant directement l'évolution de l'activation cérébrale en fonction du score de lecture (nul chez les analphabètes et variable dans les autres groupes), nous avons montré que l'impact de l'alphabétisation est bien plus étendu que les études précédentes ne le laissaient penser. Apprendre à lire augmente les réponses des aires visuelles du cortex dans une région spécialisée pour la forme écrite des lettres, l'aire de la forme visuelle des mots (*visual word form area* ; revue dans Dehaene *et al.*, 2011). Cependant, à notre surprise, même l'aire visuelle primaire voit également son activation augmenter. Ce résultat a été confirmé par une autre étude d'IRM fonctionnelle menée au laboratoire par Marcin Szwed (Szwed *et al.*, 2011a). Il est également soutenu par une étude comportementale : l'intégration des contours, une tâche dépendante des aires visuelles précoces et notamment des connexions horizontales de l'aire visuelle primaire, est sélectivement diminuée chez les analphabètes par rapport aux adultes de même âge et de même niveau socio-économique qui ont appris à lire dans l'enfance ou à l'âge adulte (Szwed *et al.*, 2011b).

L'apprentissage de la lecture augmente également les réponses au langage parlé dans une région du cortex auditif, le *planum temporale*, impliquée dans le codage des phonèmes (les plus petits éléments significatifs du langage parlé, comme « b » ou « ch »). Ce résultat pourrait correspondre au fait que les analphabètes ne parviennent pas à réaliser des jeux de langage tels que la délétion du premier son d'un mot (Paris→aris). Enfin, la lecture induit également une extension des aires du langage et une communication bidirectionnelle entre les réseaux du langage parlé et écrit : chez un bon lecteur, voir une phrase écrite active l'ensemble des aires du langage parlé, tandis qu'entendre un mot parlé permet de réactiver rapidement, si nécessaire, son code orthographique dans les aires visuelles.

En second lieu, à quoi servent les aires cérébrales impliquées dans la lecture avant qu'une personne n'apprenne à lire ? Selon le modèle du recyclage neuronal, l'apprentissage de la lecture n'implique peut-être pas toujours un gain de fonction,

mais l'augmentation des réponses aux mots pourrait s'accompagner de diminutions des réponses à d'autres catégories de connaissances. Effectivement, chez les analphabètes, nos résultats montrent que l'aire visuelle de l'hémisphère gauche qui, chez les lecteurs, décode les mots écrits, répond à une fonction proche : la reconnaissance visuelle des objets et des visages. Dans cette région, au cours de l'apprentissage, la réponse aux visages diminue légèrement à mesure que la compétence de lecture augmente, et l'activation aux visages se déplace partiellement dans l'hémisphère droit. Le cortex visuel se réorganise donc, en partie, du fait de la compétition entre l'activité nouvelle de lecture et les activités plus anciennes de reconnaissance des visages et des objets. Ces travaux sont aujourd'hui en cours de réplication chez l'enfant de 6 à 9 ans, au moment même de l'apprentissage de la lecture (équipe de G. Dehaene-Lambertz). Chez l'enfant de quatre ans, en collaboration avec Jessica Cantlon, nous sommes également parvenus à confirmer l'existence d'une compétition entre l'apprentissage des lettres et des visages : l'amélioration des scores de reconnaissance des lettres s'accompagne d'une diminution systématique de l'activité évoquée par les visages dans la région fusiforme (Cantlon *et al*., 2011). Nous ne savons pas si cette compétition corticale entraîne des conséquences fonctionnelles pour la reconnaissance ou la mémoire des visages, mais des recherches comportementales en cours, menées par Paulo Ventura chez les adultes portugais, suggèrent que le traitement des visages devient moins « holistique » chez les personnes qui ont appris à lire.

Enfin, dernière question posée par notre étude d'IRM : les modifications cérébrales liées à l'alphabétisation peuvent-elles se produire à l'âge adulte ? Ou bien existe-t-il une période sensible pour cet apprentissage dans la petite enfance ? Dans notre étude, la très grande majorité des effets de l'apprentissage de la lecture sur le cortex sont visibles autant chez les personnes scolarisées dans l'enfance que chez celles qui ont suivi des cours d'alphabétisation à l'âge adulte. Bien entendu, ces dernières n'atteignent que rarement les mêmes performances de lecture, mais cette différence pourrait n'être due qu'à leur moindre entraînement. À performances de lecture égales, il n'existe pratiquement pas de différences mesurables entre les activations cérébrales des personnes qui ont appris à lire dans l'enfance ou à l'âge adulte. Les circuits de la lecture restent donc plastiques tout au long de la vie.

Ces résultats soulignent l'impact massif de l'éducation sur le cerveau humain. Ils nous rappellent également que l'immense majorité des expériences d'IRM cérébrale portent sur le cerveau éduqué et que l'organisation cérébrale en l'absence d'éducation constitue un immense territoire largement inexploré.

Intuitions géométriques

En collaboration avec Véronique Izard et Pierre Pica (CNRS, Paris) ainsi qu'Elizabeth Spelke (Harvard), nous avons poursuivi notre étude de l'intuition mathématique spontanée (Izard *et al.*, 2011). Les concepts de la géométrie euclidienne (points, droites, plans, angles…) se développent-ils intuitivement chez tous les êtres humains, ou bien sont-ils des inventions propres à la culture occidentale et qui nécessitent un long apprentissage ? Des peuples qui ne disposent pas d'une formation en mathématiques sont-ils capables de faire preuve d'intuitions géométriques spontanées ? Pour répondre à ces questions, nous avons testé des Indiens Mundurucus d'Amazonie, non scolarisés, vivant dans un territoire isolé, et

dont le langage ne possède que peu de concepts géométriques. Leur compréhension intuitive des concepts fondamentaux de la géométrie a été comparée à celles de populations ayant étudié la géométrie à l'école. Nous avons élaboré deux tests cognitifs nouveaux. Le premier consiste à répondre à des questions sur les propriétés abstraites des droites, en particulier leur caractère infini et leurs propriétés de parallélisme. Dans le second, il s'agit de compléter un triangle en calculant la position de son sommet ainsi que l'angle au niveau de ce sommet.

Afin d'introduire la géométrie auprès des Mundurucus en quelques minutes, nous leur avons montré, sur l'écran d'un ordinateur, deux mondes, l'un plat et le second arrondi en forme de sphère, sur lesquels se trouvaient des villages (correspondants aux « points » en géométrie euclidienne) et des chemins strictement rectilignes (les « droites »). Les mêmes tests ont été effectués chez une trentaine d'adultes et d'enfants originaires de France et des États-Unis, qui, contrairement aux Mundurucus, avaient étudié la géométrie à l'école.

Les résultats indiquent que tous les êtres humains disposent d'intuitions universelles en géométrie élémentaire, quelle que soit leur culture ou leur niveau d'éducation. Les Indiens Mundurucus se sont montrés tout à fait capables de résoudre des problèmes simples de géométrie plane et de modifier leurs réponses lorsque les mêmes questions étaient posées en géométrie sphérique. Au premier test, ils ont répondu convenablement à la plupart des questions, y compris celles qui dépassaient la simple perception et portaient par exemple sur le comportement des droites parallèles à l'infini. Leurs réponses au second test, celui du triangle, ont mis en évidence l'intuition d'une propriété essentielle en géométrie plane, à savoir le fait que la somme des angles d'un triangle est constante (égale à 180°). En géométrie sphérique, les indiens d'Amazonie ont même présenté des réponses plus précises que les sujets français ou nord-américains qui avaient sans doute, au cours de leur scolarité, acquis une plus grande familiarité avec la géométrie euclidienne plane.

Ces résultats confortent nos recherches antérieures sur l'intuition arithmétique et géométrique des indiens Mundurucus. En accord avec les idées développées par Platon dans le *Ménon*, ils indiquent que des intuitions proto-mathématiques très sophistiquées peuvent être mises en évidence chez tous les êtres humains, à condition d'introduire les concepts abstraits des mathématiques sous la forme d'exemples concrets et tangibles. Cette conclusion n'est évidemment pas dépourvue de conséquences pour l'éducation aux mathématiques.

PUBLICATIONS (2010-2011)

Articles originaux

Forget J., Buiatti M., Dehaene S., « Temporal Integration in Visual Word Recognition », *J. Cogn. Neurosci.*, 22(5), 2010, 1054-68.

Dehaene S., Nakamura K., Jobert A., Kuroki C., Ogawa S., Cohen L., « Why do children make mirror errors in reading? Neural correlates of mirror invariance in the visual word form area », *NeuroImage*, 49, 2010, 1837-48.

Qiao E., Vinckier F., Szwed M., Naccache L., Valabrègue R., Dehaene S., Cohen L., « Unconsciously deciphering handwriting: Subliminal invariance for handwritten words in the visual word form area », *NeuroImage*, 49(2), 2010, 1786-99.

Kouider S., de Gardelle V., Dehaene S., Dupoux E., Pallier C., « Cerebral bases of subliminal speech priming », *Neuroimage*, 49(1), 2010, 922-929.

Marti S., Sackur J., Sigman M., Dehaene S., « Mapping the introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection », *Cognition*, 115, 2010, 303-13.

Rasanen P., Salminen J., Wilson A., Aunio P., Dehaene S., « Computer-assisted intervention for children with low numeracy skills », *Cognitive Development*, 24, 2009, 450-72.

de Lange F., Jensen O., Dehaene S., « Accumulation of Evidence during Sequential Decision Making: the Importance of Top-down Factors », *Journal of Neuroscience,* 30, 2010, 731-738.

Rusconi E., Pinel P., Dehaene S., Kleinschmidt A., « The enigma of Gerstmann's syndrome revisited: a telling tale of the vicissitudes of neuropsychology », *Brain*, 133(Pt 2), 2010, 320-32.

Berteletti I., Lucangeli D., Piazza M., Dehaene S., Zorzi M., « Numerical Estimation in Preschoolers », *Developmental Psychology*, 46, 2010, 545-551.

Viarouge A., Hubbard E.M., Dehaene S., Sackur J., « Number Line Compression and the Illusory Perception of Random Numbers », *Exp. Psychol.,* 57, 2010, 446-54.

Pegado F., Bekinschtein T., Chausson N., Dehaene S., Cohen L., Naccache L., « Probing the lifetimes of auditory novelty detection processes », *Neuropsychologia*, 48, 2010, 3145-54.

Dehaene S., Pegado F., Braga L.W., Ventura P., Filho G.N., Jobert A., Dehaene-Lambertz G., Kolinsky R., Morais J., Cohen L., « How learning to read changes the cortical networks for vision and language », *Science*, 6009, 2010, 1359-1364.

Piazza M., Facoetti A., Trussardi A.N., Berteletti I., Conte S., Lucangeli D., Dehaene S., Zorzi M., « Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia », *Cognition*, 116, 2010, 33-41.

Pegado F., Nakamura K., Cohen L., Dehaene S., « Breaking the Symmetry: Mirror discrimination for single letters but not for pictures in the Visual Word Form Area », *Neuroimage*, 55, 2011, 742-9.

Szwed M., Dehaene S., Kleinschmidt A., Eger E., Valabrègue R., Amadon A., Cohen L., « Specialization for written words over objects in the visual cortex », *Neuroimage*, 56, 2011, 330-44.

Kamienkowski J.E., Pashler H., Dehaene S., Sigman M., « Effects of practice on task architecture: Combined evidence from interference experiments and random-walk models of decision making », *Cognition*, 119(1), février 2011, 81-95.

Hesselmann G., Flandin G., Dehaene S., « Probing the cortical network underlying the psychological refractory period: A combined EEG-fMRI study », *Neuroimage*, 56(3), 2011, 1608-21

Cantlon J.F., Pinel P., Dehaene S., Pelphrey K.A., « Cortical Representations of Symbols, Objects, and Faces are Pruned Back during Early Childhood », *Cerebral cortex,* 21, 2011, 191-9.

Van Opstal F., de Lange F.P., Dehaene S., « Rapid parallel semantic processing of numbers without awareness », *Cognition,* 120, 2011, 136-47.

Dehaene S., Cohen L., « The unique role of the visual word form area in reading », *Trends in Cognitive Science,* 15(6), 2011, 254-62.

Faugeras F., Rohaut B., Weiss N., Bekinschtein T.A., Galanaud D., Puybasset L., Bolgert F., Sergent C., Cohen L., Dehaene S., Naccache L., « Probing consciousness with event-related potentials in patients who meet clinical criteria for vegetative state », *Neurology*, 77(3), 2011, 264-8.

Vinckier F., Qiao E., Pallier C., Dehaene S., Cohen L., « The impact of letter spacing on reading: a test of the bigram coding hypothesis », *Journal of Vision*, 11(6), 8, 2011, 1-21.

Izard V., Pica P., Spelke E., Dehaene S., « Flexible intuitions of Euclidean geometry in an Amazonian indigene group », *PNAS*, 108, mai 2011, 9782-87.

Markram H., Meier K., Lippert T., Grillner S., Frackowiak R., Dehaene S., Knoll A., Sompolinsky H., Verstreken K., DeFelipe J., Grant S., Changeux J.P., Sariam A., « Introducing the Human Brain Project », *Procedia Computer Science*, 2011.

Livres

Dehaene S., *La Bosse des Maths, quinze ans après*, Paris, Odile Jacob, 2010.

Dehaene S., *The number sense* (2nd edition), New York, Oxford University Press, 2011.

Dehaene S. & Brannon E., *Space, Time and Number in the brain: Searching for the foundations of mathematical thought*, Londres, Elsevier, 2011.

Chapitres de livres

Cohen L., Vinckier F., Dehaene S., « Anatomical and Functional Correlates of Acquired Peripheral Dyslexias », *in* Cornelissen P., Hansen P., Kringelbach M., Pugh K. (éd.), *The neural basis of reading*, Oxford Scholarship Online Monographs, 2010, 223-64.

Sigman M., Dehaene S., « Why does it take time to make a decision? The role of a global workspace in simple decision making », *in* Vartanian O. & Mandel D.R. (éd.), *Neurosciences of Decision Making*, Psychology Press, 2011, 11-44.

Dehaene S., « Reading as Neuronal Recycling », *in* McCardle P., Miller B., Lee J.R., Tzeng O.J.L. (éd.), *Dyslexia Across Languages*, Paul H. Brookes Publishing, 2011, 102-16.

Revues, commentaires, diffusion des connaissances

Dehaene S., Brannon E.M., « Space, time, and number: a Kantian research program », *Trends in Cognitive Science*, 14, 2010, 517-9.

Principales conférences invitees

– « A taxonomy of conscious and nonconscious visual states and its neural correlate ». Nobel Symposium *The Enlightened Brain: Evolution and Development of the Human Brain* à l'occasion du 200e anniversaire de l'Institut Carolin, Stockholm, Suède, 8 juin 2010.

– « How learning to read changes human brain networks ». Gordon conference on Cognitive Neuroscience, Waterville, États-Unis, 2-6 août 2010.

– « Cognitive Imaging: The Neural Bases of Human Brain Function ». Academia Europae, Leuven, 11 septembre 2010.

– « The brain mechanisms underlying mathematical operations ». University of Jyväskylä, Finlande, 1er octobre 2010.

– « The massive impact of literacy on the human brain ». Académie pontificale des sciences, 27-29 octobre 2010.

– « The visual word form area », *conference and public debate with Cathy Price*. Society for the Neurobiology of Language, San Diego, 12 novembre 2011.

– « Les neurones de la lecture ». Collège de Belgique, 18 janvier 2011.

– « Quantity coding and computation in the animal and human brain ». Conference COSYNE, Salt Lake City, 27 février 2011.

– « Physiological signatures of conscious access and the workspace model ». Conference COSYNE, Snowbird, États-Unis, 28 février 2011.

– « Brain mechanisms of numeracy and consequences for education ». The Latin American school for education, cognitive and neural sciences (James S. Mcdonnell Foundation and Universidad de Chile), San Pedro de Atacama, 14 mars 2011.

– « How Math Comes to Mind: Intuition, Visualization, and Teaching », discussion avec le mathématicien Stephen Strogatz. Princeton, 27 avril 2011.

– « Limits of Subliminal Processing and Signatures of Conscious Access ». Association for Psychological Science, Washington, 27 mai 2011.

– « How learning to read changes the visual system », Rank prize lecture. European Conference on Visual Perception (ECVP), Toulouse, 28 août 2011.

THÈSES SOUTENUES DANS LE LABORATOIRE

– Antoine Del Cul (21/06/2010 ; directeur : Stanislas Dehaene) : « L'accès à la conscience et ses perturbations dans la schizophrénie. Étude des mécanismes cognitifs et cérébraux du masquage visuel rétrograde ».

– Sepideh Sadaghiani (27/07/2010 ; directeur: Andreas Kleinschmidt) : « The impact of ongoing brain activity on the variability of human brain function and behavior ».

– François Leroy (15/09/2011 ; directeur : Ghislaine Dehaene-Lambertz) : « Étude méthodologique et structurale du développement cérébral en IRM : application aux aires du langage dans une population de nourrissons ».

PARTICIPATION AUX PROGRAMMES NATIONAUX ET INTERNATIONAUX

Responsabilité du pilier « neurosciences et cognition humaines » du *Human Brain Project,* dossier préparé en réponse à l'appel d'offre *Future Emerging Technologies* de la Communauté européenne.

HONNEURS ET DISTINCTIONS

Stanislas Dehaene a été nommé chevalier dans l'ordre de la Légion d'honneur, docteur *honoris causa* de l'université de Lisbonne, et professeur honoraire de l'*East China Normal University* de Shanghai.