
# LES SIGNATURES NEUROBIOLOGIQUES DE LA CONSCIENCE

## NEUROBIOLOGIE FONCTIONNELLE, PHÉNOMÈNES DE CONSCIENCE, COGNITION, AUTOMATES « INTELLIGENTS », ÉTHIQUE

Gilbert Belaubre
Eric Chenin
Victor Mastrangelo
Pierre Nabet
Alberto Oliverio
Jacques Printz
Jean Schmets
Jean-Pierre Treuil

edp sciences

# LES SIGNATURES NEUROBIOLOGIQUES DE LA CONSCIENCE

## NEUROBIOLOGIE FONCTIONNELLE, PHÉNOMÈNES DE CONSCIENCE, COGNITION, AUTOMATES « INTELLIGENTS », ÉTHIQUE

**Alberto OLIVERIO** (Laboratoire de Psychobiologie, Université La Sapienza, Rome) ; **Francis EUSTACHE** (Inserm-EPHE-Université de Caen U1077, Neuropsychologie et imagerie de la mémoire humaine, Université de Caen) ; **Armelle VIARD** (Inserm-EPHE-Université de Caen U1077, Neuropsychologie et imagerie de la mémoire humaine, Université de Caen) ; **Claire SERGENT** (Laboratoire Psychologie de la Perception UMR 8242, Université Paris Descartes) ; **Jérôme SACKUR** (Laboratoire des sciences cognitives et psycholinguistique (ENS/CNRS/E-HESS) ; **Laure ZAGO** (Groupe d'Imagerie Neurofonctionnelle, Institut des Maladies Neurodégénératives, UMR 5293, Université de Bordeaux) ; **Marie AMALRIC** (Research group The Concepts, Actions, and Objects (CAOs) Lab / The Kid Neuro Lab. Carnegie Mellon University Pittsburgh, USA) ; **Jean-Gabriel GANASCIA** (Sorbonne Université, ACASA - Agents Cognitifs et Apprentissage Symbolique Automatique) ; **Antoine BORDES** (Laboratoire Facebook Artificial Intelligence Research, Paris) ; **Luc STEELS** (ICREA, Institut de Biologia Evolutiva (UPF/CSIC), Université de Barcelone ; Université Libre de Bruxelles - VUB) ; **Jean-Paul HATON** (Université de Lorraine, Reconnaissance des formes en IA, LORIA/INRIA/NANCY) ; **Ernesto DI MAURO** (Dipartimento di Scienze Ecologiche e Biologiche, Università della Tuscia, Viterbo, Italie) ; **Franck COSSON** (Docteur en philosophie, Université de Lorraine) ; **Gérard DE BOISBOISSEL** (Centre de recherche des Écoles Saint-Cyr Coetquidan) ; **Laurence DEVILLERS** (Sorbonne Université, GEMASS/LIMSI/CNRS) ; **Raja CHATILA** (Sorbonne Université, ISIR/CNRS Institut des Systèmes Intelligents et de Robotique).

Dans le cadre de ses activités scientifiques, l'Académie Européenne Interdisciplinaire des Sciences a réuni durant la période 2016-2018, par l'intermédiaire de séminaires, conférences et colloque, divers spécialistes travaillant dans les domaines des neurosciences, de la psychologie cognitive, de l'intelligence artificielle, ou encore réfléchissant aux impacts sociétaux des avancées obtenues.

Cet ouvrage est le fruit de toutes ces contributions et a pour ambition de présenter un certain nombre de résultats, de perspectives actuellement discernables, de points de vue concernant l'état des connaissances dans ces domaines avec un lien, parfois direct, parfois implicite, avec la question de la conscience.

Une introduction générale, rédigée par le comité de lecture de l'AEIS, rappelle le contexte de l'émergence de ces résultats, perspectives et points de vue, dont le lecteur prendra connaissance dans le corps de l'ouvrage, qui comporte quatre parties :

1. Travaux en neurosciences et psychologie expérimentale.
2. Sciences cognitives et intelligence artificielle.
3. Réflexions sur l'intelligence, la conscience et l'impact de l'IA sur les activités humaines.
4. Synthèse des discussions de la table ronde tenue à l'issue du colloque de mars 2018.

Un court épilogue rédigé par le comité de lecture met en avant des réflexions et questions qu'ont soulevé la lecture des différents chapitres de l'ouvrage et la prise de connaissance d'articles scientifiques foisonnants sur les travaux actuels pluridisciplinaires autour des interrogations liées à la conscience au sens large.

ISBN : 978-2-7598-2544-8

9 782759 825448

edp sciences


# LES SIGNATURES NEUROBIOLOGIQUES DE LA CONSCIENCE

## NEUROBIOLOGIE FONCTIONNELLE, PHÈNOMENES DE CONSCIENCE, COGNITION, AUTOMATES « INTELLIGENTS », ÉTHIQUE

Gilbert Belaubre
Eric Chenin
Victor Mastrangelo
Pierre Nabet
Alberto Oliverio
Jacques Printz
Jean Schmets
Jean-Pierre Treuil


# LES SIGNATURES NEUROBIOLOGIQUES DE LA CONSCIENCE

NEUROBIOLOGIE FONCTIONNELLE,
PHÈNOMENES DE CONSCIENCE,
COGNITION, AUTOMATES « INTELLIGENTS », ÉTHIQUE

Académie Européenne Interdisciplinaire des Sciences

ISBN (papier) : 978-2-7598-2544-8

ISBN (ebook) : 978-2-7598-2612-4

# La collection de l'AEIS

Les travaux de l'**Académie Européenne Interdisciplinaire des Sciences** portent depuis plus de dix ans sur les questions majeures auxquelles est confrontée la recherche à la croisée de plusieurs disciplines. Et la présente collection a pour but de faire le point sur ces questions.

Nos ouvrages sont issus de séminaires mensuels et de congrès bisannuels pluridisciplinaires auxquels sont associés de nombreux chercheurs extérieurs. Notre ambition est de faciliter les échanges entre programmes de recherche spécialisés. Elle est aussi d'informer un public plus large sur les avancées récentes.

Pour chacun des ouvrages et en concertation avec les auteurs, un comité de lecture a pour tâche de coordonner l'ensemble des contributions conformément aux objectifs poursuivis, et d'en faire ressortir les lignes de force.

Les ouvrages de l'Académie recourront à deux formes de publication : le livre-papier
avec diffusion dans les bibliothèques des universités et des centres de recherche
et en librairie, et/ou la version électronique en ligne dans la section e-Books du site de l'éditeur *EDPSciences*.

Notre premier ouvrage de la collection : « *Formation des systèmes stellaires et planétaires-Conditions d'apparition de la vie »* est disponible en téléchargement gratuit au format PDF à l'adresse suivante

https://www.edp-open.org/books/edp-open-books/312-formation-des-systemes-stellaires-et-planetaires

Notre deuxième ouvrage : « *Ondes, Matière et Univers »* est disponible en téléchargement gratuit au format PDF à l'adresse ci-après.

https://www.edp-open.org/images/stories/books/fulldl/livreAEIS_ebook.pdf

# TABLE DES MATIERES

**Introduction générale** ..... 7

**PREMIÈRE PARTIE**

**Travaux en neurosciences et psychologie expérimentale**

**Présentation** ..... 27

**Chapitre 1.** Rôle essentiel de la mémoire dans la formation de toute représentation ..... 31

**Chapitre 2.** La mémoire humaine et ses substrats cérébraux..... 43

**Chapitre 3.** Les bases neurobiologiques de la conscience ..... 65

**Chapitre 4.** Quelles données subjectives pour l'étude du flux de conscience ? ..... 85

**Chapitre 5.** Bases cérébrales de la spécialisation hémisphérique de l'attention visuo-spatiale et des relations complémentaires entre l'attention spatiale et le langage ..... 107

**Chapitre 6.** Représentation et manipulation des concepts mathématiques par le cerveau humain ..... 117

**DEUXIÈME PARTIE**

**Sciences cognitives et Intelligence Artificielle**

**Présentation** ..... 143

**Chapitre 7.** Intelligence Artificielle : des Big-Data au Cerveau ..... 147

**Chapitre 8.** Former les machines à la compréhension du langage naturel ..... 163

**Chapitre 9.** L'origine et l'évolution du langage ..... 183

**Chapitre 10.** Les machines pensantes : un panorama de l'intelligence artificielle..... 209

## TROISIÈME PARTIE

## Intelligence, Conscience et impact de l'IA sur les activités humaines

**Présentation** ..... 235

**Chapitre 11.** A la recherche de définitions ..... 239

**Chapitre 12.** Recherches sur l'apparition de niveaux de conscience chez l'animal..... 245

**Chapitre 13.** Conscience artificielle et monde militaire ..... 277

**Chapitre 14.** Relation émotionnelle entre humains et robots : Quelle éthique ? ..... 289

## QUATRIÈME PARTIE

## Table ronde

**1** : Peut-on envisager une Conscience Artificielle ? ..... 303

**2** : Problèmes éthiques posés par les décisions prises par les machines . 307

**3** : Interactions et Interdépendances hommes-machines ..... 310

**4** : Intelligence Artificielle et Insertion dans les réseaux sociaux ..... 313

**Épilogue** : Pour conclure ..... 317

**Remerciements** ..... 327

**Présentation de l'Académie Européenne Interdisciplinaire des Sciences** ..... 329

# INTRODUCTION GENERALE

Dans le cadre de ses activités scientifiques, l'Académie Européenne Interdisciplinaire des Sciences a réuni par l'intermédiaire de séminaires, conférences et colloque divers spécialistes travaillant dans les domaines des neurosciences, de la psychologie cognitive, de l'Intelligence Artificielle, ou encore réfléchissant aux impacts sociétaux des avancées obtenues. Ce livre est le fruit de toutes ces contributions et a pour ambition de présenter un certain nombre de résultats, de perspectives actuellement discernables, de points de vue concernant l'état des connaissances dans ces domaines avec un lien, parfois direct, parfois implicite, avec la question de la conscience.

Cette introduction générale a pour but de rappeler le contexte de l'émergence de ces résultats, perspectives et points de vue, dont le lecteur prendra connaissance dans le corps de l'ouvrage. La question du fonctionnement de l'esprit humain et de ses rapports avec le cerveau et le corps ne date pas d'hier en effet et de nombreuses réflexions philosophiques ont été proposées au fil des siècles.

En 1944, le physicien et prix Nobel de Physique Erwin Schrödinger publia un ouvrage qui marqua son époque, intitulé « Qu'est-ce que la vie ? ». Il y abordait un certain nombre de questions auxquelles, selon lui, la science devait répondre pour expliquer la nature physique et chimique des êtres vivants. Par exemple, il indiquait que la matière vivante se soustrait à l'augmentation du désordre qui conduit à l'équilibre thermodynamique en se nourrissant de néguentropie. Il voulait ainsi expliquer la présence de « l'ordre » à l'intérieur des êtres vivants et leur tendance à s'opposer au chaos et à la désorganisation qui régit les systèmes physiques. La compréhension du fonctionnement du cerveau était aussi à l'époque l'un des grands problèmes de la science. Schrödinger avança quelques réflexions en rapport avec la manière dont il faut comprendre d'un point de vue mécaniste[1], la subjectivité ou la perception d'une conscience unitaire. Francis Crick (Prix Nobel de physiologie ou médecine en 1962), codécouvreur de la structure de la molécule d'ADN, a salué en ce livre «une description théorique précoce du fonctionnement du stockage de l'information génétique, qui avait été pour lui une source d'inspiration dans ses premières recherches ».

[1]Le mécanisme est une philosophie de la nature selon laquelle l'Univers et tout phénomène qui s'y produit peuvent et doivent s'expliquer d'après les lois des mouvements matériels. Les explications mécanistes ramènent le fonctionnement du vivant à l'ensemble des interactions variées entre les molécules ou les ensembles de molécules qui en font partie.

**Une brève incursion en Philosophie de l'Esprit**

La conscience humaine est très difficile à définir pour diverses raisons. Elle pose en particulier à la science un problème différent dans sa nature de l'explication de phénomènes physiques tels que la dynamique des planètes, les phénomènes atomiques et nucléaires ou bien la photosynthèse dans le monde végétal. Cette différence, on l'a caractérisée de différentes façons. La conscience est seulement accessible et intrinsèque au sujet conscient, alors qu'une expérience scientifique est accessible à tous les participants. Elle a un caractère ineffable, c'est-à-dire qu'on ne peut en rendre compte convenablement dans les termes du langage, contrairement aux phénomènes physiques qui peuvent être exprimés avec précision à partir d'unités de mesure et constantes fondamentales (masse, température, G, $k_B$).

Depuis l'Antiquité, les philosophes réfléchissent à concilier l'esprit et son substrat biologique, le cerveau. Parmi toutes les approches philosophiques proposées pour tenter de résoudre le problème difficile de la conscience humaine, le dualisme et le monisme sont les deux principales écoles de pensée qui s'y sont essayées. Ces deux écoles réunies recueillent l'adhésion d'un grand nombre de philosophes. Ces derniers ont toutefois été amenés à nuancer ces deux positions théoriques générales afin de répondre aux critiques qui leur ont été formulées.

Pour les tenants du dualisme, deux grands courants philosophiques se sont développés. D'abord, celui du dualisme de substance pour lequel le monde physique existe bel et bien mais les aspects subjectifs de la conscience sont de nature distincte et constituent l'autre grande substance dont est fait le monde. Ce courant, dont les tenants les plus célèbres dans l'antiquité ont été Platon puis Aristote, fut formulé plus précisément par René Descartes au XVIIème siècle. Dans son *Discours de la méthode*, publié en 1637, René Descartes s'interrogeait sur ce dont il ne pouvait pas douter. Il arriva à la conclusion qu'il pouvait douter de tout, sauf de son expérience subjective. D'où la célèbre formule *« Je pense, donc je suis »*. Cela soulève immédiatement la question de l'interaction entre ce monde subjectif et le monde physique. Une question éminemment difficile à laquelle Descartes essaya de répondre sans succès – comme l'on sait.

Différentes variantes du dualisme ont alors été élaborées pour conserver l'intérêt des deux entités distinctes, tout en évitant certains de leurs écueils. Parmi celles-ci, on citera le dualisme des propriétés, où l'on admet que l'être humain n'est constitué que de matière, mais celle-ci possèderait deux types bien distincts de propriétés. Cette approche fut en premier présentée par Gottfried Wilhelm Leibniz au XVIIe siècle. Une version « moderne » a été présentée par le philosophe David Chalmers de l'Université de New-York, en 1995, dans un article intitulé « Facing Up to the Problem of Consciousness » paru dans le Journal of Consciousness Studies. Il proposait de distinguer les difficultés que pose l'étude de la conscience en deux types de problèmes distincts : les « problèmes faciles » et le « problème difficile » de la conscience. Les problèmes dit « faciles » concernent certaines manifestations de la conscience qu'il parait possible de décrire par les méthodes classiques d'observations et d'expérimentation scientifique. Ainsi en est-il par exemple de

la douleur dont on peut associer l'origine à une lésion par le corps et suivre le cheminement de ses signaux à travers l'architecture neuronale. Le même genre d'investigation est possible pour décortiquer la mécanique de tous les processus qui rendent possible la conscience (la vision, la mémoire, l'attention, les émotions, etc.). Selon David Chalmers, on peut trouver des explications fonctionnelles adéquates à ces processus.

Le problème difficile, pour sa part, découle des découvertes de la physique durant la première moitié du XXe siècle, découvertes qui ont rendu difficile de trouver à la conscience une place dans un monde matériel devenu plus intelligible. Ce monde fut compris dans sa globalité, par le biais des théories de la Physique, comme la relation entre ondes, forces, atomes et molécules ; il restait bien peu de place pour l'aspect subjectif de la conscience. C'est cet aspect qui constitue le cœur du problème difficile pour David Chalmers et d'autres de cette école. Pour eux, toutes les explications sur les causes et conséquences de nos états mentaux et leur identification dans un système nerveux donné (le problème facile) ne nous renseigneront pas sur la dimension subjective de la conscience.

Pour les dualistes de propriétés, mais aussi pour tous ceux qui accordent une réalité aux états mentaux conscients, le problème réputé « difficile » est donc celui qui se heurte à une double question : de quoi les expériences de conscience phénoménale sont-elles faites (ou comment les caractériser) ? pourquoi les individus sont-ils le sujet de telles expériences ?

Pour l'autre grande option philosophique qu'est le matérialisme, déjà défendue dans l'antiquité, dans la philosophie occidentale, par Parménide, Démocrite, Épicure ou Lucrèce..., la relation entre nos états mentaux et nos comportements ne pose pas de problème puisque les deux font partie du monde physique. Elle fut adoptée au XVIIème siècle par Baruch Spinoza ; plus proche de nous, on peut citer au XIXe siècle le physicien Ernst Mach, ainsi que Williams James et Bertrand Russel au XXe siècle… Une expérience subjective comme la douleur est bien réelle, mais elle correspond tout simplement aux états neuronaux qui la font naître. Il existe en philosophie de l'esprit, un courant moderne qu'on désigne par *matérialisme éliminativiste*, comportant à son tour de multiples variantes. L'idée, sous sa forme radicale[2], est que les concepts utilisés par la psychologie, et notamment le concept de la conscience, n'ont pas de réalité. Les progrès des neurosciences élimineront peu à peu les théories anciennes au profit de nouvelles plus pertinentes. Des philosophes et chercheurs « matérialistes éliminativistes », tels que Paul et Patricia Churchland, Université de Californie à San Diego, pensent que toutes les questions sur la conscience pourront se ramener à ce que David Chalmers désigne sous la catégorie des problèmes «faciles» et devraient être normalement résolus. Il est trop aisé, selon eux, de conclure qu'un phénomène comme la conscience est inexplicable par le seul fait qu'une discipline telle que la psychologie n'est pas en mesure aujourd'hui de l'appréhender.

[2] comme celle prônée par Daniel Dennett. *La Conscience expliquée*, Odile Jacob, 1993.

Mentionnons une troisième école qui est soutenue par certains philosophes dont David Chalmers qui promeuvent l'idée de l'universalité de la conscience, idée qui relève d'une théorie assez ancienne en philosophie appelée le Panpsychisme. Elle a été défendue au fil des siècles notamment par Thalès, Baruch Spinoza et William James, l'un de ses porte-paroles contemporain étant le philosophe Galen Strawson des Universités de Reading (Angleterre) et du Texas à Austin. Elle considère que toute matière a une nature psychique.

Mentionnons également les travaux développés, depuis la seconde moitié du XXème siècle dans le cadre du courant de la philosophie analytique, autour du rapport entre *conscience et représentation.* Ces travaux ont donné lieu à diverses « théories représentationnelles de l'esprit ». Parmi ses promoteurs on trouve Thomas Nagel, Université de New-York, Franck Cameron Jackson, Université nationale australienne, David Rosenthal, Université de la ville de New York (CUNY), Uriah Kriegel, Institut Jean Nicod (ENS)… Ils sont en partie basés sur la notion d'état mental dit « intentionnel » développée dans « Psychologie d'un point de vue empirique (1874) » par Franz Brentano. Ce dernier philosophe affirme que l'intentionnalité constitue le critère pertinent pour une distinction générale des phénomènes mentaux et des phénomènes physiques. Dans ce courant, deux grandes orientations théoriques sont introduites, regroupées sous les termes de théories du premier ordre (de la conscience) et théories d'ordre supérieur. Schématiquement, les premières soutiennent « qu'être conscient équivaut à avoir une représentation de premier ordre, une représentation du monde comme étant tel ou tel ». Les secondes affirment que cette représentation de premier ordre ne constitue pas à elle seule un état mental conscient. Il faut que cette représentation soit elle-même représentée, sur un second niveau. La conscience mobiliserait ainsi la représentation d'une représentation. De telles réflexions philosophiques ne sont pas sans échos dans les théories empiriques de la conscience apportées par les neurosciences cognitives.

Pour en terminer sur ces thèmes, constatons qu'il existe présentement différentes catégorisations plutôt convergentes de nos différentes formes de conscience. Pour le philosophe Ned Block, Université de New York, les phénomènes conscients comporteraient au moins quatre aspects centraux se manifestant en état d'éveil. *La conscience d'accès*, où un état est conscient si, lorsque l'on est dans cet état, une représentation de son contenu est immédiatement disponible. Cette représentation peut alors servir de prémisses pour le raisonnement et peut jouer un rôle dans le contrôle rationnel de l'action et de la parole. Ce concept est à mettre en regard avec celui d'espace de travail neuronal utilisé pour certains modèles neuronaux. *La conscience phénoménale*, qui correspond aux aspects qualitatifs de notre vie mentale (désignés aussi par qualia). En d'autres termes, « l'effet que cela fait » de ressentir une douleur, de la joie, de percevoir une couleur, etc. *La conscience réflexive* qui est notre capacité d'inspecter délibérément le cours de nos pensées, de faire de l'introspection ou de pister notre comportement. *La conscience de soi*, c'est-à-dire la représentation de soi qui confère une certaine unité à notre vie mentale.

Mais depuis le siècle dernier, deux disciplines scientifiques se sont introduites dans le débat, qui s'en trouve considérablement affiné par leurs apports. Ces deux disciplines sont les neurosciences et les sciences cognitives, plus exactement, dans cette dernière, la recherche en Intelligence Artificielle.

**L'essor des neurosciences**

De l'Antiquité au Moyen Age, médecins et savants se sont penchés sur le cerveau et ont tenté de comprendre son fonctionnement biologique. À partir du XIXe siècle, les découvertes s'accélèrent et la connaissance du cerveau prend un essor spectaculaire. En 1861, Paul Broca, anatomiste et anthropologue fait l'autopsie du cerveau d'un malade aphasique et explique la perte de la parole par une lésion dans le cortex frontal de l'hémisphère gauche. Durant la première moitié du XXe siècle nait la biologie des cellules nerveuses. Ramón y Cajal (prix Nobel 1906), père de la neurobiologie actuelle, établit les fondements de l'étude moderne du système nerveux. En 1924, Hans Berger réalise le premier électroencéphalogramme sur un être humain. En 1936, le Prix Nobel est attribué à Otto Loewi et Henri Dale pour leurs découvertes de la nature chimique de la neurotransmission. Michel Jouvet, neurophysiologiste, est à l'origine de la découverte du sommeil paradoxal. Antonio Damasio, directeur de l'Institut pour l'étude neurologique de l'émotion et de la créativité de l'Université de Californie du Sud, a mis en évidence le poids de nos émotions. Il souligne que le fonctionnement cognitif du cerveau en est largement tributaire. Jean-Pierre Changeux, neurobiologiste de l'Institut Pasteur, inscrit l'évolution de la structure fonctionnelle du cerveau dans une interaction importante avec l'environnement du sujet. Il émet ainsi l'hypothèse de la théorie de la sélection neuronale. Marc Jeannerod, neurobiologiste de l'Institut des sciences cognitives de Bron, cherche à comprendre le lien entre l'esprit et le cerveau en considérant l'imbrication entre la biologie et la philosophie.

Les nombreux progrès effectués en matière de traitement du signal, au cours du XXe siècle, ont permis par ailleurs de développer de nouveaux outils d'exploration et d'étude de l'activité cérébrale. Durant la deuxième moitié du XXe siècle sont ainsi apparues diverses techniques qui permettent aussi bien de connaître l'anatomie, la morphologie et l'activité cérébrale que de diagnostiquer les maladies neurodégénératives. Ces techniques se classent en deux groupes. Les premières mesurent indirectement l'activité du cerveau (imagerie métabolique) telles que la tomographie par émission de positons (TEP), la tomographie par émissions de photons (TSEP) et l'imagerie fonctionnelle par résonance magnétique (IRMf). La plus récente des technologies d'imagerie cérébrale est l'ISPIf, Imagerie Spectroscopique Proche Infrarouge fonctionnelle (en anglais fNIR), qui exploite le spectre des ondes lumineuses infrarouges (700-900 nm). Ces techniques (IRMf, TEP, ISPIf) permettent d'observer à la fois une augmentation de la consommation d'oxygène et une augmentation du débit sanguin cérébral là où les neurones sont stimulés. Elles permettent de visualiser les zones du cerveau qui s'activent lors de la réalisation d'une tâche et donnent donc une vision du cerveau dans sa totalité. Les secondes mesurent directement l'activité électromagnétique du cerveau telles que l'électroencéphalographie (EEG) et la magnétoencéphalographie(MEG). L'électroencéphalographie consiste à suivre la

progression des ondes électriques nées de la propagation des signaux nerveux *via* des électrodes posées sur le crâne, mais ne détecte que les ondes du cortex, la couche la plus externe du cerveau.

Les neurosciences abordent ainsi le décryptage du fonctionnement cérébral, du cerveau entier jusqu'à l'échelle moléculaire, chez le sujet normal et dans diverses maladies du système nerveux. L'étude des fonctions cognitives prend par ailleurs un nouvel essor grâce aux méthodes de la psychologie expérimentale couplées à la linguistique, à la philosophie, à l'intelligence artificielle... La jonction de ces « sciences cognitives » avec les neurosciences a donné naissance aux neurosciences cognitives qui étudient les bases neurales des fonctions mentales. La connaissance du monde mental de l'adulte et de l'enfant en a été bouleversée.

**Les neurosciences cognitives et les tentatives de modélisation de la conscience**

Dans un bulletin de médecine/sciences de l'INSERM de 1995, il est écrit « La neurobiologie est-elle en passe de s'attaquer aux phénomènes qui conduisent à la pensée consciente ? » L'article faisait suite à la publication la même année, dans la revue Nature, d'un article de Francis Crick, Prix Nobel, et Christof Koch, alors professeur d'informatique et de neurosciences au California Institute of Technology, intitulé « Are we aware of neural activity in primary visual cortex? ». Ils proposaient pour répondre à cette question de s'orienter vers la recherche des corrélats neuronaux de la conscience, c'est-à-dire des changements neuronaux qui se produisent au moment où émerge la conscience. L'imagerie cérébrale allait permettre de « voir » ce qui se passe dans le cerveau. Et ils donnaient quelques pistes pour aller dans cette direction. Par exemple, se concentrer d'abord sur les corrélats neuronaux de la conscience visuelle des primates. Leur article a eu un tel retentissement que des psychologues, des neuroscientifiques, mais aussi des physiciens, des philosophes, des médecins ont recommencé à s'intéresser à l'étude scientifique de la conscience. D'autres neurologues, et notamment Jean-Pierre Changeux, de l'Institut Pasteur, à Paris, leur emboîtent le pas. Ils apportèrent ainsi *de facto* un éclairage nouveau sur l'antique débat entre les matérialistes et dualistes.

Les neuroscientifiques n'affirment pas forcément que la conscience est matière, mais qu'elle « émerge » de l'activité des cellules du système nerveux. Et que la science peut étudier ce phénomène d'émergence. L'idée heurta alors les psychologues comportementalistes, pour lesquels la seule manière scientifique et objective d'étudier l'esprit consiste à observer ce qui est scientifiquement rapportable : le comportement, qui est la réponse de l'esprit à un stimulus. Toutefois, le travail de Francis Crick et Christof Koch trouva un accueil favorable chez les tenants de l'intelligence artificielle.

Le développement des neurosciences cognitives à la fin du XXe siècle a ainsi rendu possible les premiers modèles empiriques de la conscience humaine. Dans la recherche de ces modèles, les neuroscientifiques adoptent une définition opératoire de la conscience, c'est-à-dire une définition qui se prête aux observations et aux expérimentations. Telle celle exposée par Lionel Naccache, neurologue à l'Institut du Cerveau et de la Moelle Épinière (ICM), où la conscience est vue comme « notre capacité à nous rapporter subjectivement

nos propres états mentaux ». La démarche est alors d'explorer ce qui se passe physiquement dans le cerveau humain lorsque le sujet est conscient.

Les neuroscientifiques se sont donc mis en quête d'une signature neuronale de la conscience : un corrélat nécessaire et suffisant pour qu'une stimulation devienne consciente. Où et quand se produit-il ? Comment progresse-t-il ? À quel moment de cette progression la conscience émerge-t-elle ? Pour ce faire, les chercheurs exploitent l'imagerie cérébrale. Celle-ci ne donne certes pas à voir la conscience elle-même, mais les changements neuronaux qui se produisent en même temps que la prise de conscience.

En se limitant ainsi à l'identification de ces corrélats neuronaux, les neuroscientifiques se restreignent à résoudre ce que certains philosophes, on l'a vu, nomment le « problème facile de la conscience ».

**Modèles principaux de la conscience**

1. *Expérimentations en neurosciences ; Modèle de l'espace de travail neuronal global*

Autour des années 1960, Allen Newell, chercheur en informatique et psychologie cognitive à la compagnie RAND Corporation, John Clifford Shaw, programmeur de systèmes à RAND Corporation, et Herbert Simon, économiste et sociologue américain, prix Nobel d'économie 1978, créent des programmes, *Logic Theorist* puis *General Problem Solver* dans le but de résoudre des problèmes formalisés, par exemple celui de démontrer des théorèmes de logique (pour le premier) ou bien d'affronter des problèmes plus complexes (pour le second). Ces programmes sont considérés comme les premiers programmes d'intelligence artificielle. Ils furent les premiers à montrer l'utilité d'un espace de travail global - contenant l'objet à traiter - dans un système complexe constitué de circuits spécialisés – les opérateurs de traitement. La mise en commun de l'information traitée par chacun de ces circuits permettait de résoudre des problèmes qu'aucun circuit n'aurait pu résoudre seul. Dans les années 1980, le psychologue Bernard Baars, de l'Institut des neurosciences de La Jolla en Californie, fut le véritable promoteur de ce modèle. En disposant d'un espace de travail où l'information traitée par les circuits spécialisés est rendue ainsi accessible à l'ensemble de la population neuronale du cerveau.

Le principe de « l'espace de travail neuronal global » n'a cessé d'être repris et perfectionné depuis. En fait, la majorité des modèles neurobiologiques de la conscience intègrent certains aspects du concept d'espace de travail neuronal global. On peut citer ceux de Gerald Edelman, Antonio Damasio, Francisco Varela. Stanislas Dehaene, Jean-Pierre Changeux et Lionel Naccache, qui en ont fait leur hypothèse de base. Selon leur théorie, l'information sensorielle qui parvient au cerveau est traitée en permanence par des ensembles de neurones qui travaillent en parallèle, de manière inconsciente. Pour que leur information accède à la conscience, il faut que leur activité soit suffisante, mais aussi qu'ils bénéficient d'une amplification de la part des réseaux neuronaux où va émerger la conscience, à la manière d'une attention préconsciente. Une activité cohérente entre plusieurs populations de neurones distribués dans le cerveau s'installe alors. Les

connexions à longues distances qui s'établissent ainsi constituent l'espace de travail global. Cet espace met à disposition du cerveau cette information consciente qui peut dès lors être évaluée, mémorisée à long terme, donner lieu à des actions intentionnelles, etc. Cette mise à disposition généralisée d'un ensemble perceptif cohérent constituerait l'état conscient.

Ainsi, selon Stanislas Dehaene, lorsque notre cerveau est sollicité, les premières aires cérébrales à s'activer sont celles des cortex visuel et auditif. Parmi ces flots de représentations mentales inconscientes qui, à tout instant, traversent nos circuits cérébraux, l'une d'entre elles est choisie pour sa pertinence. Dès lors, des groupes de neurones s'activent de manière coordonnée, d'abord au sein des aires spécialisées, puis cette activation finit par envahir les lobes pariétaux et frontaux. Un vaste ensemble de neurones, répartis dans plusieurs régions cérébrales, « s'embrase » ainsi soudainement pour former un seul état neuronal global. C'est à ce moment-là que la conscience se manifeste.

Ce processus d'accès à la conscience se produit, répétons-le, en plusieurs étapes, en mobilisant un temps de latence important, d'environ un tiers de seconde après l'intervention du stimulus. Cette hypothèse de latence dans l'apparition de la conscience a fait l'objet de nombreux protocoles expérimentaux. L'expérience qui a eu le plus d'écho est celle qui a été conduite dans le laboratoire de Giulio Tononi, à l'Université du Wisconsin, à Madison.

La voie expérimentale principale mise en œuvre sur la plateforme NEUROSPIN – dirigée par Stanislas Dehaene – consiste à présenter un stimulus sensoriel (image ou son) juste au-dessous ou au-dessus du seuil de conscience du patient, de sorte que certaines informations sont subliminales alors que d'autres accèdent à la conscience. Ainsi toute une série d'expériences vont permettre de stimuler la conscience de différentes façons. On peut ainsi faire varier les conditions expérimentales pour détecter le passage d'une non-perception à une perception non consciente, puis consciente. L'électro et la magnétoencéphalographie (EEG et MEG) permettent de suivre l'activité neuronale en quasiment temps réel (milliseconde). On a découvert ainsi la grande profondeur du traitement inconscient de l'information. Pratiquement toutes les régions du cerveau peuvent travailler de façon non consciente. On a découvert aussi et surtout, les signatures d'accès à la conscience. La première signature est une énorme et subite amplification de l'activité neuronale dans de nombreuses régions du cortex, particulièrement dans les aires pariétales (postérieures) et préfrontales bilatérales du cerveau. Lorsqu'une image ou un mot montré est vu de manière consciente, tout s'active simultanément, un phénomène *a priori* assez surprenant car il s'agit de régions très dispersées à travers le cortex. En parallèle se manifeste une deuxième signature, l'EEG montre le déploiement d'une onde électrique lente et de grande ampleur, appelée P300, au sein de la région pariétofrontale. Toujours selon Stanislas Dehaene *« L'onde P300, de l'ordre de quelques microvolts, démarre vers 270 millisecondes, atteint son pic entre 350 et 500 ms après l'arrivée du stimulus visuel, et perdure jusqu'à 600, 700 voire 900 ms. Autrement dit, notre conscience est très en retard sur le monde extérieur ! ».* Deux autres signatures ont été individualisées : une synchronisation des signaux électriques que s'échangent les aires corticales les plus éloignées les unes des autres, signe que tous

les neurones impliqués se mettent à travailler en même temps ; et une explosion tardive et soudaine d'ondes de haute fréquence.

2. *Un modèle conçu à partir d'une phénoménologie de la conscience : la théorie de l'information intégrée*

En parallèle – à ces approches basées sur les fonctions de la conscience – une théorie a été développée à partir d'une analyse en quelque sorte des propriétés *ressenties* de nos états conscients. Elle a donc des objectifs plus ambitieux mais est plus spéculative ; c'est l'approche dite de l'information intégrée, proposée en 2004 par Giulio Tononi. Il avait auparavant travaillé avec Gerald Edelmann (prix Nobel 1972) à l'Institut de neurosciences de La Jolla en Californie. Cette théorie énonce cinq propriétés « phénoménologiques » qui caractérisent selon elle toute expérience : 1) une expérience consciente est quelque chose de *réel,* elle existe *indépendamment de tout observateur* ; 2) cette expérience est composée de parties liées entre elles ; 3) cette expérience se distingue de toute autre expérience possible, sa composition lui étant spécifique ; 4) elle est unifiée, constitue un tout irréductible : tous ses composants sont interdépendants, on ne peut les séparer sans la faire disparaitre ; 5) elle est limitée quant à son contenu et sa durée. Puis la théorie postule qu'à chacune de ces cinq propriétés phénoménologiques correspond une propriété « physique » de la structure et de la dynamique du système cérébral lorsqu'il « vit » cette expérience consciente. Autrement dit elle traduit les propriétés phénoménologiques en des propriétés physiques qui doivent être satisfaites pour que l'on puisse dire qu'un substrat physique (ici le cerveau) est support de conscience. Ces propriétés physiques s'expriment mathématiquement (dans le langage des probabilités) et leur réalisation plus ou moins poussée et riche de contenu donne lieu à une évaluation quantitative appelée Information conceptuelle intégrée. Le degré de conscience est d'autant plus grand que cette information est élevée. Ainsi, tout objet, vivant (ou inerte), peut être gratifié, dans un certain état - c'est-à-dire à un certain instant de son existence - d'une conscience qui peut être nulle, ou au contraire maximale. Mais seuls certains systèmes physiques ont le degré d'architecture nécessaire pour soutenir la conscience. Cette théorie est soutenue et co-développée par Christof Koch, président de l'Institut Allen pour les sciences du cerveau, à Seattle. Il y a d'autres modèles de la conscience que nous ne développerons pas ici, comme par exemple le modèle du Schéma de l'attention, promu par Michael Steven Graziano, Université de Princeton, modèle qui se rattache par certains aspects aux théories philosophiques d'ordre supérieur évoquées plus haut.

**Mécanismes neuronaux et marqueurs somatiques**

Par ailleurs, dans ces processus de prise de conscience, les mécanismes neuronaux du cerveau ne sont peut-être pas les seuls à être mobilisés ; il semblerait que des marqueurs somatiques (liés au corps) de notre vie mentale contribueraient aussi à notre vie consciente. Catherine Tallon-Baudry et ses collègues du Laboratoire de Neurosciences Cognitives & Computationnelles de l'ENS, ont récemment découvert que la façon dont notre cerveau sent et analyse les battements du cœur pourrait jouer un rôle dans notre conscience de soi. L'équipe de recherche de Lionnel Naccache de l'ICM a également montré que lorsque nous

prenons conscience d'un stimulus, non seulement notre cerveau présente les signatures habituelles mais nos pupilles se dilatent, notre cœur accélère, etc.

**Tester ces modèles de la conscience humaine : une question d'actualité**

Dans une publication récente (2019) parue dans Nature Human Behaviour, un collectif de neuroscientifiques, biologistes et psychologues insistait sur la nécessité de tester les différents modèles de façon rigoureuse. C'est l'ambition du programme « Accelerating Research on Consciousness », lancé par la fondation Templeton World Charity. Pour la première fois, un protocole va permettre de confronter les principales théories de la conscience. Le but étant de sélectionner les modèles les plus explicatifs et scientifiquement testables. La première phase du projet a été annoncée en octobre 2019, lors de la réunion de la Society for Neuroscience, à Chicago. Ces expériences seront dirigées par Lucia Melloni, neuroscientifique à l'Institut Max-Planck. Le programme débutera par la confrontation de la théorie de l'espace de travail global à celle de l'information intégrée. Rappelons que ces deux théories sont proposées respectivement par Stanislas Dehaene, titulaire de la chaire de psychologie cognitive expérimentale au Collège de France, et par Giulio Tononi, de l'Université du Wisconsin, et Christof Koch, président de l'Institut Allen pour les sciences du cerveau, à Seattle.

En ce qui concerne la mémoire humaine, Francis Eustache de l'Université de Caen et ses collaborateurs ont montré que ce n'est pas une fonction localisée. La mémoire qui fonctionne chez le sujet sain est le résultat d'une synchronisation harmonieuse de différents réseaux. D'aucuns font ainsi un parallèle entre certains phénomènes physiques et le fonctionnement du cerveau, à savoir aussi bien la conscience que la mémoire présentent finalement un aspect de non localité. On a mentionné au début de cette introduction un grand physicien, Erwin Schrödinger ; on peut citer plus près de nous un autre grand savant Alfred Kastler (prix Nobel de Physique 1966), il publia en 1976 un ouvrage ayant pour titre « Cette étrange matière ». Dans la dernière partie de son ouvrage, il livre des réflexions que ses connaissances lui inspirent : « *La physique contemporaine me paraît confrontée à un autre problème critique, celui de la matière vivante. Il me semble que la physique pourrait jouer un rôle capital dans le développement futur de la biologie, si la recherche confirmait une idée qui se profile aujourd'hui, selon laquelle il pourrait exister un rapport entre la* cohérence *en physique et l'*ordre *biologique, caractérisés l'un et l'autre par une* néguentropie *remarquable* ».

**Le renouveau de l'Intelligence Artificielle**

Le *rêve* d'une machine *intelligente* remonte à l'Antiquité. Une des plus anciennes traces du thème « l'homme dans la machine » date de 3000 ans avant notre ère en Égypte. Durant l'année 2019 on a commémoré le 500ème anniversaire de la disparition de Léonard de Vinci à Amboise. Rappelons que l'histoire retient qu'en 1495**,** Léonard de Vinci invente le premier robot humanoïde : un chevalier activé mécaniquement. Mentionnons également les automates de Jacques Vaucanson et son fameux Canard *digérateu*r exposé en 1744 au Palais Royal**.**

En ce début du XXIe siècle, en parallèle avec les développements des neurosciences cognitives, des progrès spectaculaires ont été accomplis dans les domaines de l'Intelligence Artificielle et de la robotique. En octobre 2016, l'Académie des Sciences a organisé une conférence débat ayant pour titre « Intelligence artificielle : le renouveau ». Et de poursuivre, dans ses motivations : « découvrir, apprendre, reconnaître, juger, décider : ces tâches perceptives et cognitives – que l'on associe à l'intelligence humaine – deviennent chaque jour plus accessibles à l'automatisation. Grâce aux progrès considérables de la microélectronique, à la puissance de calcul qu'elle permet et à l'accès à des quantités gigantesques de données, l'Intelligence Artificielle (IA) vit aujourd'hui un renouveau qui s'appuie sur presque toutes les sciences et touche de plus en plus à notre vie quotidienne. »

Selon Jean-Paul Haton, il est d'usage de distinguer deux types d'IA en fonction des capacités des tâches envisagées : l'IA faible et l'IA forte. L'IA faible est celle des systèmes actuels, atteignant des résultats de très haut niveau, souvent comparables ou supérieurs à ceux d'êtres humains, mais dans des domaines restreints bien délimités (jeux, diagnostic, reconnaissance de la parole, identification d'images, etc.) pour lesquels une capacité d'apprentissage spécifique a été mise en place. Ces systèmes d'IA ne sont pas destinés à remplacer des humains, mais plutôt à coopérer avec eux pour optimiser leurs performances. L'IA forte tend vers celle de l'être humain, capable d'apprendre à mener des tâches complexes dans des domaines très différents, ou de comprendre et de raisonner sur des sujets variés en se fondant sur l'expérience acquise. L'IA forte est encore largement dans les laboratoires de recherche. Quant à savoir si l'IA risque de « dépasser » l'intelligence humaine (lors de ce qui est désigné comme une singularité) et dans cette éventualité, quels avantages et quels dangers se présenteraient à l'Humanité, il existe sur ce sujet tout un débat que cet ouvrage ne peut qu'effleurer.

**Intelligence et conscience Artificielles ?**

Ainsi, à propos de la singularité technologique, John Von Neumann, le célèbre mathématicien et physicien pionnier de l'Informatique, a-t-il pu écrire, selon Jean-Gabriel Ganascia, Sorbonne Université, spécialiste d'intelligence artificielle, Président du Comité d'éthique du CNRS – « ce n'est pas parce qu'on peut créer des capacités exponentielles de la technologie, que la technologie peut dépasser l'homme ». De même, selon Luc Steels, Université libre de Bruxelles, spécialiste dans l'intelligence et la vie artificielles appliquées à la robotique et à l'étude du langage « Il semble possible de construire des programmes hautement complexes, équivalents en termes de performance à l'intelligence humaine pour un domaine particulier, mais ces programmes seront incapables d'intégrer l'évolution ou la nature contextuelle de l'intelligence ».

Certains scientifiques estiment – à l'inverse – à la suite du mathématicien Alan Turing que les machines pourraient être un jour suffisamment perfectionnées pour être dotées d'une conscience. Tel Richard Frackowiak, pionnier de l'imagerie cérébrale, neuroscientifique au Centre hospitalier universitaire de l' Université de Lausanne. Richard Frackowiak est codirecteur du « Human Brain Project », le projet européen de simulation du cerveau humain dont le but vise à terme à simuler le fonctionnement entier du cerveau. Il est un

fervent défenseur de l'imbrication étroite entre la biologie du cerveau et l'informatique. C'est de la matière que naît la conscience dit-il. Christof Koch avance de son côté que nombre de neurologues sont convaincus que la modélisation du fonctionnement des neurones sur les ordinateurs permettra de *simuler* la conscience. Mais pour lui, l'architecture actuelle dite de Von Neuman, avec sa structure de connectivité électronique, empêche l'émergence d'une « vraie » conscience artificielle, c.-à-d. qui ne se réduise pas à une simulation ; chaque transistor n'est relié en effet qu'à quelques autres tandis que, dans nos cerveaux, chaque neurone est typiquement connecté à dix mille voire vingt mille d'entre eux. Ainsi aussi puissant soit-il, un ordinateur n'atteindra donc jamais un haut degré de conscience. Pour reprendre l'exemple du « Human Brain Project », même une simulation complète et précise du cerveau humain incluant la centaine de milliards de neurones et leur matrice de millions de milliards de connexions (les synapses) ne pourra pas être consciente. Des architectures autres que celle de Von Neuman, telles que les puces neuromorphiques, peuvent effectuer des tâches en parallèle (recevoir, transporter l'information, la traiter ou la stocker). Elles travaillent de ce fait de façon semblable aux neurones cérébraux. Si de tels éléments venaient un jour à être hautement interconnectés, on se rapprocherait davantage de l'architecture du cerveau. Une telle machine serait-elle alors capable d'expériences conscientes substantielles ? Ce changement soulèverait des questions d'éthique majeures.

*Essais de modélisation d'une certaine forme de conscience pour les « systèmes intelligents »*

Un des objectifs des recherches en IA a toujours été de comprendre les processus cognitifs humains, en en fournissant des modèles. Il est donc normal que la question des modèles de la conscience, ou du moins de certaines de ses fonctions, soit posée dans cette discipline, et en particulier dans ses applications robotiques.

En robotique la question controversée de savoir si les machines peuvent être conscientes a fait l'objet récemment d'un examen attentif dans l'article « What is consciousness, and could machines have it? » paru dans la revue *Science* d'octobre 2017 dans lequel trois chercheurs, Stanislas Dehaene, Hakwan Lau –Département de psychologie et Institut de recherche sur le cerveau, Université de Californie à Los Angeles et Département de psychologie, Université de Hong Kong – et Sid Kouider du Laboratoire de Sciences Cognitives et Psycholinguistique (École Normale Supérieure, École des Hautes Études en Sciences Sociales, CNRS) passent en revue le chemin qu'il reste à parcourir pour que les robots, les systèmes intelligents artificiels, puissent accéder à un certain niveau de conscience. Pour cela, ils se réfèrent à la conscience humaine. Les expériences menées à NEUROSPIN à l'aide d'images subliminales ont montré que le cerveau humain fonctionne principalement selon un mode non conscient, appelé $C_0$ dans l'article de *Science.* Les systèmes actuels d'IA à base de réseaux de neurones : reconnaissance des images, des sons, du langage… fonctionnent sur ce mode non conscient.

La conscience se développerait sur une couche supplémentaire. Pour accéder à cette couche consciente, le cerveau humain met en place deux sous-niveaux de traitement de

l'information aux fonctions bien distinctes et complémentaires. Dans le premier sous-niveau, C1, appelé « disponibilité totale », on peut passer d'une information consciente à une autre de manière fluide et flexible et lui donner du sens. Il correspond aux capacités du cerveau à sélectionner, parmi toutes les entrées perceptives présentes à un instant donné, celles qui sont le plus en rapport avec la poursuite d'une action en cours ou en projet ; à relier l'information sélectionnée avec d'autres présentes, avec les connaissances antérieurement acquises, les comportements requis, les rapports verbaux possibles ; à maintenir dans le temps cette attention prioritaire liée à l'action et nécessaire à son accomplissement, et à la modifier le cas échéant si cela semble nécessaire.

Le sous-niveau C2 permet d'acquérir « la conscience de soi ». Il correspond, notamment, aux capacités de disposer d'informations sur soi-même et ses processus d'inférence et donc de pouvoir les gérer ; autrement dit, aux capacités d'introspection, d'auto-évaluation, de métacognition : disposer de représentations de ses propres connaissances et aptitudes, et être à même d'en faire bon usage dans telle ou telle circonstance.

Le niveau C1 n'est pas accessible aux systèmes d'IA actuels. Ils ont la capacité - mais de manière inconsciente - de donner du sens à des informations, de prendre des décisions et aussi d'apprendre.

Par contre le sous-niveau C2 (conscience de soi) apparaît comme possible - sous certains aspects - à des robots. Sa modélisation au moyen d'algorithmes semble réalisable.

Des robots sont programmés, par exemple, pour surveiller leur propre progression d'apprentissage. D'après Stanislas Dehaene, une forme de méta-cognition serait envisageable pour nos appareils connectés.

Citons, à titre d'exemples pour illustrer ces propos, des tentatives qui sont actuellement faites pour doter les robots d'éléments de conscience dans des laboratoires d'Europe et des Etats-Unis. Pour en évaluer les résultats on peut envisager d'utiliser le test du miroir, comme on a pu le faire pour estimer la conscience de soi chez un animal ; l'entité placée face à un miroir individualisant une marque qui lui a été imprimée montre qu'elle est ainsi consciente de son propre corps.

Un autre concept est aussi utilisé. Il est issu de la philosophie de l'esprit – développé initialement par le psychologue et philosophe Georges Frederic Stout. Ce concept a vraiment émergé dans les années soixante-dix avec la parution du livre « Body and mind » du philosophe Keith Campbel, lequel introduit des individus « Zombies » qui se comportent de manière indiscernable par rapport à des personnes conscientes selon tous types de tests concevables par la science. Depuis, ce concept a été largement utilisé et développé par le philosophe David Chalmers – dont nous avons déjà parlé – pour montrer l'insuffisance des explications en termes de processus physiques lorsque celles-ci concernent les aspects subjectifs de la conscience. David Chalmers a proposé une expérience de pensée où le rôle principal est tenu par un « p-zombie » ou philosophe zombie.

En robotique le rôle de zombie peut être tenu par un robot. Les robots-zombies permettent de montrer que certains processus cognitifs décrits principalement par le sous-niveau C2 peuvent être reproduits sur un robot sans que cela soit suffisant pour le rendre conscient. C'est une contribution de la robotique à l'avancement des neurosciences et de la réflexion sur la nature de la conscience.

En France le projet ANR *Roboergosum* porté par Raja Chatila et son équipe de l'Institut des Systèmes Intelligents et de Robotique (ISIR), Sorbonne université, et le Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS/CNRS) de Toulouse « vise à comprendre les mécanismes sous-jacents à l'émergence de la conscience vue comme un processus au centre de l'interaction d'un agent et de son environnement. Le but étant de concevoir un système cognitif où seraient mis en œuvre et s'exprimeraient ces mécanismes et ayant donc une possible conscience de lui-même ».

En 2015, Selmer Bringsjord, Rensselaer Polytechnic Institute à New York, a effectué avec trois petits robots humanoïdes de type Nao un test classique appelé le puzzle des sages en vue de résoudre des énigmes logiques nécessitant un élément de conscience de soi. Le but poursuivi est de montrer la viabilité d'exemples spécifiques et limités de la conscience. Très récemment en 2019, Robert Kwiatkowski et Hod Lipson de l'Université de Columbia ont publié dans la revue *Science Robotics* des résultats obtenus sur un robot capable de s'auto-modéliser sans aide extérieure. Ils sont parvenus à doter le robot d'une représentation de lui-même sans aide extérieure ni connaissance préalable en utilisant « l'apprentissage profond », technique d'apprentissage automatique efficace. Les scientifiques ont utilisé un bras robotique articulé à quatre degrés de liberté. Initialement, le robot se déplaçait de manière aléatoire et collectait environ un millier de trajectoires, comprenant chacune une centaine de points. Le robot a ensuite utilisé l'apprentissage profond pour créer un auto-modèle. Les premiers auto-modèles étaient assez imprécis et le robot ignorait ce qu'il était, et comment ses articulations étaient connectées. Mais après une trentaine d'heures d'entraînement, l'auto-modèle est devenu compatible avec le robot physique avec une précision de quelques centimètres. Pour tester si l'auto-modèle pouvait détecter des dommages sur lui-même, les chercheurs ont imprimé en 3D une pièce déformée pour simuler des dommages et le robot a pu détecter le changement et reconfigurer son auto-modèle. Les auteurs de cette étude sont conscients des implications éthiques. « *La conscience de soi mènera à des systèmes plus résilients et adaptatifs, mais elle induira également une certaine perte de contrôle par les humains* ». C'est une technologie puissante qui devrait être manipulée dans un cadre éthique strict.

Dans les décennies à venir, les progrès rapides des algorithmes d'apprentissage profond engendreront des machines d'une intelligence « comparable » à la nôtre. Elles seront capables de parler et seront dotées d'une certaine forme de raisonnement ; et auront ainsi leur place dans de nombreux domaines, comme l'économie, la politique et… les applications sécuritaires et militaires. La naissance d'une véritable intelligence artificielle affectera l'avenir de l'humanité et conditionnera l'existence même d'un tel avenir.

**Rappel du plan de l'ouvrage**

Les textes rassemblés[3] prennent la forme de chapitres, et distribués sur quatre parties : *première partie,* travaux en neurosciences et psychologie expérimentale ; *seconde partie,* sciences cognitives et Intelligence Artificielle ; *troisième partie,* réflexions sur l'Intelligence, la Conscience et l'impact de l'IA sur les activités humaines ; *quatrième partie,* synthèse des discussions de la Table Ronde tenue à l'issue du colloque. Un aperçu de ces quatre parties est donné dans leurs présentations respectives

Enfin dans un court épilogue nous ferons part aux lecteurs des réflexions et questions que la lecture de ces textes associés à la prise de connaissance de la littérature actuelle concernant la conscience nous a inspirés.

Pour le comité de lecture de l'AEIS[4]

[3] La majorité des textes rassemblés émanent des conférenciers eux même ; dans d'autres cas, ils ont été élaborés à partir d'une transcription de la conférence concernée, puis soumis au conférencier pour correction, compléments et validation.

[4] Victor Mastrangelo (Président de l'AEIS) et Jean-Pierre Treuil (coordination comité de lecture)

# PREMIÈRE PARTIE

## TRAVAUX EN NEUROSCIENCE ET PSYCHOLOGIE EXPÉRIMENTALE

## Première Partie

# Travaux en neurosciences et psychologie expérimentale

## Présentation

La première partie de l'ouvrage présente six chapitres ayant tous en commun d'être basés sur des observations ou des expériences en neurosciences ou en psychologie cognitive, en collaboration avec des personnes, sujets de ces expériences. Ils diffèrent cependant par les fonctions ou activités cérébrales sur lesquelles ils portent leur attention, et aussi par les techniques d'investigation mises en œuvre.

Les deux premiers chapitres concernent divers modèles de la *mémoire humaine* et les travaux visant à les évaluer. Les deux chapitres suivants sont centrés sur la conscience, sous certaines des acceptions de ce terme polysémique : le premier d'entre eux traite de la *conscience perceptive* et ses corrélats neuronaux ; différentes recherches y sont abordées, toutes associant techniques d'investigation sur des sujets témoins, utilisées en psychologie expérimentale, et imagerie cérébrale ; le second traite du flux de conscience, conscience considérée comme *capacité d'attention et de concentration.* L'imagerie cérébrale en est absente, mais des expériences originales de psychologie expérimentale, notamment sur des enfants souffrant de déficit d'attention, y sont présentées. Les deux derniers chapitres sont consacrés à des questions de localisation cérébrale comparée de certains processus cognitifs : *langage versus processus visuels* pour le premier, *langage versus activités mathématiques* pour le second. Puis nous avancerons quelques questions qui nous ont paru pertinentes au fil de leur lecture.

**Mémoire et représentation**. Sous ce thème Alberto Oliverio commence par présenter différents modèles de mémoire et les structures cérébrales qui ont pu leur être associées : modèle de Squire, modèle anatomo-fonctionnel, modèle représentationnel hiérarchique… Un modèle de mémoire est d'abord une différenciation de la mémoire en différents types. On distinguera par exemple la mémoire épisodique ou mémoire contextualisée, la mémoire sémantique ou déconceptualisée, ou encore la mémoire procédurale, liée à la mise en pratique quasi-inconsciente de certains apprentissages. C'est ensuite une description des relations - éventuellement hiérarchiques - entre ces différents types de mémoire, et des propositions quant aux structures cérébrales qui les implémentent. A partir notamment de ses propres travaux, Alberto Oliverio interroge ensuite la pertinence d'une distinction trop stricte entre les différents types de mémoire proposés par ces modèles, et corrélativement la pertinence d'une attribution trop rigide de tel type de mémoire à telle structure cérébrale.

Il souligne ainsi que certaines structures cérébrales sont des fonctions multiples. Il termine en traitant de la différence entre une mémoire « archivage » - semblable à celle d'un ordinateur, et la mémoire humaine, où les souvenirs sont en perpétuel restructuration.

**La Mémoire Humaine et ses substrats cérébraux**. C'est à nouveau au fonctionnement de la mémoire humaine que Francis Eustache et Armelle Viard nous invitent à nous intéresser. Francis Eustache, dans une première partie du chapitre, rappelle comment les troubles pathologiques de mémoire, spontanés ou consécutifs à des opérations chirurgicales ou traumatismes cérébraux, ont constitué peu à peu, depuis les premières observations de Paul Broca en 1861 et très fortement depuis les années 1960, une source d'inférences dans l'élaboration des modèles théoriques de la mémoire humaine. Il développe une analyse de la genèse et de l'évolution de ces modèles et des concepts qui en sont le socle ; il termine par une brève introduction au modèle MNESIS qu'il a élaboré avec son équipe en montrant ses relations avec les modèles qui ont précédé. Dans la seconde partie du même chapitre, Armelle Viard traite spécifiquement du fonctionnement de la mémoire « autobiographique » et des structures qui y sont impliquées, en comparaison avec la mémoire « sémantique ». Elle dresse un tableau de différents travaux expérimentaux qui ont été menés sur ces thèmes, notamment en utilisant la neuroimagerie, en présente et en discute les résultats. Comment s'effectue la reconstruction/récupération d'un souvenir, quels paramètres interviennent dans de tels processus, comme l'émotion, l'ancienneté, quelles sont les régions cérébrales impliquées, telles sont quelques unes des questions traitées. Elle termine par une analyse des problèmes de la mémoire autobiographique dans les pathologies neurodégénératives comme l'Alzheimer ou la démence sémantique.

**Les bases Neurobiologiques de la conscience**. La conférence faite par Claire Sergent transcrite ici par nous-même dans un chapitre qu'elle a validé, aborde un point central. Claire Sergent y présente une synthèse des avancées récentes obtenues, par elle en particulier, au sein de la grande équipe française (Collège de France, CNRS, Université Paris-Descartes, Hôpital Pitié-Salpêtrière, Institut du cerveau) qui, depuis Jean-Pierre Changeux, passant par Stanislas Dehaene et par Lionel Naccache, avec leurs collaborateurs, s'est intéressée au fonctionnement du cerveau humain et aux corrélats neuronaux de la conscience. Utilisant les techniques modernes d'imagerie médicale et pour certaines expériences celles d'Électro-encéphalographie, Claire Sergent montre, à travers diverses expériences de masquage/démasquage, comment on a pu mettre en évidence (méthode contrastive), caractériser qualitativement et quantitativement, les différences d'activations cérébrales entre processus perceptifs conscients et processus perceptifs inconscients. Elle poursuit en indiquant les interprétations données à ces résultats, sous forme d'un *modèle de prise de conscience* impliquant l'activation et la réactivation d'un *espace de travail global* en contact avec plusieurs aires cérébrales n'ayant pas de contact direct entre elles ; elle évoque à ce propos un modèle informatique de connexion neuronal, permettant d'étudier par simulation la survenue de tels processus d'activation globale. Elle termine en montrant quels apports peuvent avoir ces avancées dans le contexte du diagnostic des états conscients en clinique humaine. Dans sa conclusion, elle affirme que les recherches à venir

permettront d'explorer le chemin inverse de celui suivi jusqu'ici, et donc «de partir de l'activité cérébrale pour aller vers l'expérience subjective » de l'individu.

**Quelles données subjectives pour l'étude du flux de conscience ?** Dans le second chapitre spécifiquement centré sur la conscience, Jérôme Sackur revient sur les résultats présentés par Claire Sergent (et par d'autres chercheurs utilisant la méthode dite contrastive). Il souligne que, avec cette méthode, sans nier l'importance des résultats obtenus, n'apparaissent que des moments arrêtés, instantanés de la conscience, alors que d'après lui la conscience est un flux permanent, « une rivière qui coule ». Il rappelle d'ailleurs que W. James, un des pères de la psychologie, disait cela il y a 100 ans et que lui-même, Jérôme Sackur, n'avait compris ce que James voulait dire, que récemment ; ce flux de conscience varie en permanence tant en intensité qu'en rapidité. Il met alors au point des techniques particulières - dont il discute les faiblesses et les forces - pour explorer chez les enfants, à chaque instant, la réaction à des stimuli intervenants au cours d'une tâche que ces enfants effectuent. Il met ainsi en évidence, dans le flux de conscience, des périodes de trouble de l'attention, comme par exemple « la rêverie éveillée ». C'est, dit-il, un caractère essentiel de l'esprit humain et de sa conscience, de ne pas être systématiquement et en permanence concentrés sur une tâche, quoiqu'il puisse exister sur ce plan de grandes variations entre individus. Appliquant ses découvertes aux enfants ayant des troubles de l'attention, Jérôme Sackur montre que cette étude a aussi des applications pratiques, en particulier dans la thérapeutique des enfants inattentifs à l'école ou dans la vie.

**Spécialisation hémisphérique de l'attention visuo-spatiale**. **Relations complémentaires entre visualisation spatiale et langage**. Laure Zago étudie la propriété la plus spécifique du cerveau humain : la différence fonctionnelle et la complémentarité des fonctions des deux hémisphères du cerveau, qui donnent à l'Homme un avantage décisif parmi les espèces animales. Cette dissymétrie a un impact sur la phénoménologie de la conscience, comme le montre par exemple la non prise de conscience de certains stimuli chez des patients affectés de certaines lésions cérébrales unilatérales. C'est d'ailleurs par des questions concernant une notion reliée à la conscience, savoir l'attention, que Laure Zago aborde son propos. Elle traite ainsi des *biais attentionnels*, supposés liés à la latéralisation ; biais comportementaux dont elle liste différents types et dont elle décrit les caractéristiques chez des patients ayant subi des lésions cérébrales affectant l'hémisphère droit, et en comparaison chez les sujets sains. Elle montre certains résultats de ses travaux autour de la corrélation entre l'importance de ces biais et l'intensité de la latéralisation des processus concernés. Une dernière section reprend la question de l'origine de la spécialisation confiant les processus visuo-spatiaux à l'hémisphère droit et ceux du langage à l'hémisphère gauche. Elle présente des résultats obtenus sur un échantillon de sujets gauchers, qui tendraient bien à prouver que cette population minoritaire (moins de 10% de la population générale) est bien une population à privilégier pour « la compréhension des règles de mise en place de la spécialisation hémisphérique des fonctions cognitives latéralisées »

**Représentation et manipulation des concepts mathématiques par le cerveau humain**. Le chapitre rédigé par Marie Amalric trouve sa place logique à la suite du précédent ; il concerne en effet la localisation de processus cognitifs précis, savoir ceux liés aux activités mathématiques. Certains auteurs pensaient et peut-être pensent toujours, que ces processus sont proches de ceux concernés par le langage, alors que d'autres hypothèses – et aussi l'intuition de certains chercheurs - en font des processus apparentés aux processus visuo-spatiaux. C'est à de telles questions que Marie Amalric s'est confrontée dans ses travaux, menés sous la direction de Stanislas Dehaene et utilisant des techniques d'imagerie médicale. Marie Amalric rappelle d'abord que des travaux antérieurs avec des enfants de quelques heures, et avec des enfants indiens Murukus qui n'apprennent pas à lire, ont montré que nous possédons dès la naissance un noyau minimal donnant une approche intuitive de la quantité, du nombre, et du positionnement dans l'espace. Ces études suggèrent qu'il existe une capacité innée que l'on peut qualifier de proto-mathématique, qui s'étend ensuite par apprentissage. Puis elle présente les expériences qu'elle a conduite, menées comparativement avec des sujets mathématiciens et non mathématiciens, leur soumettant ce qu'on pourrait appeler des exercices de compréhension d'énoncés de différents types, mathématiques et non mathématiques, et observant ce qui se passe dans le cerveau à l'aide d'imagerie IRMf. Les tests sous IRMf montrent bien que les zones impliquées dans la résolution des questions mathématiques sont différentes de celles impliquées dans la compréhension syntaxique et sémantique du langage ordinaire. C'est ce qui est observé, autant avec des mathématiciens professionnels qu'avec des non-mathématiciens. Ils montrent aussi qu'à l'échelle à laquelle on peut les observer, les mêmes zones du cerveau sont sollicitées par les questions mathématiques, quel que soit leur niveau et quel que soit la discipline mathématique concernée : de manière restreinte pour des questions basiques, et plus étendue pour des questions de plus haut niveau. Marie Amalric expose ensuite des investigations complémentaires et leurs conclusions, touchant la localisation du raisonnement logique et les liens avec les processus visuo-spatiaux. L'ensemble des expériences menées par Marie Amalric semblerait ainsi indiquer que le noyau de départ qui héberge les capacités proto-mathématiques dès la naissance, reste ensuite impliqué dans l'exercice des mathématiques au sein d'un secteur étendu qui croîtrait autour de ce noyau initial. Stanislas Dehaene parle à ce propos de « recyclage neuronal »

Pour le comité de lecture de l'AEIS[1]

[1] Gilbert Belaubre, Eric Chenin, Pierre Nabet, Alberto Oliverio,

# 1 Rôle essentiel de la mémoire dans la formation de toute représentation

**Alberto Oliverio**

Laboratoire de Psychobiologie
Université La Sapienza, Rome

## Abstract

The relationships between memory and representation raise a number of problems today. The rather classical model postulates that the various structures of the medial temporal lobe play different roles in declarative memory, a form of memory accessible to conscience. The representational-hierarchical view of amnesia, on the other hand, rejects a strong relationship between declarative memory and medial temporal lobe structures. As a consequence of such an approach these structures would not be involved in declarative memory only but in any task, cognitive, even perceptive, requiring complex representations.

However, there are other brain structures that play an important role in memory, such as the basal ganglia, which control cognitive activities such as procedural memories, motivational components of learning, and the execution of motor actions. In addition to that, even if the subdivision between procedural and declarative memories has its own rationality it must not be considered in absolute terms: many declarative memories, repeated and recurrent over time, can be proceduralised, that is to say, transferred to another register belonging to the basal ganglia. These structures, aside from their classic role in motor function, also mediate a variety of learning and memory processes.

A final point concerns the subtle transformations of procedural into declarative memories in the child, a fact emphasizing the importance of motor skills within mental representative processes.

# 1. Introduction

Il y a environ cinquante ans, l'Association de Psychologie de Langue Française avait organisé une conférence commémorative à Genève avec des intervenants de renom tels que Jean Piaget, Alfred Fessard, Daniel Bovet, Nico Frijda, Bärbel Inhelder, César Florès et bien d'autres. Je me souviens de cette conférence pour sa vivacité mais aussi pour les problèmes qu'elle soulevait en relation avec les bases biologiques de la mémoire et de leur rôle dans la formation de chaque représentation. À cet égard, je cite une phrase de l'un des premiers experts dans le domaine des analogies entre mémoires biologiques et mémoires artificielles, Nico Frijda (1970), selon lequel « dans la mémoire humaine il y a probablement un travail assez complexe pour transformer les données brutes en représentations internes ». Pour sa part, Jean Piaget (1970) observait avec plus d'optimisme que « Il n'existe aucune raison d'admettre l'existence de deux catégories de schèmes, l'un concernant la mémoire, l'autre l'intelligence ». Et Piaget, qui était renommé pour ses calambours disait aussi « que sans la mémoire il n'y a pas d'intelligence mais que la mémoire n'est pas l'intelligence ».

Bien des années plus tard, de nombreux aspects de la mémoire ont été heureusement clarifiés surtout en termes de bases neurobiologiques, alors que, je crois, il reste encore du chemin à parcourir pour ce qui concerne le thème de la représentation. En termes indicatifs, la définition la plus commune de la notion de représentation est celle qui la considère comme une connaissance ou un savoir sur quelque chose (un objet, une personne, un évènement...). La représentation cognitive serait la représentation en mémoire à long terme d'un savoir acquis par un individu. Pour François Bresson il s'agit d'une connaissance basée sur la relation entre deux systèmes d'objets (réels ou mentaux) : l'un étant le représentant de l'autre, le représenté.

Les relations entre la mémoire et la représentation posent aujourd'hui un certain nombre de problèmes. Une représentation, par exemple, impliquerait une activité symbolique, activité qui est certainement au centre de notre aptitude sémantique et syntaxique au langage. Il n'est donc pas étonnant que, lorsque nous pensons à la façon dont le cerveau peut répéter une action, nous soyons tentés de dire que le cerveau a des représentations. Pour ce qui concerne la représentation distribuée des connaissances l'information serait traitée par différentes unités (neurones) qui sont activées en même temps. Selon plusieurs auteurs, comme Dehaene, Changeux et Nadal (1987), dans un réseau neuronal c'est la configuration globale d'activation qui détermine la représentation d'un concept. Dans ce cadre, le rôle de la mémoire de reconnaissance serait de s'assurer que des nouveaux objets et événements sont reconnus car ils sont insérés dans des catégories mentales déjà existantes dans la mémoire.

A ce sujet il faut souligner que les définitions habituelles de la mémoire soulignent la faculté qui permet d'encoder, stocker et rappeler des expériences passées : mais cette définition met l'accent sur la capacité de rappel conscient, or de nombreux phénomènes mnésiques sont inconscients et s'expriment de manière implicite. Une définition de la

mémoire plus complète serait, par conséquent, qu'il s'agit de l'ensemble des mécanismes par lesquels une expérience peut modifier un comportement ultérieur, ce qui signifie, en termes cognitifs, les différentes fonctions exécutives, de l'attention à l'apprentissage aux décisions.

## 2. Modèles de mémoire ; mémoires déclaratives et procédurales

Jusqu'à ici, il n'y a pas de gros désaccords. Cependant, quand on suit une approche neuroscientifique, le problème de la multiplicité des modèles de la représentation mnésique émerge de manière significative (Voire Emmanuel Barbeau 2011). A cet égard on peut considérer quatre modèles différents qui ne concernent pas seulement les réseaux neuronaux impliqués mais aussi la clinique des troubles de la mémoire :

1. Un modèle influent de la mémoire, celui de Squire, (Squire 2004), met en avant une dichotomie anatomo-fonctionnelle assez franche entre la mémoire déclarative et les autres mémoires. Dans cette conception, la mémoire déclarative dépend des structures temporales internes et, réciproquement, ces dernières ne sont impliquées que dans la mémoire déclarative.

2. Le modèle anatomo-fonctionnel, soutenu avant tout par Mortimer Mishkin (Mishkin et al., 1984), propose que les différentes structures du lobe temporal interne contribuent de manière différente à la mémoire déclarative. L'hippocampe, en particulier, supporterait la mémoire contextualisée (mémoire épisodique et spatiale), alors que le cortex périrhinal supporterait la mémoire décontextualisée (mémoire sémantique et mémoire de reconnaissance basée sur la familiarité). Ce modèle, à l'opposé de celui de Squire, propose une relative ségrégation anatomo-fonctionnelle des différentes structures temporales internes, prédisant ainsi qu'il pourrait exister différentes « amnésies » ou formes de troubles de la mémoire en fonction de la lésion.

3. Selon le modèle représentationnel hiérarchique : (Murray, Bussey & Saskida 2007) les voies ventrales, dorsales et les structures temporales internes forment des systèmes représentationnels et hiérarchiques dans la mesure où les structures postérieures de ces systèmes traiteraient des représentations simples, les structures intermédiaires des représentations plus élaborées et les structures les plus antérieures (correspondant aux structures temporales internes) les représentations les plus complexes.

4. Le réseau cortical/sous-cortical. Malgré ce qu'en disent le modèle anatomo-fonctionnel et le modèle représentationnel hiérarchique, nous ne pouvons pas minimiser le fait que d'autres formations jouent un rôle important dans la mémoire comme le cervelet et les ganglions de la base, qui contrôlent des activités cognitives telles que les mémoires procédurales, les composantes motivationnelles de l'apprentissage et l'exécution des actions motrices. Pour ce qui concerne les relations fonctionnelles entre le cortex et les structures sous-corticales Philip Lieberman (2002) souligne comment la sélection naturelle darwinienne et des événements aléatoires ont modifié les mécanismes cérébraux dont la fonction originale était le contrôle moteur. Ce processus a produit des circuits neuronaux qui règlent les actes cognitifs, y compris le langage, en reliant l'activité du cortex préfrontal

avec d'autres aires corticales à travers des structures sous-corticales communes, telles que les ganglions de la base. Dans l'espèce humaine, la coexistence des ganglions de la base et d'un large cortex frontal a conduit à une différenciation progressive entre fonctions cognitives explicites (survenant principalement au niveau conscient) et implicites (souvent réalisées au niveau automatique et inconscient): mais nous savons aujourd'hui que de nombreuses mémoires déclaratives, répétées et récurrentes dans le temps, peuvent être « procéduralisées », c'est-à-dire transférées dans un autre registre qui, en général, dépend des ganglions de la base.

Il faut souligner à ce propos un autre point en faveur d'une absence évidente de démarcations entre les mémoires procédurales et déclaratives. Un cas exemplaire est celui du nouveau-né qui, au début, est marqué par des mouvements maternels qui exercent un effet profond sur les structures cognitives. Les temps des mouvements maternels et de ceux que l'enfant va progressivement mettre en place (avant et après) et leurs conséquences (liens de causes et d'effets) sont à la base des catégories temporelles et causales des activités cognitives telles que les structures linguistiques et donc à la base d'une pluralité de fonctions cognitives. Nous avons déjà pris en considération qu'il n'y a pas qu'un seul système de mémoire à long terme - celle explicite et verbalisée - mais qu'existe aussi une mémoire souterraine, inconsciente, implicite, non verbalisée. Cette mémoire implicite est la seule mémoire qui se développe tôt, elle est déjà présente et active dans les dernières semaines de gestation et est la seule mémoire disponible pour le nouveau-né au cours de ses deux premières années de vie : sa dimension procédurale et émotionnelle-affective permet au bébé de stocker ses premières expériences. Cette mémoire implicite peut être considérée comme la première forme de représentation ou comme la fonction inconsciente de la pensée à "l'état naissant". La transition « souple » entre mémoires procédurales et déclaratives est une des caractéristiques du développement cognitif chez l'enfant, un point qui renforce aussi l'importance de la motricité dans les processus représentatifs mentaux (Oliverio 2007).

## 3. Les ganglions de la base : des structures aux fonctions multiples

Le développement du cortex préfrontal a donc conduit à une restructuration des réseaux cognitifs dans lesquels les deux systèmes (cortex préfrontal et ganglions de la base) peuvent être activés en parallèle. Le striatum, qui chez les mammifères est l'une des structures des noyaux gris centraux en charge des fonctions cognitives implicites, est également impliqué dans des fonctions explicites et interagit avec le cortex préfrontal dans un certain nombre de tâches liées à la production de nouvelles réponses et des stratégies novatrices (Oliverio 2008). Par exemple, le striatum ventral (noyau accumbens) est impliqué dans la transition d'une stratégie cognitive à l'autre en fonction des besoins du moment et de la situation de l'environnement (Ferretti et al. 2010). En outre, dans certaines situations, le striatum peut prendre en charge les fonctions cognitives explicites, telles que les fonctions linguistiques ou, de façon plus générale, les tâches déclaratives.

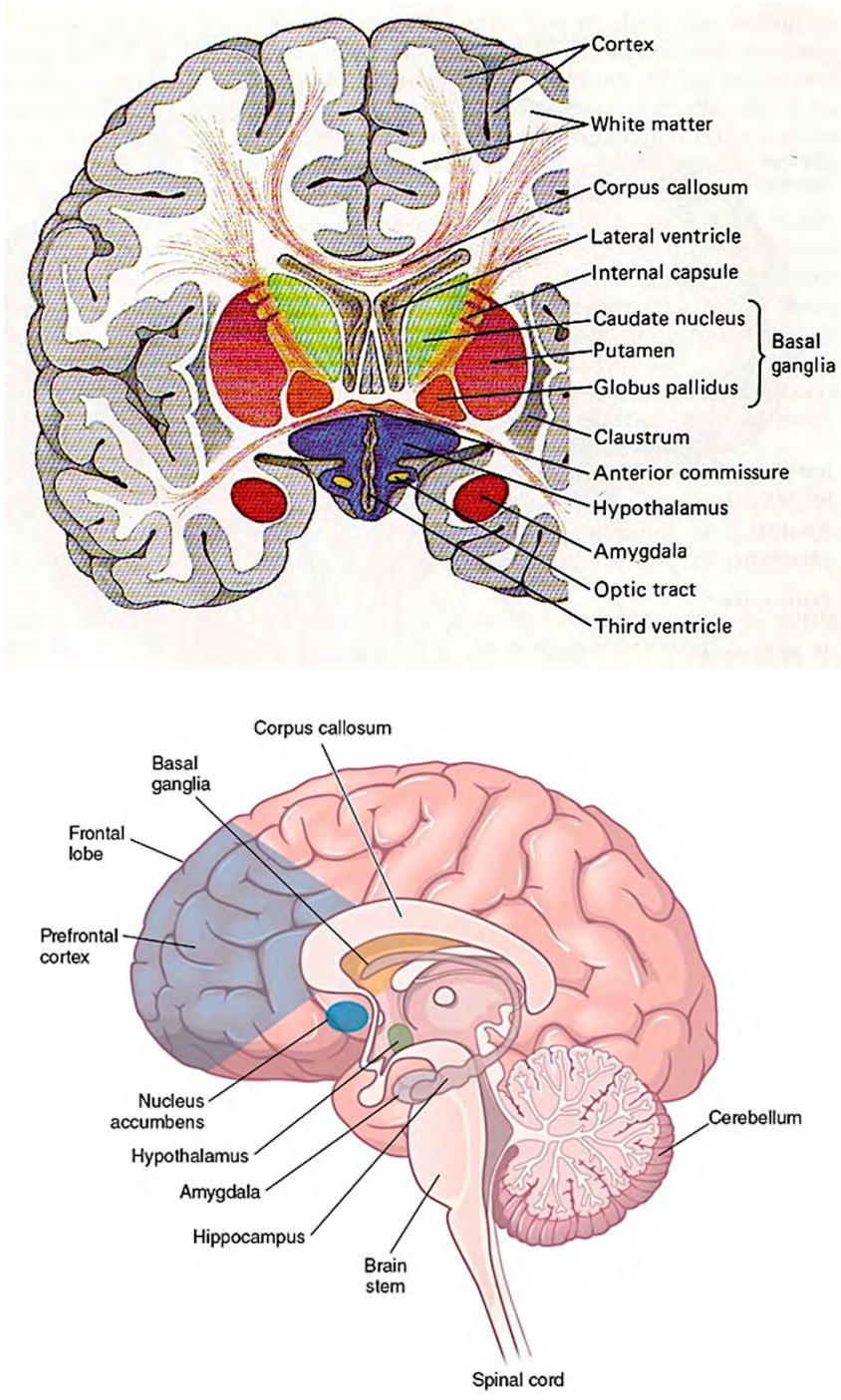



**Figure 1.** Localisation de différentes structures cérébrales (Figures fournies par l'auteur)

Le rôle des ganglions de la base dans les mémoires non procédurales et déclaratives ressort d'une série d'études conduites sur les animaux et sur les êtres humains. Ces études sont basées sur des évaluations de la mémoire spatiale, sur les réponses à des stimuli nouveaux ou connus, et sur des formes de navigation qui répondent à des références égocentriques ou allocentriques. De l'avis de plusieurs auteurs (voir Buzsáki et Moser 2013) notre mémoire sémantique dériverait de nos capacités de navigation allocentrique tandis que notre mémoire épisodique (celle de notre « parcours » autobiographique) dériverait de nos capacités de navigation egocentrique (celle des parcours dans l'espace). Les mêmes réseaux de neurones pourraient donc fournir des algorithmes capables de traiter les deux types de voyage, spatiaux et temporels, une interprétation qui indique l'unité des processus représentatifs de la mémoire, dans le prolongement de ce qui se passe chez le nouveau-né qui extrait des mouvements les notions de temps et d'espace des processus mentaux.

En ce qui concerne le striatum ventral, spécifiquement le noyau accumbens, cette structure peut participer aux formes de navigation qui ne sont généralement pas associées à la fonction de l'hippocampe, telles que les réponses de navigation dépendant des références égocentriques (par opposition à celles allocentriques) (De Leonibus et al., 2005, Floresco et al., 2006). De cette manière, le noyau accumbens peut être similaire au striatum dorsal, en ce que le striatum dorsal aussi a été impliqué dans l'apprentissage égocentrique (Abraham et al., 1983 ; Cook et Kesner, 1988 ; Potegal, 1969). Un rôle comparable pour le striatum dorsal et le nucleus accumbens dans l'apprentissage égocentrique a été démontré dans une tâche de déplacement d'objet (De Leonibus et al., 2005). Ensemble, plusieurs résultats indiquent que le striatum dorso-latéral est nécessaire seulement pour la « réponse » égocentrique de localisation, tandis que le nucleus accumbens peut participer à la fois aux formes de localisation « spatiale » allocentrique et de « réponse » égocentrique (De Leonibus et al., 2005, Ferretti et al., 2015 ; Rinaldi et al. 2012)

Les fonctions mnémoniques des ganglions de la base qui ont été étudiées principalement chez les animaux inférieurs ont également été démontrées chez les humains. Par exemple, la mémoire des tâches procédurales telles que le labyrinthe radial a été associée à une plus grande activation du striatum dorsal (Bohbot et al., 2004, Iaria et al., 2003 ; Konishi et al., 2013) et à une augmentation de la connectivité corticostriatale (Horga et al., 2014). De même, dans des études de neuroimagerie chez l'homme, les sous-régions dorsales striatales ont été impliquées dans l'apprentissage instrumental (Brovelli et al., 2011 ; Tricomi et al., 2009). Des preuves électrophysiologiques chez les humains indiquent également un rôle important du noyau accumbens dans l'apprentissage stimulus-réponse (Cohen et al., 2009). Ainsi, les ganglions de la base, en dehors de leur rôle classique dans la fonction motrice, sont impliqués dans une variété d'apprentissages et des processus de mémoire.

Les ganglions de la base sont un lien au travers duquel les régions corticales et sous-corticales qui exercent des fonctions exécutives, limbiques, sensorielles et motivationnelles peuvent influer la motricité et générer ainsi des stratégies comportementales pour atteindre des résultats favorables. Il faut souligner que bien que les ganglions de la base fonctionnent

souvent de concert avec des systèmes de mémoire d'ordre supérieur, la contribution des ganglions de la base à une situation d'apprentissage donnée reste unique. Dans certaines situations d'apprentissage, les lésions des ganglions de la base produisent des altérations plus graves (par exemple, Fouquet et al., 2013, Pistell et al., 2009) et dans d'autres situations, des déficiences moins graves (Devan et White, 1999).

En outre, conjointement avec l'existence d'interactions compétitives entre les systèmes de mémoire (pour une discussion récente, voir Poldrack et Packard, 2003), les lésions des ganglions de la base peuvent être associées à une fonction accrue d'autres systèmes de mémoire (Bradfield et Balleine, 2013 ; Kosaki et al., 2015). De même, les lésions ou l'engagement réduit d'autres systèmes de la mémoire (par exemple, l'hippocampe) ont été associés à une augmentation des fonctions de l'apprentissage et de la mémoire au niveau des ganglions de la base (Packard et Goodman, 2013). Ainsi, les ganglions de la base peuvent intervenir sur les fonctions dissociables de la mémoire qui, selon les exigences de la tâche, peuvent soit rivaliser ou coopérer avec des régions alternatives du cerveau pour le contrôle de la mémoire et de l'apprentissage.

## 4. Mémoire : simple archivage ou perpétuelle restructuration ?

La multiplicité des structures impliquées dans la mémoire et dans ses fonctions représentationnelles nie l'existence d'un « centre » de la mémoire et nous parle de réseaux complexes et d'un dynamisme qui est aussi évident dans les processus de reconsolidation à travers lesquels la représentation d'une expérience est mise à jour (Squire et Oliverio 1991).

La mémoire est souvent présentée comme une archive dans laquelle des expériences sont déposées : une archive durable qui contient les souvenirs dits à long terme, consolidés et stabilisés à partir de la forme à court terme ou « de travail ». Cette conception à deux niveaux de la mémoire a été établie suivant les théories de Donald O. Hebb (1904-1985) qui a d'abord soutenu que les souvenirs à court terme dépendaient des altérations électriques d'un circuit nerveux et à long terme des changements structurels. Conformément aux théories hebbiennes, les neuroscientifiques ont montré que la phase de consolidation de la mémoire est fragile et que de nombreux traitements physiques, comme un électrochoc administré immédiatement après une expérience ou l'administration d'antibiotiques qui bloquent la synthèse des protéines, et donc la production de nouvelles synapses, empêchent le passage de la mémoire à court terme vers la mémoire à long terme: mais une fois la consolidation effectuée, rien ne peut troubler les souvenirs stables, sauf un lent et inexorable processus d'oubli, plus évident dans les années de vieillesse.

Autrefois, la psychobiologie de la mémoire était donc basée sur le principe de stabilité des souvenirs, codifiés sous forme stable dans les circuits cérébraux : mais ce principe a été contesté il y a quelques années par les études d'une psychologue, Elisabeth Loftus (2005), qui a étudié la mémoire autobiographique et a démontré que les souvenirs dépendent d'un travail complexe de remaniement de "fragments" liés à différents niveaux autobiographiques. L'immutabilité et la stabilité de la mémoire à long terme seraient donc un mythe, et le processus de consolidation ne garantirait pas une cohérence des expériences

codifiées sous une forme « stable ». Ce nouveau concept est lié à une série de recherches, principalement conduites par Karin Nader, George Schafe et Joseph LeDoux (2000) et par Susan Sara (2010) : les résultats de ces études montrent qu'en plus de la consolidation, il y a aussi la reconsolidation, caractérisée par la restructuration des expériences antérieures, ce qui implique une nouvelle représentation cognitive. Le terme de reconsolidation indique donc que le fait de se souvenir de quelque chose rend la trace de la mémoire flexible, sujette aux re-manipulations et à la restructuration. Par conséquent, les représentations de la mémoire, plutôt que d'être stables, sont dynamiques.

## 5. Perspectives de recherche

Pour revenir au problème initial, c'est-à-dire la représentation d'un concept à travers l'activation globale d'un réseau neuronal envisagée par Dehaene et al. (1987), à l'avenir, nous serons en mesure d'utiliser de nouvelles techniques capables de marquer les synapses et les réseaux activés par une nouvelle expérience ou l'appel à la mémoire d'un événement. De plus en plus de preuves indiquent l'importance des épines dendritiques dans la formation et l'attribution des mémoires, et les modifications du nombre des épines et de la physiologie sont associées à la mémoire et aux troubles cognitifs. Les changements de l'activité des sous-ensembles de synapses sont considérés comme cruciaux pour l'établissement de la mémoire. Une nouvelle méthode pour tester directement cette hypothèse, en contrôlant sélectivement l'activité des épines potentialisées, a été proposée, basée sur une approche hybride ARN / protéine pour réguler l'expression d'un canal membranaire sensible à la lumière au niveau des synapses activées (Gobbo et al. 2017). Cette technique permet de marquer les épines potentialisées à la suite de l'encodage d'un nouveau contexte dans l'hippocampe. Cette approche a été utilisée avec succès pour cartographier les synapses potentialisées dans le cerveau en fonction de la mémoire et rendra possible de réactiver un neurone seulement aux synapses précédemment activées, ce qui est critique dans l'étude des processus de la mémoire.

En conclusion, la connaissance de la pluralité des circuits à la base de la mémoire, de leur plasticité et restructuration sur la base de nouvelles expériences nous fournissent une image de plus en plus complète des fonctions représentatives de la mémoire et de ses bases neurobiologiques.

**Références**

Abraham, L., Potegal, M., Miller, S., Evidence for caudate nucleus involvement in an egocentric spatial task—return from passive transport. Physiological Psychology. 11, 11-17, 1983.

Barbeau, E. Les modèles de la mémoire: approches anatomo-fonctionnelle et représentationnelle-hiérarchique. Revue de neuropsychologie, 3, 104-111 2011.

Bohbot, V.D., Iaria, G., Petrides, M., Hippocampal function and spatial memory: evidence from functional neuroimaging in healthy participants and performance of patients with medial temporal lobe resections. Neuropsychology. 18, 418-425, 2004.

Bovet D., Fessard A., Florés C. Frijda N.H., Inhelder B., Milner B. et Piaget J. La mémoire, Presses Universitaires de France, Paris 1970, pp. 301.

Bradfield, L.A., Balleine, B.W., 2013. Hierarchical and binary associations compete for behavioral control during instrumental biconditional discrimination. Journal of Experimental Psychology and Animal Behavioral Processes 39, 2-13, 2013.

Brovelli, A., Nazarian, B., Meunier, M., Boussaoud, D., Differential roles of caudate nucleus and putamen during instrumental learning. Neuroimage. 57, 1580-1590, 2011.

Buzsáki G. et Moser E. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. Nature Neuroscience 16, 130-138 2013.

Cohen, M.X., Axmacher, N., Lenartz, D., Elger, C.E., Sturm, V., Schlaepfer, T.E., Neuroelectric signatures of reward learning and decision-making in the human nucleus accumbens. Neuropsychopharmacology. 34, 1649-1658, 2009.

Cook, D., Kesner, R.P., Caudate nucleus and memory for egocentric localization. Behavioral and Neural Biololgy 49, 332-343, 1988.

Dehaene S., Changeux J.-P., Nadal J.P. Neural networks that learn temporal sequences by selection. Proceedings of the National Academy of Science, USA 84: 2727-2731, 1987.

De Leonibus E. Oliverio A. Mele A. A study on the role of dorsal striatum and nucleus accumbens in allocentric and egocentric spatial memory consolidation. Learning and Memory, 12, 491-503, 2005.

Devan, B.D., White, N.M., 1999. Parallel information processing in the dorsal striatum: relation to hippocampal function. Journal of Neuroscience, 19, 2789-2798, 1999.

Ferretti, V., Perri, V., Cristofoli, A., Vetere, G., Fragapane, P., Oliverio, A., Mele, A., Phosphorylation of S845 GluA1 AMPA receptors modulates spatial memory and structural plasticity in the ventral striatum. Brain Structure & Function, 220, 2653-2661, 2015.

Ferretti V., Roullet P., Sargolini F., Rinaldi A., Perri V., Del Fabbro M., Costantini V.J.A., Annese V., Scesa G. De Stefano M.E., Oliverio A., Mele A. Ventral striatal plasticity and spatial memory, Proceedings National Academy of Sciences, 107, 7945-7950, 2010.

Ferretti, V., Sargolini, F., Oliverio, A., Mele, A., Roullet, P., 2007. Effects of intra-accumbens NMDA and AMPA receptor antagonists on short-term spatial learning in the Morris water maze task. Behavioral Brain Research 179, 43-49, 2007.

Floresco, S.B., Ghods-Sharifi, S., Vexelman, C., Magyar, O. Dissociable roles for the nucleus accumbens core and shell in regulating set shifting. Journal of Neuroscience 26, 2449-2457, 2006

Fouquet, C., Babayan, B.M., Watilliaux, A., Bontempi, B., Tobin, C., Rondi-Reig, L., Complementary roles of the hippocampus and the dorsomedial striatum during spatial and sequence-based navigation behavior. PLoS One. 8, 1-11, 2013.

Gobbo F., Marchetti L., Jacob A., Pinto B., Binini N., Pecoraro Bisogni F., Alia C., Luin S., Caleo M., Fellin T., Cancedda L. et Cattaneo A. Activity-dependent expression of Channelrhodopsin at neuronal synapses. Nature Communications 8, 1629-1643, 2017. DOI: 10.1038/s41467-017-01699-7.

Horga, G., Maia, T.V., Marsh, R., Hao, X., Xu, D., Duan, Y., Tau, G. Z., Graniello, B., Wang, Z., Kangarlu, A., Martinez, D., Packard, M.G., Peterson, B.S., 2014. Changes in corticostriatal connectivity during reinforcement learning in humans. Human Brain Mapping. 36, 793-803, 2014.

Iaria, G., Petrides, M., Dagher, A., Pike, B., Bohbot, V.D., Cognitive strategies dependent on the hippocampus and caudate nucleus in human navigation: variability and change with practice. Journal of Neuroscience 23, 5945-5952, 2003.

J. Goodman and M.G. Packard Memory Systems of the Basal Ganglia in H. Steiner and K. Tseng (Eds): Handbook of Basal Ganglia Structure and Function, Second edition. Elsevier, p 725-735, 2017.
DOI: http://dx.doi.org/10.1016/B978-0-12-802206-1.00035-0

Konishi, K., Etchamendy, N., Roy, S., Marighetto, A., Rajah, N., Bohbot, V.D., Decreased functional magnetic resonance imaging activity in the hippocampus in favor of the caudate nucleus in older adults tested in a virtual navigation task. Hippocampus. 23, 1005-1014, 2013.

Kosaki, Y., Poulter, S.L., Austen, J.M., McGregor, A., Dorsolateral striatal lesions impair navigation based on landmark-goal vectors but facilitate spatial learning based on a "cognitive map". Learning and Memory 22, 179-191, 2015.

Lieberman P. On the Nature and Evolution of the Neural Bases of Human Language. Yearbook of Physical Anthropology, 45, 36-62, 2002.

Loftus, E. (2005). "Planting misinformation in the human mind: A 30-year investigation of the malleability of memory". Learning & Memory 12 (4): 361–366

Mishkin, M., Malamut, B., Bachevalier, J., Memories and habit: two neural systems. In: Lynch, G., McGaugh, J.L., Weinberger, N. M. (Eds.), Neurobiology of Human Learning and Memory. Guilford Press, New York, pp. 65-77, 1984.

Murray E., Bussey T. J. et Saksida L. M. Visual perception and memory: a new view of medial temporal lobe function in primates and rodents. Annual Review of Neuroscience 30:99-122, 2007.

Nader K., Schafe G., LeDoux J.E., Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval, Nature, 406, 722-6, 2000.

Oliverio A, Der handelnde Geist. Über die Bedeutung motorischer Abläufe für mentale Repräsentationsprozesse. Das Kind 41, 51-63, 2007.

Oliverio A. Brain and creativity. Progress of Theoretical Physics Suppl.173, 66-78, 2008.

Packard, M.G., Goodman, J., Factors that influence the relative use of multiple memory systems. Hippocampus. 23, 1044-1052, 2013.

Pistell, P.J., Nelson, C.M., Miller, M.G., Spangler, E.L., Ingram, D.K., Devan, B.D., Striatal lesions interfere with acquisition of a complex maze task in rats. Behav. Brain Res. 197, 138-143, 2009.

Poldrack, R.A., Packard, M.G., Competition among multiple memory systems: converging evidence from animal and human brain studies. Neuropsychologia. 41, 245-251, 2003.

Potegal, M., Role of the caudate nucleus in spatial orientation of rats. Journal of Comparative and Physiological Psychology, 69, 756-764, 1969.

Rinaldi, A., Oliverio, A., & Mele, A. Spatial memory, plasticity and nucleus accumbens. Reviews in the neurosciences. doi:10.1515/revneuro-2012-0070, 2012

Sara S. J. Reactivation, Retrieval, Replay and Reconsolidation in and Out of Sleep: Connecting the Dots Frontiers in Behavioral Neuroscience; 4, 185-193, 2010

Squire, L. Memory systems of the brain: a brief history and current perspective. Neurobiology of Learning and Memory, 82, 171-177, 2004.

Squire, L. R. et Oliverio, A. Biological memory in P. Corsi (Ed.) Chapters in the history of neuroscience. Oxford University Press, New York p. 338-340, 1991.

Tricomi, E., Balleine, B.W., O'Doherty, J.P., A specific role for posterior dorsolateral striatum in human habit learning. European Journal of Neuroscience 29, 2225-2232, 2009.

# 2

# La Mémoire Humaine et ses substrats cérébraux

**Francis Eustache et Armelle Viard**

Inserm-EPHE-Université de Caen U1077
Neuropsychologie et Imagerie de la Mémoire Humaine

## Abstract

Francis Eustache and Armelle Viard invite us to take an interest in the research on human memory. Francis Eustache recalls how pathological memory disorders have gradually, and very strongly since the 1960s, been a source of information in the development of theoretical models of human memory. He develops an analysis of the genesis and evolution of these models and the concepts which underpin them; he ends with a brief introduction to the MNESIS model that he developed with his team, showing its relationships with the previous models. Armelle Viard deals specifically with the running of "autobiographical" memory and the structures involved in it, in comparison with "semantic" memory. She draws up a table of various experimental works which have been carried out on these themes, in particular using neuroimaging; she presents and discusses the results. She thus deals with how the reconstruction / recovery of a memory is carried out, the parameters which intervene in such processes and the brain regions involved. She ends with an analysis of autobiographical memory disorders in neurodegenerative pathologies such as Alzheimer's and semantic dementia.

Partie A :

# Neuropsychologie de la mémoire humaine

**Francis Eustache**

## 1. Comment définit-on la mémoire ?

### 1.1. Introduction

A partir des années 1960, les syndromes amnésiques ont constitué la source d'inférence principale pour élaborer des modèles théoriques de la mémoire humaine. Ils se caractérisent par une atteinte de la mémoire, d'origine organique, disproportionnée par rapport à d'autres troubles cognitifs. La recherche de dissociations entre capacités mnésiques perturbées et capacités mnésiques préservées devient le grand paradigme de la neuropsychologie de la mémoire (Eustache et Desgranges, 2012 ; Eustache et al, 2018 ; Eustache et Guillery-Girard, 2016).

L'exemple emblématique est la dissociation, mise en évidence chez le patient amnésique H.M. par Brenda Milner, psychologue canadienne, entre l'atteinte de la mémoire à long terme et la préservation de la mémoire à court terme. La dissociation inverse est rapportée par les psychologues britanniques Shallice et Warrington à propos du patient K.F., présentant une préservation de la mémoire à long terme et une réduction de l'empan auditivo-verbal. Cette opposition est confortée par de nombreuses données expérimentales, comme les effets de primauté et de récence qui se traduisent par le rappel préférentiel des mots du début et de la fin d'une liste reflétant respectivement la fonctionnalité de la mémoire à long terme et de la mémoire à court terme.

La notion de mémoire à court terme s'est complexifiée pour aboutir au concept de mémoire de travail à composantes multiples. La mémoire à long terme a également donné lieu à plusieurs fractionnements. Les études de patients amnésiques montrent en effet qu'en dépit de difficultés à acquérir volontairement des informations nouvelles, ils sont capables d'apprentissage ou témoignent par leur comportement de la rétention d'expériences antérieures, souvent à leur insu.

A partir des années 1970, la neuropsychologie de la mémoire connaît ainsi deux évolutions parallèles : l'élaboration du concept de mémoire de travail et le démembrement de la mémoire à long terme.

### 1.2. Mémoire à court terme et mémoire de travail

La mémoire de travail, définie par les psychologues britanniques Baddeley et Hitch en 1974, est un système responsable du traitement et du maintien temporaire des

informations nécessaires à la réalisation d'activités aussi diverses que la compréhension, l'apprentissage et le raisonnement.

Elle est composée de deux sous-systèmes satellites de stockage (la boucle phonologique et le calepin visuo-spatial), coordonnés et supervisés par une composante attentionnelle, l'administrateur central. La boucle phonologique est responsable du stockage d'informations verbales, de leur manipulation et de leur rafraîchissement. Elle est constituée d'un registre phonologique de stockage passif, de capacité limitée et d'un processus d'autorépétition subvocale, la récapitulation articulatoire, permettant le rafraîchissement de l'information et la conversion d'un stimulus présenté visuellement en un code phonologique. Le calepin visuo-spatial est impliqué dans le stockage des informations spatiales et visuelles ainsi que dans la formation et la manipulation des images mentales.

L'administrateur central supervise et coordonne l'information en provenance des systèmes satellites et gère le passage de l'information vers la mémoire à long terme. Il joue un rôle dans la focalisation et le partage de l'attention, dans la sélection des informations en mémoire à long terme, dans la manipulation de ces informations et dans l'intégration en mémoire à long terme des nouvelles informations. Ainsi, la mémoire de travail n'est pas uniquement une voie de passage des entrées sensorielles en mémoire à long terme mais un espace de travail entre les données issues de l'environnement et les connaissances en mémoire à long terme.

Au début des années 2000, Baddeley a postulé l'existence d'un nouveau système temporaire de stockage, le buffer épisodique, chargé du stockage temporaire d'informations intégrées provenant de différentes sources. Il est contrôlé par l'administrateur central, qui récupère ces informations depuis les systèmes de stockage sous la forme de processus conscients, traite ces informations et, si nécessaire, les manipule et les modifie. Ce buffer est épisodique car il stocke des épisodes dans lesquels l'information est intégrée dans l'espace et le temps. Il se rapproche du concept de mémoire épisodique mais en diffère car il s'agit d'un système de stockage temporaire qui peut être préservé chez des amnésiques ayant des troubles de la mémoire épisodique.

### 1.3. La mémoire à long terme

L'une des distinctions les plus importantes concerne la mémoire épisodique et la mémoire sémantique, proposée initialement par le psychologue canadien Tulving en 1972. La mémoire épisodique est définie comme la mémoire des événements personnellement vécus, situés dans leur contexte temporo-spatial d'acquisition. Sa caractéristique fondamentale est de permettre le souvenir conscient d'une expérience antérieure : l'événement lui-même (quoi), le lieu (où) et le moment (quand) où il s'est produit. En plus de l'exactitude du souvenir de l'événement rappelé, ce qui caractérise cette mémoire est l'expérience subjective, l'impression de revivre l'événement. La récupération d'un souvenir en mémoire épisodique implique ainsi un « voyage mental dans le temps » associé

à la conscience autonoétique (ou conscience de soi). Cette définition met l'accent sur la conjonction de trois idées : l'identité (ou self), la conscience autonoétique et le temps subjectif. La situation du patient amnésique K.C., victime de plusieurs traumatismes crâniens, décrit par Tulving, permet de comprendre à quel point l'absence de mémoire épisodique et de conscience autonoétique engendre une impression de vide sans retour vers le passé ni projection dans le futur.

La mémoire sémantique a d'abord été définie comme la mémoire des mots, des concepts, des « connaissances du monde », indépendamment de leur contexte d'acquisition. Elle s'est vue ensuite attribuer la notion de conscience noétique ou conscience de l'existence du monde, des objets, des événements et de diverses régularités. La mémoire sémantique permet ainsi une conduite introspective sur le monde. Le concept comprend également les connaissances générales sur soi (ou sémantique personnelle).

L'opposition entre mémoire déclarative et mémoire procédurale, proposée dans les années 1980 par le neuroscientifique américain Squire et ses collaborateurs, se situe à un autre niveau et correspond à la distinction, plus ancienne, entre mémoire et habitude. L'information stockée en mémoire déclarative est facilement verbalisable et accessible à la conscience. Les représentations peuvent être générales (sémantiques) ou spécifiques (épisodiques). La mémoire procédurale permet d'acquérir des habiletés progressivement, au fil de nombreux essais, de les stocker et de les restituer sans faire référence aux expériences antérieures. Elle s'exprime dans l'activité du sujet et ses contenus sont difficiles à verbaliser. La mémoire procédurale est une mémoire automatique et difficilement accessible à la conscience.

Une autre distinction, opposant mémoire explicite et mémoire implicite, a été proposée dans sa version moderne par le psychologue américain Schacter, dans les années 1980. Certains résultats ont conduit à la proposition, par Tulving et Schacter en 1990, d'un nouveau système de mémoire, le système de représentations perceptives. La mémoire implicite est mise en jeu quand des expériences préalables modifient la performance dans une tâche qui ne requiert pas le rappel conscient de ces expériences. Par exemple, le fait de voir une image une première fois facilite l'identification ultérieure de cette image, y compris si elle est présentée sous une forme dégradée, et ceci sans que le sujet ait conscience de faire appel à sa mémoire. Au contraire, la mémoire explicite fait référence aux situations dans lesquelles un sujet rappelle volontairement des informations stockées en mémoire.

### 1.4. Des concepts aux modèles

Malgré la pléthore de qualificatifs associés au mot « mémoire », seules certaines composantes de la mémoire ont accédé au statut de système de mémoire, chacun ayant sa relative indépendance, ses caractéristiques propres et ses règles de fonctionnement. En accord avec la proposition de Tulving, nous retenons une organisation de la mémoire

formée de cinq systèmes : la mémoire de travail, la mémoire procédurale, la mémoire perceptive, la mémoire sémantique et la mémoire épisodique. Nous reprenons les terminologies proposées par Tulving et par d'autres auteurs, notamment Squire et Baddeley qui ont laissé leur empreinte sur ces différents concepts.

MNESIS (pour Modèle NEoStuctural InterSystémique ; Eustache et al, 2016 ; voir figure 1) intègre les éléments les plus robustes des conceptions multi-systèmes proposées par ces auteurs, tout en spécifiant davantage les relations entre les systèmes et les apports récents de la neuroimagerie. Les trois systèmes de représentation à long terme (mémoire perceptive, mémoire sémantique, mémoire épisodique) sont présentés en respectant l'organisation hiérarchique proposée par Tulving. Les traces mnésiques transitent par les mémoires perceptives avant d'accéder, éventuellement, au statut de représentations sémantiques (dans la mémoire sémantique) et, éventuellement, à celui de souvenirs (dans la mémoire épisodique). Cette organisation rend compte du fait que des patients amnésiques (avec un trouble de la mémoire épisodique) continuent de former des connaissances sémantiques.

A gauche de ces trois systèmes de représentation figurent deux flèches qui doivent être considérées comme des rétroactions. L'une (allant de la mémoire épisodique à la mémoire sémantique) désigne le processus de sémantisation des souvenirs et insiste sur le fait que les souvenirs font l'objet d'un processus de sémantisation au fil du temps. La seconde flèche (de la mémoire épisodique à la mémoire perceptive) met l'accent sur les phénomènes de reviviscence, conscients et inconscients, indispensables à la consolidation mnésique. Il s'agit de processus très divers allant de la ré-évocation de la scène initiale émaillée de détails sensoriels à des mécanismes moins contrôlés se produisant pendant des rêveries ou certains stades de sommeil. Ces rétroactions soulignent le caractère dynamique et reconstructif de la mémoire et leur corollaire, les transformations de la trace mnésique et la formation possible de faux souvenirs.

Au centre de la figure se trouve la mémoire de travail avec, d'une part, les composantes du modèle « classique » de Baddeley : administrateur central, boucle phonologique, calepin visuo-spatial et, d'autre part, le buffer épisodique. Le buffer épisodique place la mémoire au centre du psychisme et doit être rapproché de la notion de conscience de soi, qui donne au sujet une impression subjective de soi dans le temps et qui préside au sentiment d'intégrité et de continuité.

A droite du modèle est représentée la mémoire procédurale intégrant les supports d'habiletés motrices, perceptivo-motrices et cognitives. Les interactions entre ce système d'action et les systèmes de représentation sont matérialisées par les différentes flèches. Les liens avec la mémoire perceptive sont privilégiés pour la mémoire procédurale perceptivo-motrice, et avec les systèmes déclaratifs pour la mémoire procédurale cognitive. Dans tous les cas, les interactions avec les systèmes de représentation (y compris la mémoire de travail) sont particulièrement importantes lors de la phase d'apprentissage procédural.

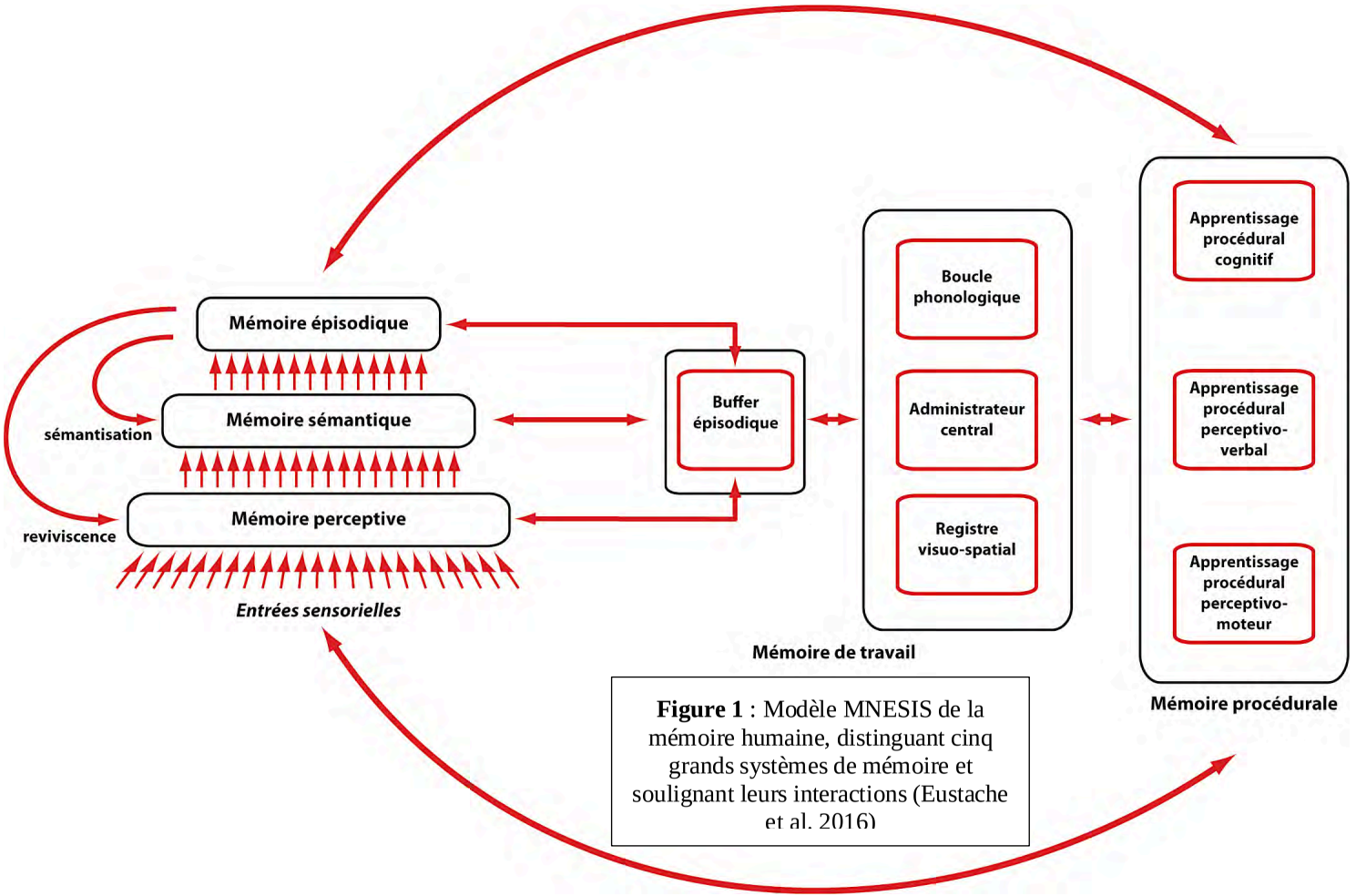



**Figure 1** : Modèle MNESIS de la mémoire humaine, distinguant cinq grands systèmes de mémoire et soulignant leurs interactions (Eustache et al. 2016)

Partie B :

# Neuroimagerie de la mémoire humaine

**Armelle Viard**

## 2. Mémoire autobiographique

### 2.1. Définition

La mémoire autobiographique est un domaine d'investigation qui a donné lieu à des travaux très variés. La mémoire autobiographique permet de voyager dans le temps subjectif et donne l'impression de revivre mentalement les détails phénoménologiques des événements du passé. Toutefois, les représentations de la mémoire autobiographique ne sont pas uniquement épisodiques : nombre d'entre elles sont sémantiques. Au total, cette mémoire correspond à la mémoire à long terme qui permet d'encoder, de stocker et de récupérer des informations dont le Self est le sujet central : il s'agit, en raccourci, de la « mémoire du Self ».

### 2.2. Modèle de Conway

Un modèle d'organisation de la mémoire autobiographique a été proposé par Conway et collaborateurs (Conway et Pleydell-Pearce, 2000). Conway considère que les souvenirs autobiographiques mettent en jeu des processus mnésiques complexes et reconstructifs : le souvenir est construit de façon dynamique à partir de quatre types de représentations organisées hiérarchiquement, du plus général au plus spécifique : le schéma historique personnel, les périodes de vie, les événements généraux et les détails spécifiques. Le *schéma historique personnel*, les *périodes de vie* et les *événements généraux* constituent les connaissances autobiographiques de base (« *autobiographical knowledge base* ») et les *détails perceptivo-sensoriels* d'événements spécifiques sont regroupés dans ce que les auteurs appellent le système de mémoire épisodique.

Plus spécifiquement, le schéma historique personnel (« life-story schema ») représente le niveau le plus global et correspond à des généralisations de thèmes dominants ou de chapitres de vie (par exemple, « ma vie en tant que femme anglaise du 21ème siècle » ou « ma carrière en tant que professeur dans une université américaine »). Les périodes de vie (« life-time periods ») constituent un niveau abstrait et étendu dans le temps, mesuré en années ou en décennies (e.g., la période de « l'école maternelle » se caractérise par un contenu thématique - les images génériques des professeurs, des salles de classe - et par une durée avec un début et une fin spécifiés). Les événements généraux (« general events ») sont plus spécifiques et en même temps plus hétérogènes que les périodes de vie et se mesurent en jours, en semaines ou en mois (e.g. « les week-ends à la campagne », « mon week-end en Italie »). Les détails perceptivo-sensoriels d'événements spécifiques (« event specific knowledge »), mesurés en secondes, en minutes ou en heures, correspondent au registre phénoménologique de la trace mnésique (images, sentiments, odeurs…). Ces détails sont rapidement oubliés, à moins d'être répétés ou associés à des connaissances générales.

À chaque étape du processus mnésique (encodage, stockage, récupération), ces quatre formes de représentation sont organisées dans la mémoire à long terme sous l'influence du modèle d'intégrité personnelle du sujet (ou « *self* »). À l'encodage, les événements sont retenus s'ils sont en accord avec les buts actuels du self (Conway et Pleydell-Pearce, 2000). Lors de la récupération, la reconstruction du souvenir est soumise à un processus central de contrôle, modulé par le self. Le processus de récupération est cyclique et dépend de l'administrateur central de la mémoire de travail.

## 2.3. Réseau cérébral

### *2.3.1. Régions sous-tendant la reconstruction de souvenir*

La récupération de souvenirs autobiographiques peut se faire selon deux modes : 1/ la récupération *directe* qui se produit de façon automatique et involontaire et 2/ la récupération générative qui est intentionnelle et contrôlée. Le processus de récupération directe (« *direct retrieval* ») consiste en l'accès involontaire aux détails spécifiques se faisant souvent en réponse à des indices très particuliers (une odeur, une saveur, une mélodie). Ce type de récupération n'est pas contrôlé par des processus exécutifs superviseurs. Le processus de récupération *générative* (« *generative retrieval* ») consiste en un processus complexe qui exige que le sujet se trouve dans un mode de récupération (« *retrieval mode* » caractérise l'état dans lequel le sujet se trouve quand il cherche à se rappeler volontairement un événement personnel). Ce processus de reconstruction se décompose en trois phases : *l'élaboration* d'indices qui permettra d'initier un contexte de recherche, la *recherche* d'un événement dans ce contexte, la *vérification* du résultat de l'étape précédente. Si le résultat n'est pas satisfaisant, le processus recommencera en utilisant l'information issue de la deuxième étape comme un nouvel indice.

Ce modèle met l'accent sur le rôle prédominant du lobe frontal dans la récupération autobiographique. Le contrôle est modulé par les buts actuels du sujet. Un souvenir récupéré est donc un profil d'activation particulier et stable des quatre types de connaissances, maintenu de façon transitoire en mémoire de travail. Ce processus rend compte de la déformation des souvenirs puisque le souvenir est encodé puis reconstruit et interprété à chaque fois en fonction du self actuel du sujet. Ainsi, l'événement encodé et récupéré est une interprétation propre au sujet.

Dans une étude réalisée en électroencéphalographie (EEG), Conway et collaborateurs ont proposé à des participants une tâche qui consistait à rechercher des souvenirs autobiographiques à partir de mots indices (phase de récupération de 5 secondes), puis à maintenir chaque souvenir en mémoire (phase de maintien de 5 secondes) et enfin à se préparer pour la présentation du mot indice suivant (phase de préparation de 5 secondes). Les résultats de cette étude EEG sont résumés dans la Table 1. Conway a proposé une interprétation de ces résultats : dans la phase de récupération, le déplacement des activations des régions antérieures aux régions postérieures reflèterait l'initiation (lobe frontal gauche), puis le résultat du processus de contrôle lorsque le souvenir est reconstruit (régions postérieures). Les régions occipitales semblent particulièrement impliquées pour les souvenirs vivaces, importants ou spécifiques. Par ailleurs, l'activation plus faible du lobe frontal droit traduirait l'accès à la connaissance de base (sémantique personnelle). Dans la phase de maintien, les activations latéralisées à droite indiqueraient les régions de stockage des connaissances autobiographiques. La phase de préparation est considérée comme une phase d'inhibition du processus de récupération. Ces résultats confortent le modèle de la mémoire autobiographique de Conway.

**Table 1** : Principaux résultats du décours temporel des activations cérébrales au cours d'une tâche d'évocation de souvenirs autobiographiques.

| | |
|---|---|
| Phase de récupération | Hémisphère gauche (lobe **frontal** et **temporal** antérieur) et régions postérieures bilatérales (**temporales** et **occipitales**) |
| Phase de maintien | Hémisphère droit (lobe **frontal** et régions **postérieures** temporales et occipitales) |
| Phase de préparation | Hémisphère droit (régions **pariétales**)<br>Désactivation des régions **temporales** et **occipitales** |

Une méta-analyse de données issues d'études en imagerie par résonance magnétique fonctionnelle (IRMf) a permis d'établir la localisation précise des activations cérébrales observées pendant le rappel autobiographique. A l'inverse de l'EEG, l'IRMf a une haute résolution spatiale qui permet de localiser précisément les régions cérébrales activées lors

d'une tâche cognitive particulière. Svoboda et al. (2006) ont analysé les résultats de 24 études en IRMf focalisées sur le rappel en mémoire autobiographique et montrent que la récupération d'un souvenir active un réseau cérébral (« *core network* » ou réseau cérébral principal), latéralisé à gauche et comprenant les cortex préfrontal médian et ventrolatéral, les cortex temporal médian (incluant l'hippocampe et le gyrus parahippocampique) et temporal latéral, les cortex retrosplénial et cingulaire postérieur, la jonction temporo-pariétale et le cervelet. Des régions moins fréquemment retrouvées sont également mentionnées.

Parmi ces études en IRMf, certaines ont permis de préciser la localisation cérébrale sous-tendant une étape particulière du rappel autobiographique, et ce de manière plus précise que les études en EEG. Steinvorth et al. (2006) ont effectué une étude en IRMf en demandant aux participants de récupérer des souvenirs autobiographiques. Les auteurs ont comparé la phase initiale de recherche des souvenirs à la phase de réminiscence et ont observé l'activation du cortex préfrontal ventro-latéral (aires de cérébrales Brodmann 44, 45, 47). Cette région serait donc spécifiquement activée lors de la phase initiale de recherche d'un souvenir. Cabeza et al. (2004) se sont plutôt intéressés aux régions activées lorsqu'on évoque un souvenir propre à soi par rapport à un événement non personnel. Ils ont demandé à des étudiants de prendre des photos d'endroits précis du campus universitaire et ont comparé ces photos à celles prises par un autre participant. Les auteurs ont montré que les photos prises par soi-même activaient plus le cortex préfrontal médian (aires de Brodmann 10, 32) par rapport à des photos prises par un autre individu. Cette région serait donc spécialement dédiée au traitement d'informations relatives à soi.

### *2.3.2. Régions influencées par la qualité des souvenirs : émotion*

D'autres études en IRMf ont évalué le rôle de la qualité des souvenirs rappelés et les régions cérébrales associées. L'émotion est une caractéristique phénoménologique importante des souvenirs autobiographiques persistants et vivaces. En effet, une étroite relation fonctionnelle existe entre mémoire et émotion comme le soulignent des travaux montrant un rappel préférentiel de souvenirs autobiographiques émotionnels. Les souvenirs émotionnels perdurent et gardent leur vivacité, ce qui semble manquer aux autres souvenirs, et sont mieux mémorisés que des événements neutres.

Au plan anatomique, le complexe amygdalien joue un rôle important dans le traitement d'informations émotionnelles, comme le confirment des études en neuroimagerie. Cette structure est connectée avec des régions fronto-temorales impliquées dans la mémoire autobiographique. Certaines études d'IRMf ont comparé le rappel de souvenirs émotionnels à celui de souvenirs sémantiques (Greenberg et al., 2005) ou moins émotionnels (Viard et al., 2010) et ont confirmé l'implication de l'amygdale. Grâce à des analyses de connectivité fonctionnelle, ces études ont également montré l'existence d'une co-activation entre l'amygdale et l'hippocampe pendant le rappel de souvenirs émotionnels.

### *2.3.3. Régions influencées par la qualité des souvenirs : imagerie visuelle mentale*

L'imagerie mentale visuelle joue également un rôle important dans la récupération en mémoire autobiographique et serait liée à l'épisodicité du souvenir : l'imagerie visuelle augmente le rappel de détails spécifiques et l'expérience subjective du souvenir. L'imagerie visuelle serait le meilleur facteur de spécificité par rapport à d'autres indices (olfactifs, tactiles, auditifs, moteurs). L'imagerie est un indice de récupération efficace et économique contenant des informations génériques sur les principaux personnages, le lieu et l'organisation temporelle de l'événement vécu, ces indices pouvant être utilisés lors de la récupération. La quantité de détails phénoménologiques présents lors du rappel autobiographique pourrait refléter son exactitude (Conway and Pleydell-Pearce, 2000) et distinguer les événements vécus des événements imaginés.

Gardini et al. (2006) se sont intéressés au rôle de l'imagerie visuelle mentale dans le rappel autobiographique. Suite à la présentation d'un mot cible (e.g., voiture), les participants devaient générer une image visuelle mentale en rapport avec le mot présenté (e.g., ma voiture) ou dans un contexte dépersonnalisé (e.g., une voiture). Les auteurs ont montré que la génération d'images mentales relatives à soi activait préférentiellement le cunéus, le précuneus et le gyrus parahippocampique par rapport à la génération d'images impersonnelles. Ainsi, ces régions auraient un rôle spécifique dans l'imagerie visuelle mentale d'items relatifs à soi.

### *2.3.4. Régions influencées par l'ancienneté du souvenir (récent vs. ancien)*

Différents travaux ont montré que l'effet de l'âge du souvenir est hétérogène selon les systèmes de mémoire (Eustache et al., 1998). La mémoire épisodique est particulièrement sensible à l'effet de l'âge, contrairement à la mémoire sémantique. Deux types de consolidation mnésique sont différenciés dans la littérature : d'une part la ***consolidation à court terme*** qui dure quelques secondes ou quelques minutes et qui permet le passage en mémoire à long terme, et d'autre part la ***consolidation à long terme*** qui porte sur des mois, des années ou des décennies et qui aboutit à un stockage durable des représentations mnésiques. Nous nous intéresserons essentiellement à la consolidation mnésique à long terme et au rôle du lobe temporal interne (ou LTI), qui comprend notamment l'hippocampe, dans le stockage et la récupération des souvenirs. Le premier courant théorique de la consolidation mnésique insiste sur le rôle temporellement limité du LTI dans le stockage et la récupération des événements et des informations, alors que le deuxième courant postule l'existence du rôle permanent du LTI dans la récupération des souvenirs autobiographiques.

### *2.3.4.1. Le modèle de Squire et Alvarez*

Squire et Alvarez (1995) ont développé l'un des premiers modèles de la consolidation à long terme pour expliquer les formes d'amnésie rétrograde avec gradient temporel observées chez des patients atteints d'une lésion du LTI. Cette structure jouerait un rôle à la fois dans l'encodage des informations et dans leur récupération pendant un certain temps. Selon cette conception, le LTI sert à indexer les multiples éléments néocorticaux qui constituent la trace mnésique de l'événement vécu. Lors des évocations successives de l'événement ou pendant le sommeil, la co-activation répétée des différents éléments néocorticaux de la trace mnésique, par l'intermédiaire du LTI, crée et renforce graduellement les interconnections néocorticales. Lorsque la consolidation est complète, les interconnections représentant l'événement sont devenues permanentes. L'évocation du souvenir s'effectue alors indépendamment du LTI. Ce mécanisme permet d'expliquer la préservation des souvenirs anciens en cas d'atteinte du LTI dans les formes d'amnésie rétrograde avec gradient temporel de Ribot (rencontrées dans la maladie d'Alzheimer débutante, voir ci-dessous), la préservation des souvenirs récents en cas d'atteinte néocorticale dans les formes avec gradient temporel inversé (observées dans la démence sémantique) et la perturbation globale des souvenirs lorsque les lésions concernent à la fois le LTI et le néocortex.

Néanmoins, ce modèle repose sur une conception de la mémoire déclarative qui ne prend pas en compte la dissociation entre les composantes épisodique et sémantique dont la pertinence a été récemment réaffirmée dans la littérature (Vargha-Khadem et al., 1997 ; Guillery et al., 2001). Ainsi, Vargha-Khadem et al. (1997) ont rapporté les cas de trois jeunes patients atteints de lésions précoces limitées à l'hippocampe. Ces patients présentaient une amnésie antérograde massive et une amnésie rétrograde autobiographique. Ils étaient incapables de rappeler le moindre événement de leur vie mais leur mémoire sémantique était bonne (connaissances sur le monde, apprentissage didactique, langage), ce qui leur avait permis de suivre un cursus scolaire pratiquement normal. Le modèle de Squire et Alvarez ne permet pas d'expliquer l'amnésie rétrograde épisodique globale observée chez les patients ayant une atteinte circonscrite au LTI, ni leur capacité à acquérir de nouvelles connaissances en mémoire sémantique malgré l'atteinte du LTI. Une approche théorique alternative a été proposée par Nadel et Moscovitch (1997) où le lobe temporal interne joue un rôle permanent dans la récupération des souvenirs autobiographiques épisodiques.

### *2.3.4.2. Modèle du rôle permanent du LTI de Nadel et Moscovitch*

Nadel et ses collaborateurs (Nadel et Moscovitch, 1997) ont proposé un modèle alternatif de consolidation qui tient compte de la dichotomie épisodique/sémantique et de l'observation d'amnésies rétrogrades épisodiques globales chez des patients atteints d'une lésion temporale interne, ainsi que des études en neuroimagerie fonctionnelle de la mémoire autobiographique chez des sujets sains montrant l'absence d'activation différentielle de cette région cérébrale en fonction de l'intervalle de rétention (Gilboa et al.,

2004 ; Viard et al., 2007). Ces observations suggèrent, en effet, que le LTI intervient temporairement dans la récupération des informations sémantiques mais de manière permanente dans la récupération des souvenirs épisodiques. Nadel et collaborateurs proposent que le LTI et le néocortex interagissent en continu. Le stockage et la récupération en mémoire autobiographique épisodique dépendraient du LTI de façon permanente. En effet, lorsque la consolidation « standard » est complète, le LTI continue d'indexer les divers éléments de la trace mnésique dans les différentes régions cérébrales impliquées.

Dans ce contexte, nous avons élaboré une étude en IRMf dans laquelle les participants devaient rappeler des souvenirs autobiographiques en fonction de cinq périodes de vie : 3 périodes anciennes (0-17ans, 18-30 ans, 31 ans-5 dernières années) et 2 périodes récentes (5 dernières années, 12 derniers mois). Les résultats ont montré l'activation d'un réseau cérébral similaire pendant le rappel de souvenirs anciens et récents, comprenant le cortex préfrontal médian, le précunéus, le cortex cingulaire postérieur et l'hippocampe (structure faisant partie du LTI ; figure 2). Ce résultat est donc en faveur du modèle alternatif de consolidation de Nadel et Moscovitch (1997).

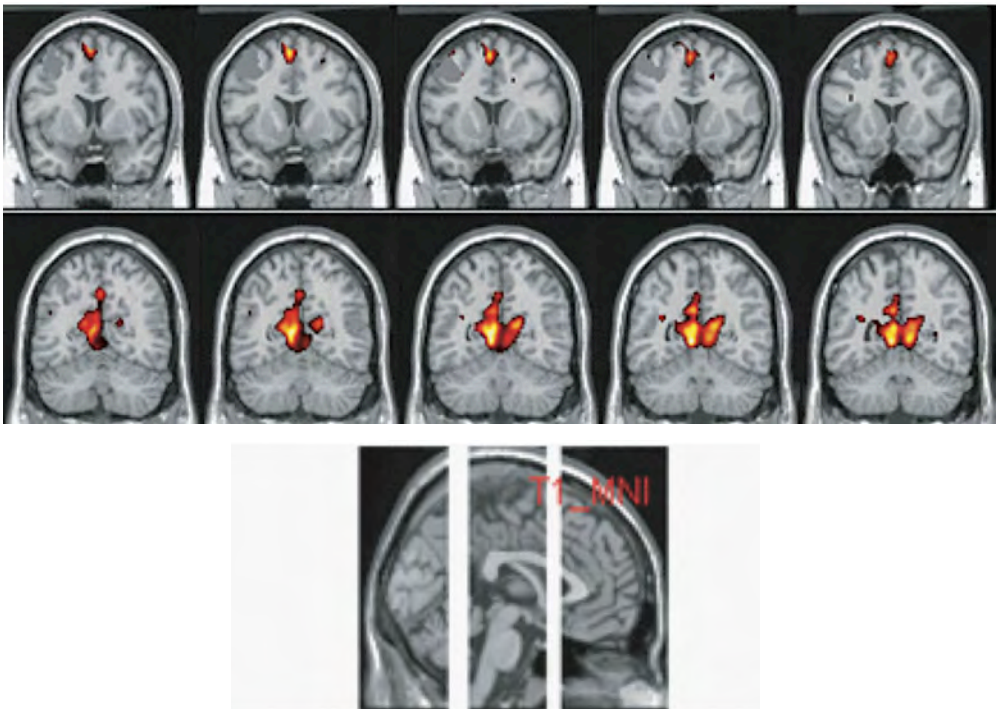



**Figure 2** : Régions cérébrales activées pendant le rappel de souvenirs autobiographiques anciens et récents (cinq périodes de vie) : chez un sujet sain, le même réseau cérébral est activé quelle que soit l'ancienneté du souvenir (d'après Viard et al., 2007).

Selon Nadel et Moscovitch (1997), le LTI intervient tout au long du stockage dans la réactivation interne de la trace mnésique de l'événement (réactualisation du souvenir) qui conduit à la création d'une trace qui sera à son tour encodée et consolidée. La répétition de ce processus produit des *traces multiples* liées à l'événement originel situées dans le LTI et le néocortex. Ainsi, chaque réactivation d'un souvenir crée une nouvelle trace hippocampique : les souvenirs anciens seraient donc associés par un plus grand nombre de traces mnésiques que les souvenirs récents. Gilboa et al. (2004) ont en effet montré que le rappel de souvenirs autobiographiques anciens activait l'hippocampe antérieur et postérieur, alors que le rappel de souvenirs récents activait uniquement l'hippocampe antérieur, suggérant qu'il y aurait plus de traces mnésiques lorsque les souvenirs sont anciens.

Les théories de la consolidation mettent en avant le rôle majeur du LTI et du cortex préfrontal dans le rappel autobiographique. Elles soulignent aussi l'idée qu'une atteinte quelconque du processus de récupération se traduit par une perturbation de la mémoire autobiographique.

# 3. Mémoire autobiographique dans les pathologies neurodégénératives

## 3.1. Amnésies rétrogrades et gradients temporels

Dans la littérature, il existe trois grandes catégories d'amnésies rétrogrades : les amnésies rétrogrades présentant un *gradient temporel de Ribot* (préservation des souvenirs anciens au détriment des souvenirs récents), les amnésies rétrogrades avec un *gradient temporel inverse* (préservation des souvenirs récents au détriment des souvenirs anciens) et les amnésies rétrogrades *sans gradient temporel* (figure 3). Ces trois catégories d'amnésies rétrogrades sont étayées par l'étude des pathologies cérébrales localisées (LTI, lobe temporal externe...). L'étude de l'amnésie rétrograde permet d'évaluer l'influence de l'amnésie sur la dichotomie épisodique/sémantique et de mieux connaître les différents sous-systèmes mnésiques ainsi que leurs substrats neuronaux.

## 3.2. Maladie d'Alzheimer

La maladie d'Alzheimer débutante se manifeste par un déficit prépondérant de la mémoire épisodique et une atteinte du LTI. Elle se caractérise généralement par une amnésie antérograde et rétrograde. Concernant le versant antérograde, la mémoire épisodique est majoritairement perturbée alors que l'atteinte de la mémoire sémantique est variable et dépend du stade d'évolution de la maladie. L'amnésie rétrograde touche aussi bien la mémoire autobiographique que la mémoire des événements publics, quel que soit l'intervalle de rétention.

A l'aide d'un questionnaire semi-structuré, le TEMPau, Piolino et al. (1999) ont confirmé l'atteinte de la mémoire autobiographique épisodique à un stade modéré de la démence. Le rappel total d'événements autobiographiques indique que les souvenirs anciens sont mieux préservés que les souvenirs récents, contrairement aux sujets contrôles. Sur l'ensemble des patients, 80% des souvenirs étaient génériques contre seulement 20% d'épisodiques. Ainsi, la relative préservation des souvenirs anciens chez les patients atteints de la maladie d'Alzheimer concernerait majoritairement des souvenirs sémantisés et non strictement épisodiques. Le gradient de Ribot semble refléter le gradient de sémantisation des souvenirs.

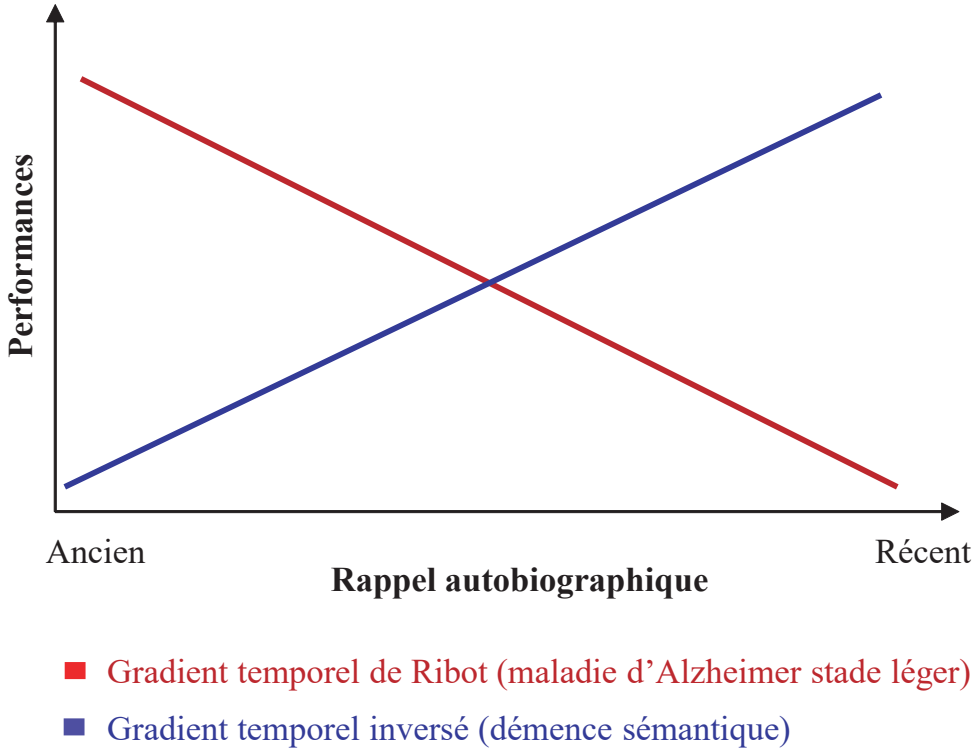



**Figure 3 :** Rappel autobiographique suivant un gradient temporel de Ribot observé dans la maladie d'Alzheimer (rouge) et gradient temporel inversé observé dans la démence sémantique (bleu).

L'étude d'imagerie cérébrale d'Eustache et al. (2004) visait à évaluer la mémoire autobiographique de 17 patients atteints de la maladie d'Alzheimer à l'aide du questionnaire TEMPau en fonction de 3 périodes de vie (enfance/adolescence, période intermédiaire, 5 dernières années), couplée à un examen en tomographie par émissions de positions (TEP). Au niveau comportemental, les analyses ont révélé une perte de mémoire

autobiographique suivant le gradient temporel de Ribot et une analyse qualitative a montré que les souvenirs anciens étaient plus génériques (i.e., sémantiques) que spécifiques (i.e., épisodiques). Au niveau cérébral, des analyses de corrélation ont montré le désengagement de l'hippocampe (partie du LTI) avec l'intervalle de rétention : le rappel de souvenirs récents était corrélé à l'activité de l'hippocampe, ce qui n'était pas le cas pour le rappel de souvenirs anciens. Ces analyses ont également indiqué un passage de gauche à droite de l'engagement du cortex préfrontal : le rappel de souvenirs récents était corrélé avec le cortex préfrontal droit, alors que le rappel de souvenirs anciens était corrélé avec le cortex préfrontal gauche. Les résultats de cette étude sont en accord avec les deux modèles de la consolidation évoqués précédemment : le rappel de souvenirs récents requiert le LTI, notamment l'hippocampe, mais les souvenirs anciens sémantisés ne dépendent plus que de régions néocorticales, notamment le cortex préfrontal.

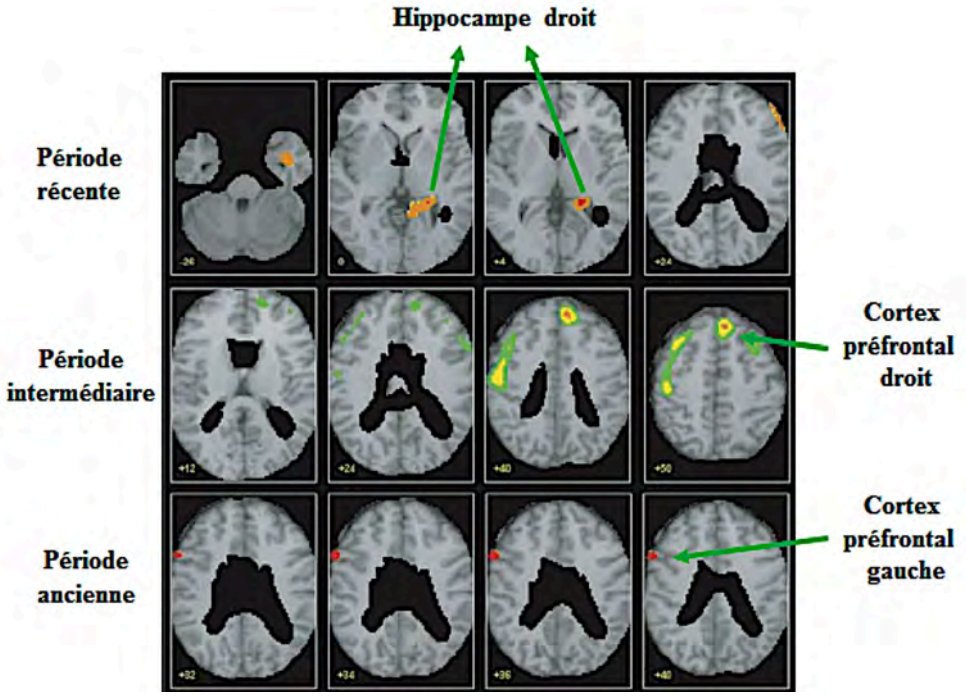



**Figure 4 :** Résultats d'une étude en tomographie par émissions de positions (TEP) montrant les régions cérébrales corrélées avec la performance en rappel autobiographique en fonction de 3 périodes de vie (enfance/adolescence, période intermédiaire, 5 dernières années) chez des patients atteints de la maladie d'Alzheimer (d'après Eustache et al., 2004).

### 3.3. Démence sémantique

La démence sémantique se manifeste essentiellement par un trouble de la mémoire sémantique (verbale et non verbale) alors que la mémoire épisodique, et plus généralement

la mémoire quotidienne, sont préservées chez ces patients présentant une atrophie corticale des parties inféro-latérales du lobe temporal surtout gauche, épargnant le LTI. Cependant, de récentes études montrent que le LTI, notamment l'hippocampe, serait atrophié chez ces patients (La Joie et al., 2013 ; Viard et al., 2013). Ainsi, cette pathologie présente un profil cognitif et neuroanatomique différent de celui de la maladie d'Alzheimer débutante. Cette différence se confirme également dans la composante rétrograde puisque le profil d'amnésie autobiographique dans la démence sémantique, contrairement à la maladie d'Alzheimer débutante, ne semble pas obéir à la loi de Ribot mais présente un gradient temporel inverse où les souvenirs récents sont mieux rappelés que les souvenirs anciens.

Maguire et al. (2010) ont effectué un suivi longitudinal d'un patient atteint de démence sémantique avec un examen d'IRMf. Le patient devait ré-évoquer ses souvenirs autobiographiques dans l'IRM. Les résultats ont montré que sa mémoire autobiographique déclinait progressivement au cours des trois années consécutives qui coïncidait avec une atrophie cérébrale accrue, notamment de l'hippocampe. Dans ce même contexte, nous avons comparé le profil d'atrophie de deux patients atteints de démence sémantique au cours d'un examen d'IRMf pendant lequel les patients devaient ré-évoquer leurs souvenirs autobiographiques en fonction de cinq périodes de vie (Viard et al., 2014) : 3 périodes anciennes (0--17 ans, 18--30 ans, 31ans--5 dernières années) et 2 périodes récentes (5 dernières années, 12 derniers mois). Les deux patients (JPL et EP) présentaient une atrophie du cortex temporal latéral, caractéristique de cette pathologie, et le patient JPL présentait en plus une atrophie de l'hippocampe bilatéral. Au niveau comportemental, JPL avait un déficit dans le rappel de souvenirs récents et anciens, contrairement à EP qui avait des performances correctes. Au niveau des résultats d'imagerie, JPL hyperactivait certaines régions néocorticales mais ce processus de compensation restait inefficace puisqu'il n'arrivait pas à évoquer des souvenirs autobiographiques de manière épisodique. Cette étude a permis de montrer que l'intensification de l'atrophie hippocampique affecte profondément le rappel de souvenirs (récents et anciens) dans cette pathologie et que l'hyperactivation de régions néocorticales est parfois insuffisante pour compenser efficacement le déficit autobiographique chez les patients avec une atteinte hippocampique trop importante. Nous avons revu ces deux patients, ainsi que deux nouveaux patients, pour leur proposer une tâche de projection dans le futur, associée à un examen d'IRMf (voir ci-dessous).

# 4. Projection dans le futur

## 4.1. Définition et réseau cérébral

Selon Tulving, la mémoire épisodique nous permet de voyager mentalement dans le temps, qu'il soit passé (i.e., revivre mentalement des expériences passées) ou futur (i.e., vivre mentalement des expériences futures). Ainsi, cette mémoire épisodique nous permet de nous représenter dans le futur par projection mentale. Buckner et Carroll (2007) montrent que le réseau cérébral impliqué dans le rappel autobiographique est très similaire au réseau cérébral impliqué dans la projection vers le futur. Dans une étude en IRMf, nous

avons demandé à des volontaires sains d'évoquer des souvenirs passés ayant eu lieu dans les 12 derniers mois et des événements futurs qu'ils avaient projeté d'effectuer dans les 12 prochains mois (Viard et al., 2011). Confortant le constat de Buckner et Carroll (2007), nos résultats d'imagerie ont montré un réseau d'activation commune entre le passé et le futur, comprenant l'hippocampe, le cortex préfrontal médian et latéral, le cortex cingulaire postérieur, le précuneus et des régions postérieures (occipitales et pariétales). Ces résultats confortent la théorie de Schacter et Addis (2007) selon laquelle la mémoire épisodique serait une base sur laquelle se construisent nos simulations futures. Ainsi, les événements passés et futurs reposent sur les mêmes informations stockées en mémoire épisodique et dépendent des même processus cognitifs (implication du soi, imagerie visuelle mentale, processus d'assemblage d'éléments disparates…).

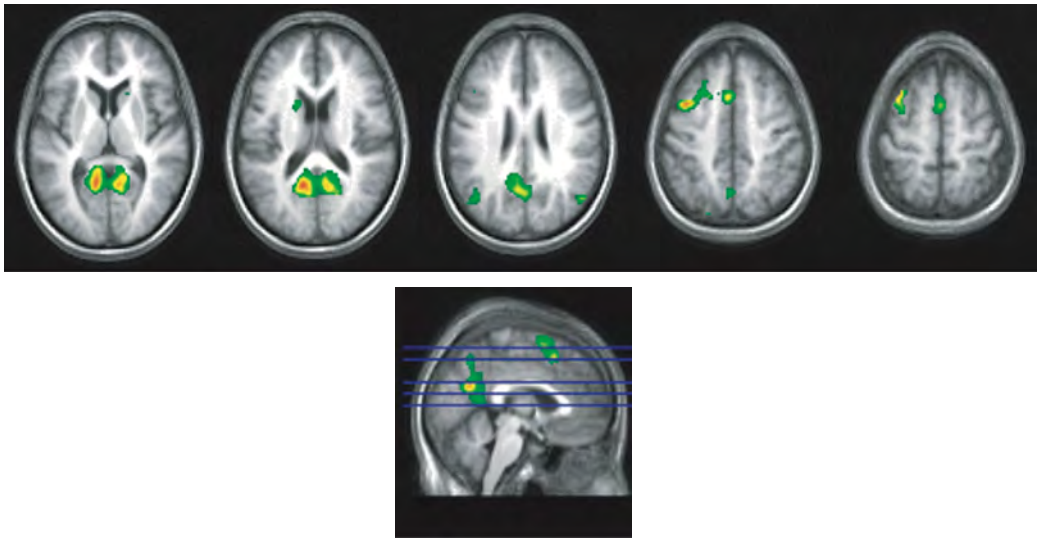


**Figure 5** : Régions activées en commun lors du rappel en mémoire autobiographique et de la projection dans le futur (d'après Viard et al., 2011).

### 4.2. Projection dans le futur et démence sémantique

Peu d'études se sont intéressées à la projection dans le futur dans les pathologies neurodégénératives en imagerie cérébrale. Nous avons proposé une étude en IRMf à quatre patients atteints de démence sémantique dans laquelle ils devaient évoquer des souvenirs passés ayant eu lieu dans les 12 derniers mois et des événements futurs qu'ils avaient projeté d'effectuer dans les 12 prochains mois (Viard et al., 2014). Les résultats ont montré que tous les patients avaient une atrophie du cortex temporal latéral, caractéristique de cette pathologie. Deux patients présentaient en plus des atrophies importantes de l'hippocampe (patient JPL) et du cortex préfrontal médian (patients JPL et EP). Chez ces deux patients, l'évocation de souvenirs passés ou d'événements futurs était fortement altérée. A l'inverse, chez les patients qui ne présentaient pas d'atrophie dans ces deux régions (patient EG) ou

une atrophie restreinte de l'hippocampe (patient LL), l'hyperactivation du tissu résiduel ou d'autres régions ayant un rôle compensatoire leur permettait de se rappeler et de se projeter efficacement dans le futur. Cette étude a permis de montrer que l'intégrité fonctionnelle de l'hippocampe et du cortex préfrontal médian était cruciale pour la projection dans le futur : l'atrophie de ces deux structures empêche cette projection alors que l'intégrité de ces deux structures, ou l'hyperactivation du tissu résiduel, permet de normaliser les capacités de projection dans le futur.

## 5. Conclusion et perspectives

Dans ce chapitre, nous avons présenté la mémoire humaine et ses troubles dans le cadre théorique actuellement le plus admis en neuropsychologie, celui des systèmes de mémoire et de leurs mécanismes de fonctionnement. Nous avons ensuite décrit le concept de mémoire autobiographique, les contributions de l'imagerie cérébrale fonctionnelle à sa compréhension, puis ses modifications dans différents cadres pathologiques. En fait, des troubles de mémoire, d'intensités diverses et avec des symptômes multiples, peuvent survenir dans des situations cliniques variées. En dehors de toute situation conduisant à une lésion cérébrale (traumatisme cranio-cérébral, accident vasculaire cérébral…), il s'agit des grandes pathologies psychiatriques comme la schizophrénie ou la dépression, mais également des pathologies non-cérébrales comme les conséquences psychosociologiques d'un cancer, le cancer du sein chez la femme ayant été le plus étudié. En fait, toute situation qui met en cause l'identité de la personne et son lien avec son environnement peut entrainer des troubles de la mémoire. Un autre exemple est le trouble de stress post-traumatique, dont les critères de diagnostic ont évolué au fil du temps et mettent aujourd'hui l'accent sur les distorsions de la mémoire, notamment des images et des pensées intrusives, qui se trouvent au cœur de la sémiologie de ce syndrome. Ainsi, au-delà des pathologies qui atteignent les mécanismes centraux de « l'instrument mémoire » et leurs substrats cérébraux, de nombreuses situations qui remettent en cause une relation harmonieuse avec soi et avec les autres sont susceptibles d'entrainer des altérations de la mémoire. Pour ces raisons, de nombreuses recherches se développent aujourd'hui aux confins des neurosciences et sciences humaines et sociales. Il en est ainsi des stéréotypes, comme ceux associés au vieillissement, qui peuvent exercer un effet significatif sur les performances de mémoire des personnes âgées, si elles sont placées dans une situation qui suggère que la tâche qu'elles réalisent est sensible aux effets de l'âge. Ces effets ne sont pas marginaux et permettent de nuancer une vision trop mécaniciste du fonctionnement de la mémoire. Un autre exemple est l'étude des interactions entre la mémoire individuelle et la mémoire collective. On connait de mieux en mieux les mécanismes qui concourent à l'élaboration de ces mémoires, mais jusqu'à présent, elles donnaient lieu à des études bien séparées, les premières relevant de la psychologie et de la biologie et les secondes de la sociologie et de l'histoire. L'approche transdisciplinaire permet de comprendre leurs interactions, mais aussi d'appréhender différemment des situations cliniques où un individu singulier éprouve des difficultés à inscrire son histoire personnelle dans l'histoire de la société dans laquelle il évolue (Legrand et al, 2015 ; Eustache et al, 2017 ; Eustache et al, 2018). La nouvelle science de la mémoire se trouve ainsi aux confins de multiples disciplines et courants de

pensée et prend une place importante dans le monde moderne face aux nouveaux moyens de communication et aux déplacements massifs de populations.

**Références**


Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. Trends in Cognitive Science 11:49–57.

Cabeza R, Prince SE, Daselaar SM, Greenberg DL, Budde M, Dolcos F, LaBar KS, Rubin DC (2004) Brain activity during episodic retrieval of autobiographical and laboratory events: an fMRI study using a novel photo paradigm. J Cogn Neurosci 16:1583-94.

Conway MA, Pleydell-Pearce CW (2000) The construction of autobiographical memories in the self-memory system. Psychological Review 107:261-188.

Eustache F, Amieva H, Thomas-Antérion C, Ganascia J-G, Jaffard R, Peschanski D, Stiegler B. (2017) Ma mémoire et les autres. Paris : Le Pommier.

Eustache F, Amieva H, Thomas-Antérion C, Ganascia J-G, Jaffard R, Peschanski D, Stiegler B. (2018) La mémoire au futur. Paris : Le Pommier.

Eustache F, Desgranges B. Les chemins de la mémoire. Paris: Le Pommier/Inserm; 2012.

Eustache F, Desgranges B, Lalevée C (1998) L'évaluation clinique de la mémoire. Revue neurologique 154S:18-32.

Eustache F, Faure S, Desgranges B. Manuel de Neuropsychologie. 5 ed. Paris: Dunod; 2018.

Eustache F, Guillery-Girard B. La Neuroéducation, La mémoire au cœur des apprentissages. Paris : Ed Odile Jacob, 2016.

Eustache F, Piolino P, Giffard B, Viader F, de La Sayette V, Baron JC, Desgranges B (2004) In the course of time: a PET study of the cerebral substrates of autobiographical amnesia in Alzheimer's disease. Brain 127:1549-1560.

Eustache F, Viard A, Desgranges B (2016) The MNESIS model: Memory systems and processes, identity and future thinking. Neuropsychologia 87:96-109.

Gardini S, Cornoldi C, De Beni R, Venneri A (2006) Left mediotemporal structures mediate the retrieval of episodic autobiographical mental images. NeuroImage 30:645–655.

Gilboa A, Winocur G, Grady CL, Hevenor SJ, Moscovitch M (2004) Remembering our past: functional neuroanatomy of recollection of recent and very remote personal events. Cerebral Cortex 14:1214-1225.

Greenberg DL, Rice HJ, Cooper JJ, Cabeza R, Rubin DC, Labar KS (2005) Co-activation of the amygdala, hippocampus and inferior frontal gyrus during autobiographical memory retrieval. Neuropsychologia 43:659-674.

Guillery B, Desgranges B, Katis S, de La Sayette V, Viader F, Eustache F (2001) Semantic acquisition without memories: evidence from transient global amnesia. NeuroReport 12:3865-3869.

La Joie R, Perrotin A, de La Sayette V, Egret S, Doeuvre L, et al. (2013) Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. NeuroImage Clin 3:155–162.

Legrand N, Gagnepain P, Peschanski D, Eustache F (2015) Neurosciences et mémoires collectives : les schémas entre cerveau, sociétés et cultures. Biol Aujourdhui 209:273-86

Maguire EA, Kumaran D, Hassabis D, Kopelman MD (2010) Autobiographical memory in semantic dementia: a longitudinal fMRI study. Neuropsychologia 48:123-136.
Nadel L, Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. Current Opinion in Neurobiology 7:217-227.
Piolino P, Desgranges B, Giffard B, Guillery B, Benali K, Lalevée C, de La Sayette V, Eustache F (1999) La mémoire autobiographique dans le vieillissement normal et dans la maladie d'Alzheimer. Revue de Neuropsychologie 9:452-453.
Schacter DL, Addis DR (2007) The cognitive neuroscience of constructive memory: remembering the past and imagining the future. Philos Trans R Soc Lond B Biol Sci 362:773–786.
Squire LR, Alvarez P (1995) Retrograde amnesia and memory consolidation: a neurobiological perspective. Current Opinion in Neurobiology 5:169-177.
Steinvorth S, Corkin S, Halgren E. 2006. Ecphory of autobiographical memories: An fMRI study of recent and remote memory retrieval. Neuroimage 30:285–298.
Svoboda E, McKinnon MC, Levine B (2006) The functional neuroanatomy of autobiographical memory: A meta-analysis. Neuropsychologia 44:2189–2208.
Vargha-Khadem F, Gadian DG, Watkins KE, Connely A, Van Paesschen W, Mishkin M (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. Science 277:376-380.
Viard A, Chételat G, Lebreton K, Desgranges B, Landeau B, et al. (2011) Mental time travel into the past and the future in healthy aged adults: an fMRI study. Brain Cogn 75:1–9.
Viard A, Desgranges B, Matuszewski V, Lebreton K, Belliard S, et al. (2013) Autobiographical memory in semantic dementia: New insights from two patients using fMRI. Neuropsychologia 51:2620–2632.
Viard A, Lebreton K, Chételat G, Desgranges B, Landeau B, Young A, De La Sayette V, Eustache F, Piolino P (2010) Patterns of hippocampal-neocortical interactions in the retrieval of episodic autobiographical memories across the entire life-span of aged adults. Hippocampus 20:153-165.
Viard A, Piolino P, Desgranges B, Chételat G, Lebreton K, et al. (2007) Hippocampal activation for autobiographical memories over the entire lifetime in healthy aged subjects: An fMRI study. Cereb Cortex 17:2453–2467.
Viard A, Piolino P, Belliard S, de La Sayette V, Desgranges B, Eustache F (2014) Episodic future thinking in semantic dementia: a cognitive and FMRI study. PLoS One 9:e111046.

# 3

# Les Bases Neurobiologiques de la Conscience une revue des avancées récentes[1]

**Claire Sergent**

Laboratoire
Psychologie de la Perception
UMR 8242, Université Paris Descartes

## Abstract

On the theme of neural correlates of consciousness, Claire Sergent presents a synthesis of recent advances obtained within the large French team (Collège de France, CNRS, Paris-Descartes University, Pitié-Salpêtrière Hospital, Institut du cerveau) which, since Jean-Pierre Changeux, passing by Stanislas Dehaene and Lionel Naccache, with their collaborators, was interested in the running of the human brain. Using modern medical imaging techniques and for certain experiments those of electroencephalography, Claire Sergent shows how we have been able to highlight, to characterize qualitatively and quantitatively, the differences in brain activations between conscious perceptual processes and unconscious perceptual processes. She continues by indicating the interpretations given to these results, in the form of an *awareness model* involving the activation and reactivation of a global workspace in contact with several areas of the brain that have no direct contact with each other. She ends by showing what contributions these advances can make in the context of the diagnosis of conscious states in human clinic. In her conclusion, she affirms that the research to come will make it possible to explore the opposite path from the one followed so far, and therefore "to start from the brain activity to go towards the subjective experience" of the individual.

[1] Ce chapitre est la transcription, effectuée par Pierre Nabet et Jean Pierre Treuil, membres de l'AEIS, de la conférence de Claire Sergent faite au colloque organisé par l'AEIS, à l'Institut Henri Poincarré, le 15 mars 2018 ; le texte a été relu et amendé par la conférencière. Il est publié avec son accord.

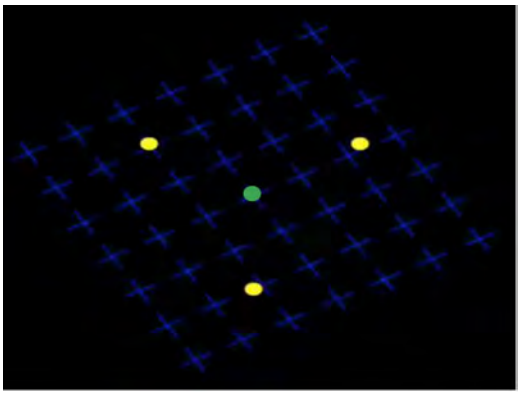
# 1. Introduction

Comme il a été dit en introduction de ce colloque, la période scientifique que nous vivons actuellement semble bien être unique dans l'évolution de notre société : d'une part nous avons les outils conceptuels de la psychologie expérimentale, le développement de l'intelligence artificielle et, par ailleurs, des outils techniques très évolués, dont l'imagerie cérébrale, pour comprendre la neurologie de la cognition. Il s'agit donc de faire ici le point sur ce que l'on sait, d'un point de vue des neurosciences, sur la conscience chez l'homme.

Nous examinerons successivement :

- La cognition inconsciente,
- Les corrélats neuronaux de la conscience,
- Les mécanismes et modélisations, et enfin
- Les signatures de la conscience

# 2. La cognition inconsciente



Une expérience simple peut nous donner une première idée de ce qu'on appelle la conscience : La figure ci-contre représente un écran d'ordinateur avec une texture de fond constituée de petites croix bleues. Sur ce fond bleu figurent trois points jaunes représentant les sommets d'un triangle. Au milieu de ce triangle apparaît un point vert intense et on demande au sujet de fixer fortement son regard sur ce point aussi longtemps que possible en essayant de ne pas ciller. Au bout d'un moment il se rend compte que les points jaunes disparaissent de la vue, totalement ou partiellement, il n'en a plus conscience, alors que l'on sait bien qu'ils sont toujours présents. D'ailleurs, si à ce moment il ferme et rouvre les yeux, les points jaunes réapparaissent car cette interruption suffit à lever l'effet de cette illusion. Ce phénomène a été baptisé « cécité induite par le mouvement » (Bonneh, Cooperman & Sagi Nature 2001). C'est un des nombreux exemples d'expériences qui nous indique que, même à l'éveil, notre cerveau peut traiter une information sans que nous en ayons conscience.

Que se passe-t-il donc dans le cerveau, à l'éveil, lorsque ce dernier traite une information de manière non consciente ?

*La figure 2* illustre une expérience de masquage. On présente au sujet un mot (un stimulus), ici le mot *neuf* (nine=9), pendant quelques dizaines de millisecondes, ici 43 millisecondes (*figure 2a, partie supérieure*). Ce temps peut paraitre très court, mais il est largement suffisant pour que le système perceptif permette de le voir et d'en comprendre le sens. Le sujet en est conscient. Par contre si ce stimulus de 43 millisecondes est immédiatement précédé et suivi de masques (dans le cas présent des images de 71 millisecondes), il disparait de la conscience et devient subliminal, *le sujet n'en n'a plus conscience*. Cependant, on montre que ce mot caché et en particulier le sens de ce mot caché, va pouvoir influencer le comportement du sujet vis-à-vis de stimuli conscients.

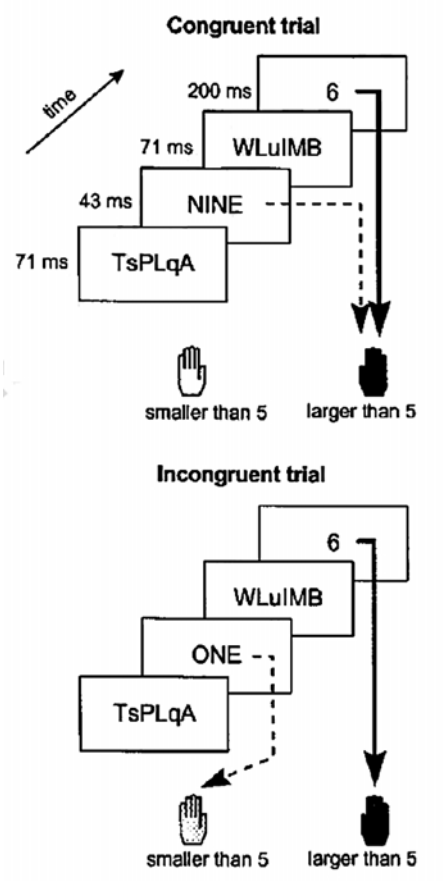



*Figure 2a*

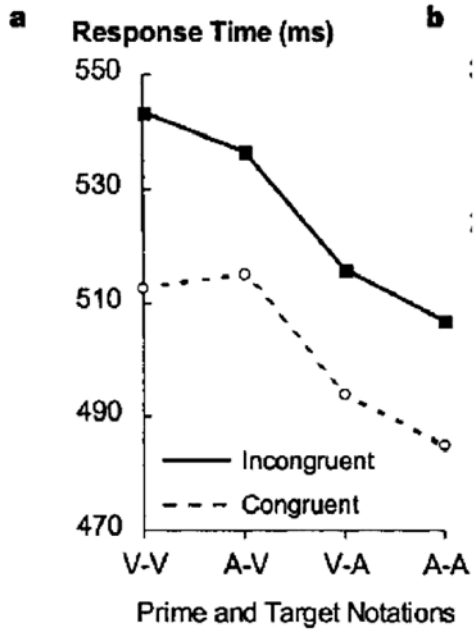



*Figure 2b*

**Figure 2** : expérience de masquage ; dans cette expérience, le sujet est placé devant un ordinateur et on lui fait voir une image (dans le cas présent un chiffre) seule ou encadrée par d'autres images. Image extraite de Dehaene et al, Nature 1998

Dans cette expérience en effet, on demande au sujet de comparer des symboles parfaitement visibles (par exemple ici le nombre 6) au nombre 5, et on regarde si la réponse est influencée par la présence d'un mot caché qui précède (le 9 dans la partie supérieure de la figure 2a, le 1 dans la partie inférieure). On observe effectivement une influence en terme de temps de réaction (***figure 2b***) : il y a une accélération du temps de réponse dans le cas où ce nombre visible a été précédé d'un mot invisible votant pour la même réponse que lui (par exemple le mot « neuf »/ « nine » dans la figure 2a), par rapport au temps de réponse constaté lorsqu'il a été précédé d'un mot invisible votant pour la réponse opposée (par exemple le mot « un »/ « one » dans la figure 2a). Bien entendu, l'expérience a été répétée avec différents formats de ces chiffres (notation arabe ou en toutes lettres dénotés respectivement A et V pour verbal dans la figure 2b) : quel que soit le format de présentation du stimulus invisible et du stimulus visible on obtient le même résultat, c'est bien le sens des stimuli qui est reconnu, au-delà de leur format physique.

Ces travaux de masquage, qui sont tout à fait significatifs et emblématiques, démontrent qu'un sujet humain est capable d'extraire le sens d'un mot subliminal, perçu de manière non consciente.

Par delà ces premières constatations, on regarde comment le cerveau traite ces informations et, en particulier, on analyse quelle est l'influence du mot caché sur la préparation motrice, sur le système moteur (*Figure 3*).

La réponse correcte au nombre visible est, dans le cas de cette expérience, une action de la main droite s'il est plus grand que 5, et de la main gauche s'il est plus petit ; les préparations motrices de ces deux réponses produisent, en électro-encéphalographie, des signatures très connues : au clic avec la main droite, une petite négativité apparait au dessus du cortex moteur gauche, et inversement. Mais que se passe-t-il dans le cerveau dans sa réaction au nombre invisible ? Plus précisément, voit-on un début de préparation motrice qui serait la bonne dans la comparaison de ce nombre au nombre 5 ? Sur la topographie toujours enregistrée en électro-encéphalographie, on voit effectivement apparaître une petite préparation motrice en réponse au nombre invisible, cohérente avec une réponse à droite, s'il est plus grand que 5 et une réponse à gauche s'il est plus petit que 5.

Donc, non seulement le cerveau a réussi à extraire le sens du stimulus mais de plus il a su appliquer une consigne arbitraire qui était de le comparer à 5. On voit ainsi à quelle complexité peut accéder le traitement cérébral d'un stimulus inconscient.

En conclusion rapide sur le traitement non-conscient, contrairement à la position qui a été longtemps celle des neurosciences, qui était de dire que finalement un processus inconscient était un réflexe, puis, dans les années 1970, que les processus inconscients sont des réflexes qui peuvent être médiés par le cortex, les études actuelles montrent qu'il faut penser beaucoup plus large et qu'il existe réellement une cognition inconsciente. De plus,

et de manière intéressante, on montre aussi qu'il ne semble pas y avoir de territoire interdit dans le cortex, pour des représentations non conscientes de quelque nature qu'elles soient. Les études récentes ont montré, par exemple, qu'il peut y avoir des activations frontales inconscientes (van Gaal et al., Journal of Neuroscience 2010).

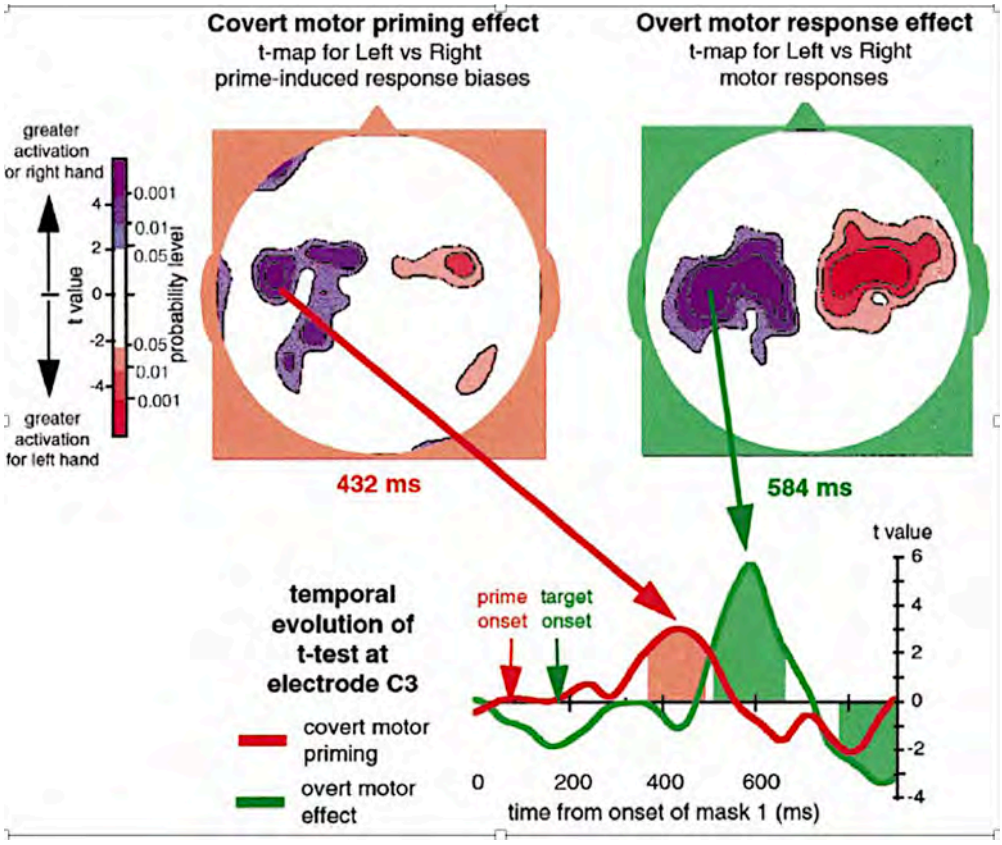



**Figure 3** : réponses du cerveau aux expériences de masquage démasquage. On regarde ici la signature, dans le cerveau, du geste effectué par le sujet en réponse au stimulus conscient qu'il perçoit (en vert), et, en amont, l'amorce de préparation motrice engendrée par le stimulus amorce masqué (en rouge/rose). Image extraite de Dehaene et al, Nature 1998

## 3. Corrélats neuronaux de la conscience

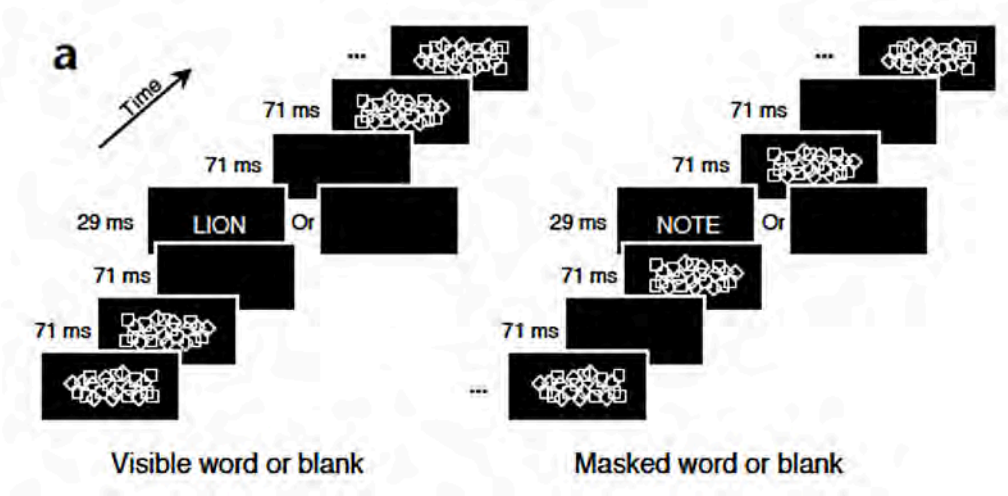



*Figure 4a* *Figure 4b*

**Figure 4** : expérience de masquage et démasquage des stimuli ; principe. Image extraite de Dehaene et al, Nature Neuroscience 2001

*La figure 4* illustre à nouveau une expérience de mots masqués. Dans la partie droite (*figure 4b*), le stimulus (ici le mot NOTE) est masqué et n'est pas du tout perçu par le sujet. Par un tout petit changement de stimulation on peut démasquer ce mot (*figure 4a*). Pour ce faire, on intercale entre le mot masqué et les masques qui précèdent et suivent immédiatement le mot, une petite période - d'une durée de 71 ms, identique à celle des masques - d'écran blanc. Le mot est alors perçu par le sujet, il a été démasqué. On enregistre par IRM fonctionnelle l'activité cérébrale dans le premier cas et dans le second pour contraster les deux situations (c'est la méthode contrastive classique de la science).

Les résultats obtenus concernant l'activité cérébrale sont représentés dans la *figure 5* : dans le cas subliminal - où le mot n'est pas perçu consciemment - la forme du mot a cependant bien été extraite, mais les activations restent très localisées dans le lobe temporal, au niveau de l'aire de la forme visuelle des mots (*figure 5b*). Lorsqu'au contraire, le même mot est démasqué par la manipulation décrite plus haut (*figure 5a*), on constate que l'activation envahit une beaucoup plus grande partie du cortex. Cette activation est plus forte que dans le cas masqué, comme on le voit sur la courbe verte au centre de la figure. On constate une extension du réseau d'activation, qui est intéressé par cette information, par exemple sur l'aire de BROCA (activation en vert à la base du cortex préfrontal sur la figure 5a), extension contenant aussi le cortex préfrontal et des aires pariétales, et semblant donc toucher aussi le système attentionnel du cerveau.

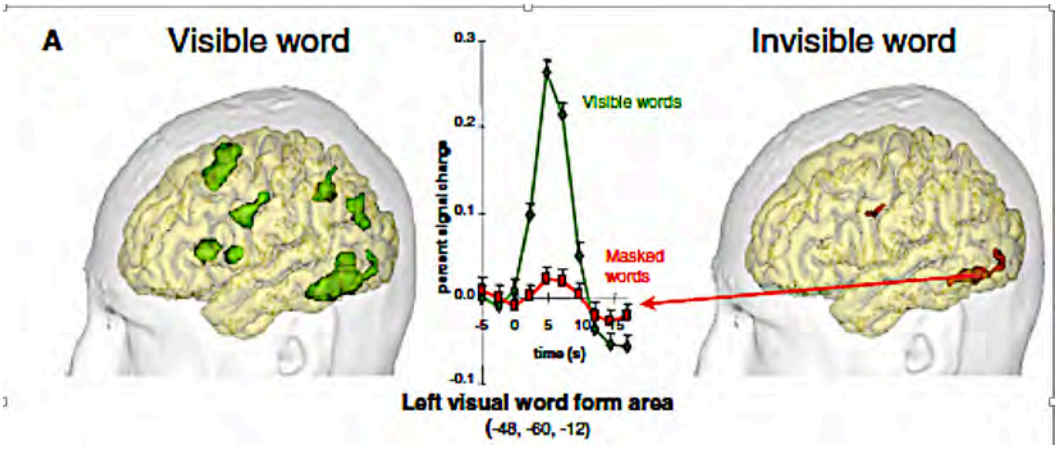



**Figure 5** : expérience de masquage et démasquage des stimuli ; imagerie cérébrale. Image extraite de Dehaene et al, Nature Neuroscience 200

Il est intéressant de noter que ce genre d'études, qui font contraster traitement conscient et non conscient à stimulation égale ou quasiment égale, ont été réalisées, maintenant, avec de nombreux types de stimulus différents, convergent vers des corrélats neuronaux qui semblent être communs et indépendants de la modalité de stimulation.

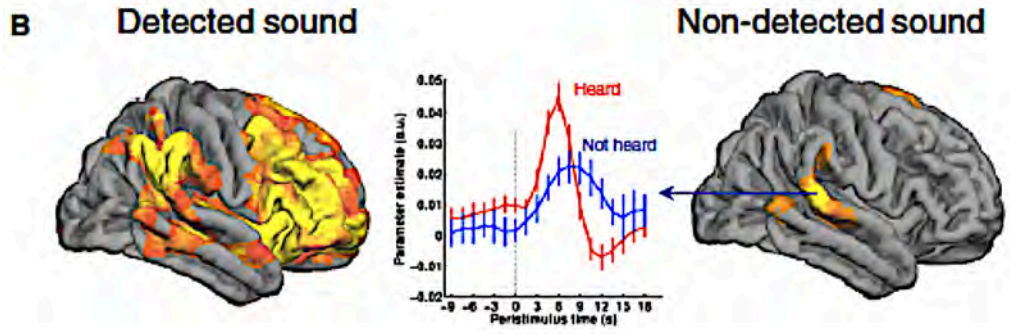



**Figure 6** : expérience de présentation de son au seuil de perception consciente; IRM fonctionnelle. Image extraite de Sadghiani et al, Journal of Neuroscience 2009

La figure 6 illustre ainsi le cas de stimuli auditifs, des sons présentés au seuil de perception consciente, et qui sont parfois rapportés comme entendus ou non-entendus. Lorsque ces stimuli sont rapportés comme non-entendus, on voit effectivement le cortex auditif répondre seul ; lorsqu'ils sont au contraire rapportés comme entendus (pour une stimulation cette fois strictement identique), le réseau des réactions à ces stimuli auditifs s'étend de façon notable. Il est intéressant de souligner que même en l'absence de manipulation de

masquage, et simplement sur la base de ce que rapporte le sujet (perception consciente ou non), on aboutit à un résultat similaire : un stimulus identique est traité par un réseau beaucoup plus vaste lorsqu'il est perçu consciemment que lorsqu'il est perçu inconsciemment.

Dans les quelques dizaines d'années qui viennent de s'écouler, ces genres d'expériences ont apporté un consensus sur les corrélats neuronaux de la conscience : si des débats se font jour, actuellement, sur des questions plus complexes, en tout cas, sur les corrélats neuronaux de la conscience, une espèce de consensus s'est établie, qui est que lorsqu'on prend conscience d'une information on constate (cf schéma représenté *figure 7*) une augmentation du signal sur les aires sensorielles, une extension du réseau qui s'intéresse à cette information et qui va inclure notamment des aires frontales et pariétales, mais aussi le cortex cingulaire antérieur (impliqué dans le système limbique, lié à l'évaluation des stimuli et à la motivation, donc également aux émotions). De plus, on voit aussi une augmentation du dialogue au sein de ce réseau. Afin de comprendre quels sont les mécanismes à l'œuvre derrière ces corrélats, il nous faut étudier plus avant le scénario de leur mise en place au moment de la prise de conscience.

## 4. Mécanismes et Modélisations

Pour aller au-delà de cette vision statique, il faut s'interroger sur le scénario de mise en place de ces phénomènes lorsqu'un sujet prend conscience d'une information, autrement dit en nous interrogeant sur les mécanismes en œuvre.

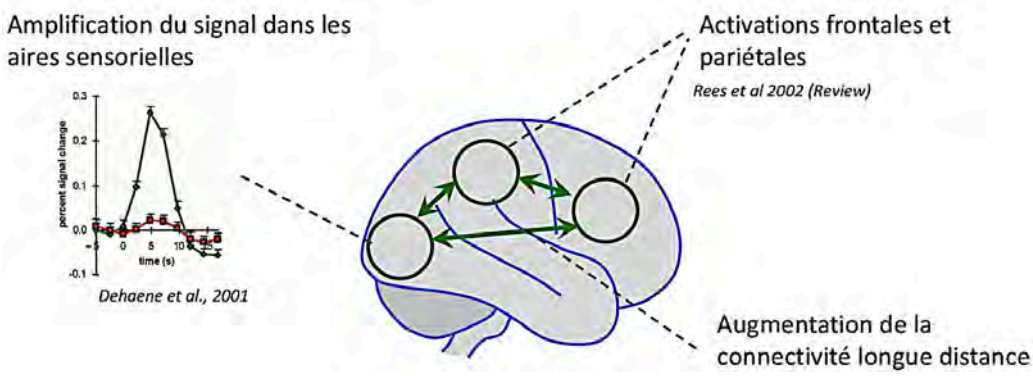



**Figure 7** : Corrélats neuronaux de la conscience. (cf. Gross et al. 2004, Gaillard et al.2009*)*

C'est pour répondre à cette interrogation que nous avons mené une expérience un peu différente du masquage, un phénomène qu'on appelle *clignement attentionnel* ou *attentionnal blink.* Dans ce genre d'expérience, le sujet est encore une fois devant un écran d'ordinateur et voit défiler des items de manière rapide, mais ici, l'intervalle entre les différents items est suffisant pour que le sujet soit capable de rapporter ces items indépendamment, sans difficulté, malgré les masques situés avant et après le mot, comme

dans l'expérience de masquage. On demande d'abord au sujet de rapporter le chiffre qui lui est présenté (le chiffre cinq sur la figure 8a) il n'a bien sûr aucun mal à le faire. On lui présente alors exactement la même chose, mais en lui demandant aussi de prêter attention à une cible qui est présentée avant ce mot. On constate alors que le fait de porter attention à cette première cible va faire qu'il sera incapable de voir la deuxième cible si elle arrive dans un intervalle de temps critique autour de 300 millisecondes après la première : comme s'il avait cligné à ce moment là. Bien sûr on vérifie que le sujet ne cligne pas véritablement, c'est un *clignement attentionnel.*

Le fait de porter attention à un premier stimulus va ouvrir un intervalle de temps où le second stimulus ne sera pas perçu. Il est alors intéressant d'étudier quelle est la perception de ce deuxième stimulus au moment du clignement attentionnel : au lieu de demander au sujet de dire seulement s'il a vu ou pas vu le deuxième stimulus, on lui donne une échelle subjective pour rapporter *à quel point* il a bien vu ce deuxième stimulus ; la partie supérieure de la *figure 8b* montre cette échelle de visibilité subjective qui va de 0 (je n'ai rien vu du tout) à 100 (j'ai très bien vu). Les distributions de réponses sur cette échelle, sont surprenantes : pour un même participant, un même individu, et pour exactement la même stimulation, dans certains essais le sujet dira *je n'ai rien vu*, (comme dans les essais témoins où rien n'a été présenté), et dans d'autres essais il dira *oui j'ai très bien vu*. Ceci suggère la présence d'une dynamique sous-jacente, peut-être une dynamique non linéaire.

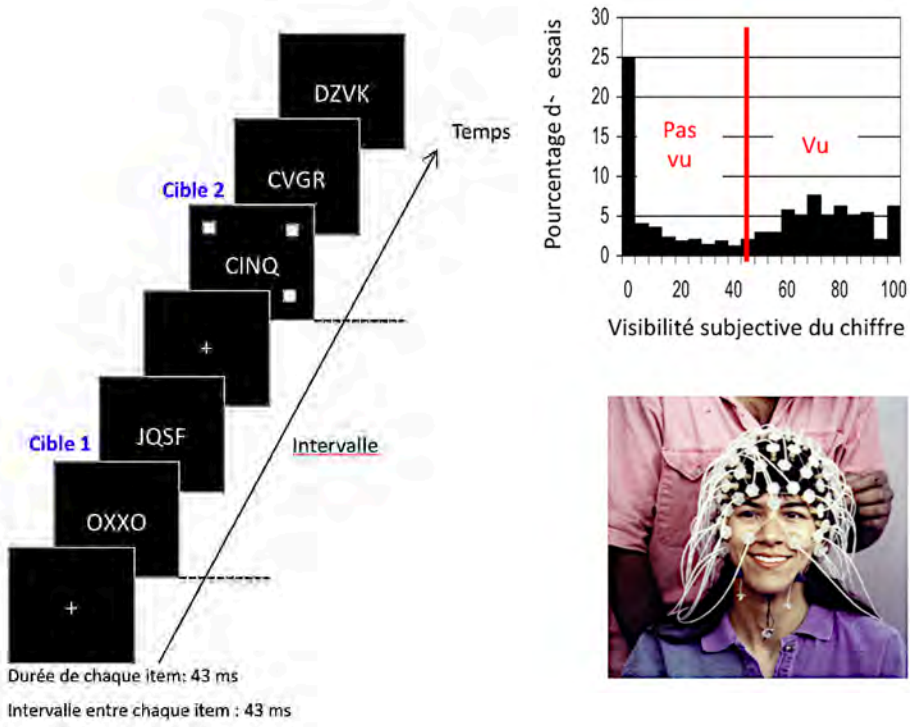



**Figure 8** : Clignement attentionnel (attentional blink). Images extraites de Sergent et Dehaene, Psych. Science 2004. Sergent, Baillet et Dehaene, Nature Neuroscience 2005

Nous avons réalisé des électro-encéphalogrammes pour savoir ce qui se passe au niveau de l'activité cérébrale lorsque le sujet dit qu'il a vu ou pas vu le stimulus. *La figure 9* montre les résultats obtenus (rappelons que dans les deux cas la stimulation est identique). Elle montre à différents instants, sur une image « plate » l'activité des différentes électrodes posées sur le crâne du sujet. Les premières ondes cérébrales observées le sont à 96 ms : des activations sur les électrodes occipitales (donc le système visuel) apparaissent de manière identique lorsque le stimulus est vu ou pas vu, absence surprenante de différence, contrairement à ce qu'on obtenait en IRM fonctionnelle. Il en est de même encore à 180 ms où les activations sont très fortes, très riches, identiques dans les deux cas. C'est seulement à 250 ms après la présentation du stimulus, qu'une divergence apparaît : lorsque le sujet à la fin de l'essai dit *j'ai vu le stimulus* on voit l'enclenchement de toute une série d'ondes cérébrales qui ne sont pas déclenchées lorsque le sujet dit ne pas avoir vu le stimulus et ceci jusqu'à plus d'une demi-seconde après la présentation du stimulus.

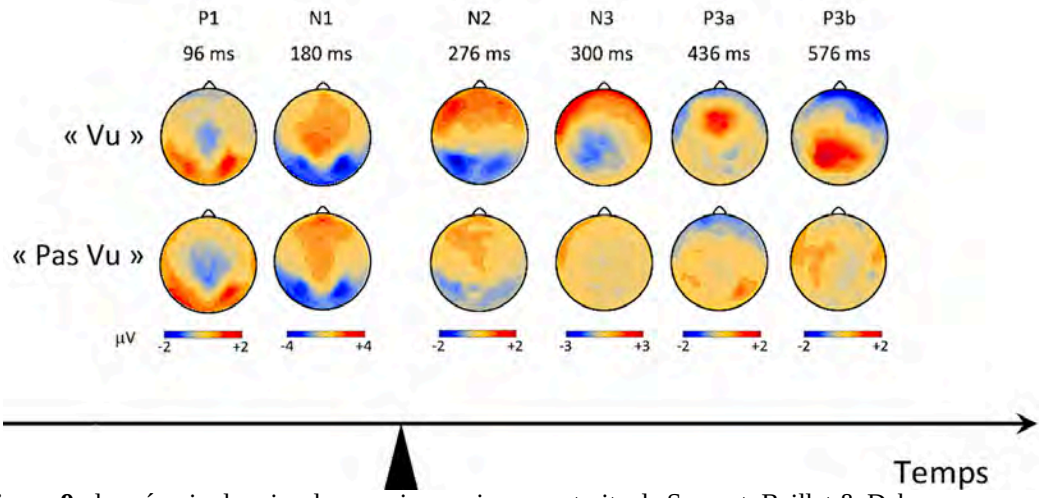



**Figure 9** : le scénario de prise de conscience ; image extraite de Sergent, Baillet & Dehaene, Nature Neuroscience 2005

On peut présenter ces résultats de la prise de conscience par le cerveau d'une information sous forme d'un film. Il est bien sûr, impossible de le monter dans un texte écrit, mais on peut le décrire : Ce film montre les faces gauche et droite du cerveau ainsi que la face du dessous. Sur les images qui défilent figurent, en bas, les activations lorsque le sujet dit qu'il a vu le stimulus et en haut lorsqu'il dit ne pas l'avoir vu. Un petit compteur de temps figure en bas et à gauche. Le film commence à 40 ms avant la présentation du stimulus. Les premières activations se font jour autour de 100 ms après la présentation du stimulus, ce sont de belles activations dans le cortex occipital : ce système neuronal, lieu d'intégration des stimuli visuels, est donc en train de répondre et ce de manière similaire pour les stimuli conscients et inconscients.

On voit très bien la progression de ces activations dans le cortex temporal, ce qui est attendu et conforme à ce que l'on sait. On remarque d'ailleurs que cette progression est un peu plus forte à gauche qu'à droite car, rappelons-le, ce sont *des mots qui sont présentés*. Cependant aucune différence n'existe, que le sujet dise qu'il a vu le stimulus ou qu'il dise qu'il n'a rien vu. La transition va se faire un peu plus loin, autour de 250 ms, où l'on commence à voir des activations frontales soutenues et également des réactivations dans le cortex temporal ; de nombreuses aires, à travers tout le cerveau, semblent ainsi être concernées par le traitement du stimulus lorsque le sujet, à la fin de l'essai, dit qu'il a bien vu le stimulus, alors que dans le cas où il dit qu'il n'a pas vu le stimulus ces réactivations n'ont pas été déclenchées.

**Comment interpréter ces résultats ?**

Un des modèles sur lequel nous travaillons actuellement, est celui décrit et formulé par Bernard Baars, modèle par la suite « mis en neurones » pourrait-on dire, par Stanislas Dehaene et Jean-Pierre Changeux, et après eux par plusieurs autres collaboratrices et collaborateurs, dont l'auteure de cet article. Le scénario de prise de conscience, tel qu'on peut l'interpréter maintenant, serait le suivant (cf *figure 10*) : les premiers traitements sensoriels initiaux de l'information, les 200 premières millisecondes, semblent correspondre à une étape de traitement préconsciente ; il n'est pas encore décidé si la personne prendra ou non conscience du stimulus et nous proposons de considérer que le facteur qui déclenche la prise de conscience du stimulus est la réactivation, notamment dans le système attentionnel, de ces représentations qui sont initialement locales. Si cette réactivation est couronnée de succès, elle permettra de conduire à un état particulier où il existe un partage global de l'information sensorielle que le sujet ressentira comme *accès conscient*.

Si on représente cette dernière étape sous la forme d'un réseau à plat (*Figure 11*), l'étape préconsciente correspondrait à une représentation qui emprunte les autoroutes automatiques (dans des modules périphériques, sur la figure) du traitement de l'information, mais qui n'ont pas de fortes connexions à longue distance. Mais à partir du moment où cette représentation connecte un « hub » , une espèce d'échangeur qui a des connexions beaucoup plus riches à travers différents modules, cette information va pouvoir être partagée beaucoup plus largement, par exemple au niveau des aires du langage pour pouvoir la rapporter verbalement, au niveau des aires d'évaluation et de planification pour intégrer cette information à la planification courante, etc. Finalement ce partage d'information est d'un point de vue neuronal et fonctionnel ce que nous expérimentons comme une accessibilité de l'information à la conscience.

Ce modèle peut donner lieu à une implémentation sur ordinateur, ce qui permet la simulation de son fonctionnement ; on peut alors étudier si on retrouve une dynamique abrupte, « tout ou rien », entre déclenchement ou non de la deuxième phase de partage de l'information. L'implémentation met en œuvre quatre niveaux hiérarchiques, depuis le

niveau sensoriel (niveaux A et B) jusqu'au niveau des « hubs » possédant une riche connectivité longue distance (réseau fronto-pariétal, niveaux C et D), Chaque module de ce réseau est constitué d'un petit circuit thalamo-cortical (cf *figure 12, partie supérieure* – **B**). Deux éléments sont importants dans la manière dont ces modules sont connectés entre eux : d'abord la connectivité ascendante qui va permettre le traitement ascendant, rapide, de l'information, implémentée par des synapses de type « AMPA » ; ensuite la connectivité descendante, extrêmement riche, en provenance des aires de plus haut niveau (niveaux C et D), implémentée par des synapses de type NMDA plus lentes et plus soutenues, qui correspondrait à ces étapes de maintien de l'information et d'établissement d'un réseau global.

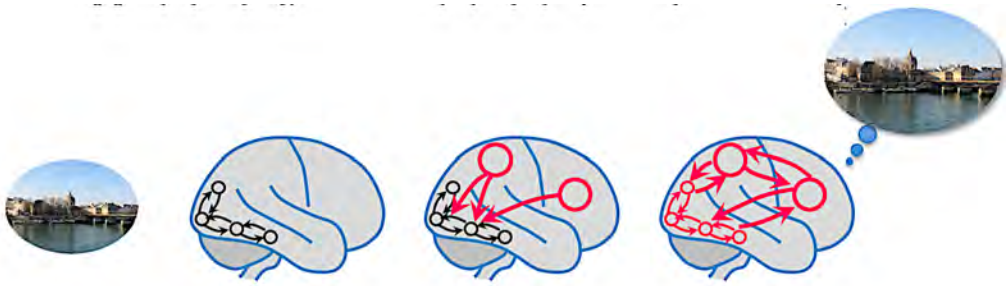


**Figure 10** : Modèle de l'espace global de travail conscient.
Images extraites de Baars 1988, Dehaene et al, *TICS 2006*

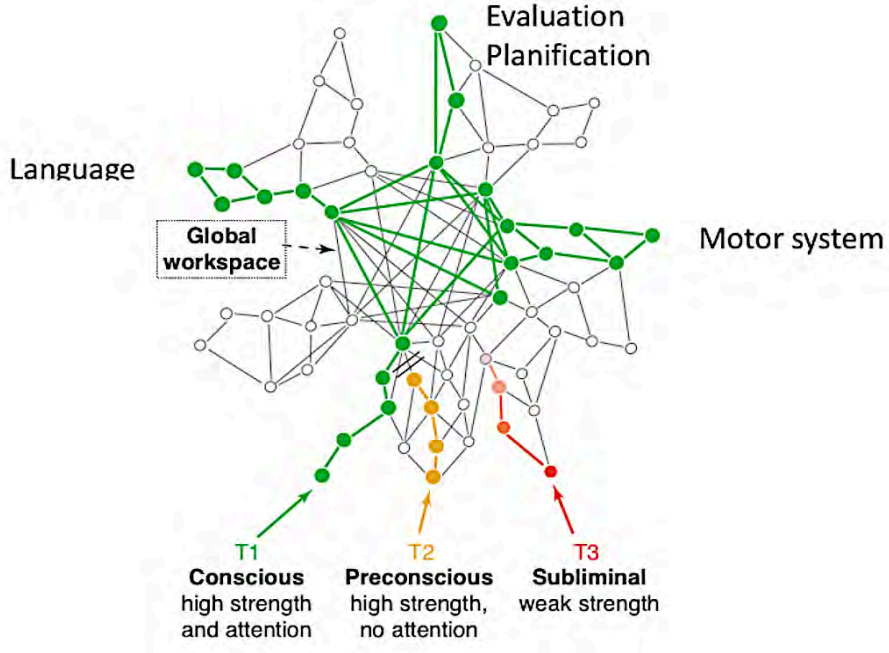



**Figure 11 :** prise de conscience : connexion à un échangeur conduisant au partage de l'information.
Images extraites de Baars 1988, Dehaene et al, *TICS 2006*

Par le biais de cette implémentation une fois mise en place, on simule les expériences faites en EEG : on entre une première cible (au niveau de l'aire A1), puis une seconde cible (au niveau de l'aire A2) et l'on regarde comment ces deux activations se propagent dans le réseau (cf *figure 12*, partie inférieure - **C**). Ce que l'on constate, c'est bien l'existence de deux groupes de réactions du système à une même stimulation : dans certains essais on voit l'information de la deuxième cible monter jusqu'aux dernières étapes et puis s'affadir et cesser assez rapidement. Dans d'autres essais la même stimulation va provoquer une dynamique très différente : une montée de l'information *puis son maintien au niveau le plus haut*, accompagnée d'une redescente de l'activation ; c'est cette redescente de l'activation portée par la connectivité descendante qui permet cette deuxième étape de traitement, partagée cette fois par tout le réseau et maintenue sur une très longue période, au-delà de 300 millisecondes après la présentation du stimulus.

# 5. Vers des signatures de la conscience

Il serait particulièrement intéressant en clinique humaine d'avoir une signature électrophysiologique de la prise de conscience, notamment pour le diagnostic des patients non-communicants, comme les patients végétatifs ou minimalement conscients. Ce sont des patients qui, à la suite d'un coma dû à un traumatisme crânien, une anoxie cérébrale (ou autre), vont ouvrir spontanément les yeux mais avec lesquels on a du mal à communiquer et chez qui l'examen clinique seul ne permet pas de déterminer si ils ont conscience d'eux même ou de leur environnement. C'est un problème très actuel (par exemple l'affaire Vincent Lambert) et les praticiens médicaux ont vraiment besoin d'outils qui viendraient des connaissances scientifiques actuelles sur les mécanismes de prises de conscience, pour informer le diagnostic et le pronostic de ces patients.

## 5.1 Connectivité fonctionnelle et diagnostic des états de conscience

Un exemple de connaissances théoriques qui ont été transférées dans la pratique clinique et qui est utilisé maintenant en routine pour informer le diagnostic de ces patients, notamment chez Jacobo Sitt et Lionel Naccache à l'hôpital de la Pitié-Salpétrière (Paris), est cette propriété de communication longue distance entre des aires cérébrales distantes, très associée à la prise de conscience.

La méthode employée consiste à réaliser un EEG de l'activité spontanée et voir s'il existe une communication, un transfert d'information à longue distance, sur des électrodes éloignées. C'est bien le cas pour des sujets sains, c'est aussi le cas pour des sujets dont on sait qu'ils sont conscients mais qui sont paralysés, par contre ces communications longues distances disparaissent chez des patients végétatifs dont on sait qu'ils sont dans un état grave (il n'y a plus que des communications locales, cf *figure 13*). Mais ces résultats ont été obtenus avec l'activité spontanée, c'est-à-dire qu'ils permettent de détecter si les capacités de communication fonctionnelle sont préservées chez un patient, ce qui est une condition nécessaire pour pouvoir prendre conscience de son environnement. Cependant

elles ne permettent pas d'affirmer avec certitude que le patient prend bien conscience de son environnement extérieur. Peut-on aller plus loin et lire dans le cerveau la prise de conscience d'un stimulus particulier ?

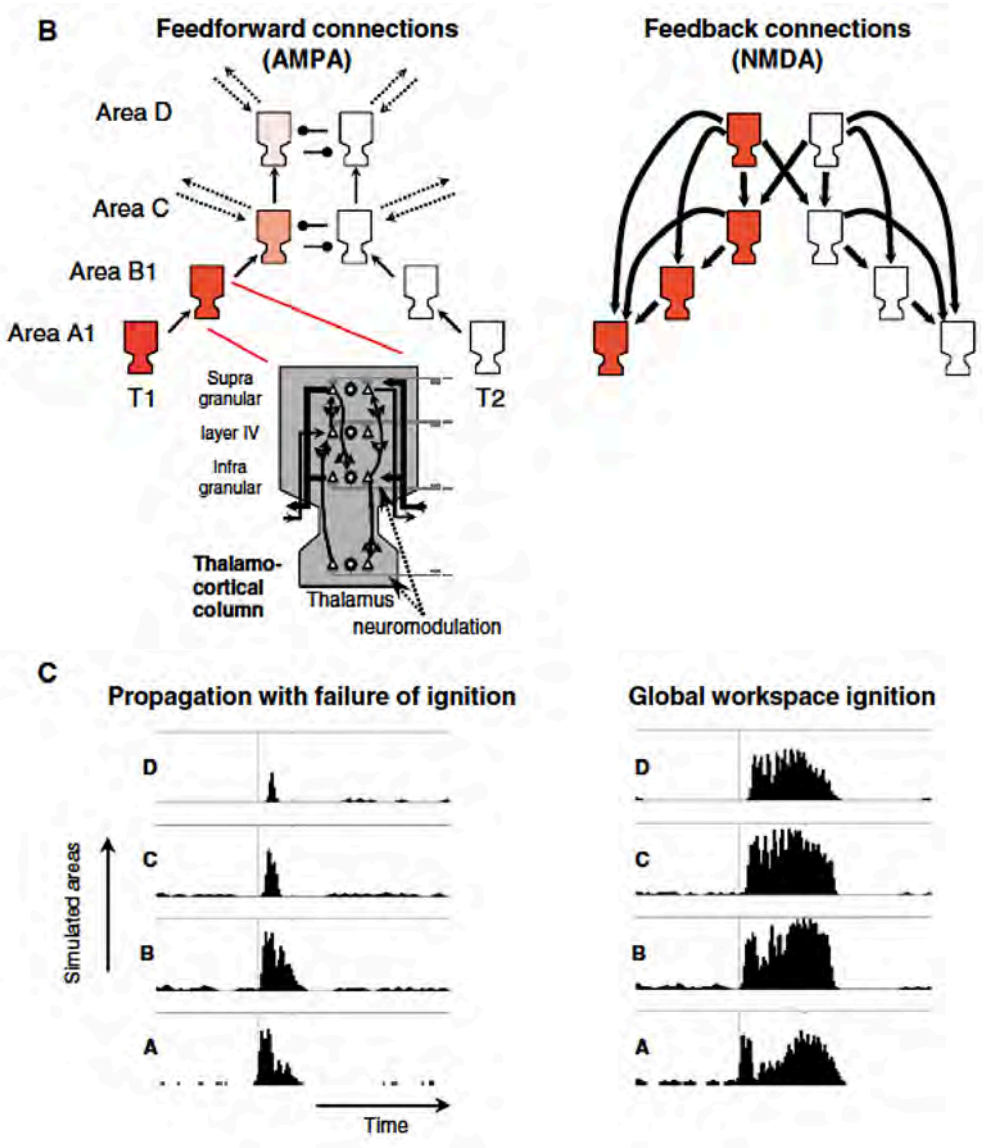



**Figure 12** Implémentation du modèle neuronal (**B**) et résultats des simulations (**C**) montrant le phénomène de tout ou rien. Images extraites de Dehaene, Sergent, Changeux, Pnas 2004 et Dehaene et Changeux, Review Neuron 2010

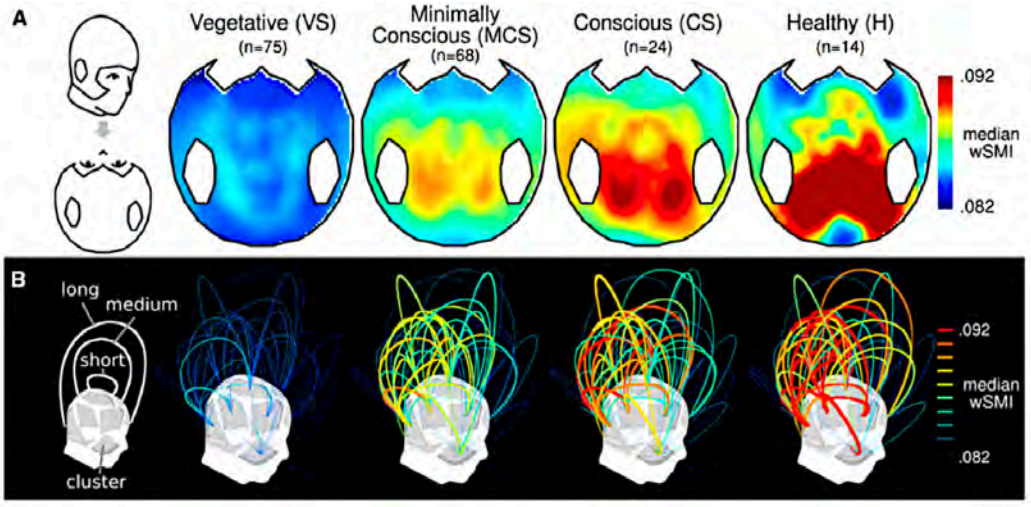



**Figure 13** : Connectivité fonctionnelle et diagnostic de l'état de conscience. Image extraite de King, Sitt et al. Current Biology 2013

### 5.2 Repérer une dynamique de saut dans le signal cérébral

Nous présentons ici des résultats d'une étude non encore publiée à ce jour (mars 2018) et qui vient d'être soumise à une revue. Dans cette étude, on a cherché à mettre en évidence ce comportement d'activité cérébrale de type *tout ou rien* rencontré dans les expériences d'EEG spontané, mais cette fois-ci en réaction à des stimuli auditifs de différents niveaux, plus pratiques à utiliser chez les patients : des stimuli auditifs de très faibles niveaux, donc inaudibles, puis des stimuli au seuil d'audibilité (parfois conscients parfois non), et enfin des stimuli très audibles. On parcours ainsi une gamme d'intensités sonores, à différents niveaux de Signal/ Bruit : -13dB, -11dB, -9db ( seuil), -7dB et -5dB, et on tente de voir si cette montée en gamme nous permet de détecter une dynamique non-linéaire. La *figure 14* ci-dessous montre des simulations de deux types de dynamique que l'on pourrait observer dans le signal cérébral en réponse à ces stimuli de différentes intensités.

Dans un modèle classique de *dynamique unimodale*, l'activité cérébrale globale augmente avec la force du stimulus autour d'une distribution unimodale, c'est-à-dire que, pour un même niveau de stimulation sonore, les variations d'activité à travers les essais restent centrées autour d'une même moyenne (un seul « mode » d'où le terme unimodal). C'est cette moyenne qui augmente avec la force du stimulus (Fig 14 partie gauche). Ici, nous envisageons aussi un autre modèle, inspiré des observations précédentes qui suggèrent des phénomènes de saut dynamique autour de la prise de conscience. Ce deuxième modèle est celui d'une *dynamique bimodale*.

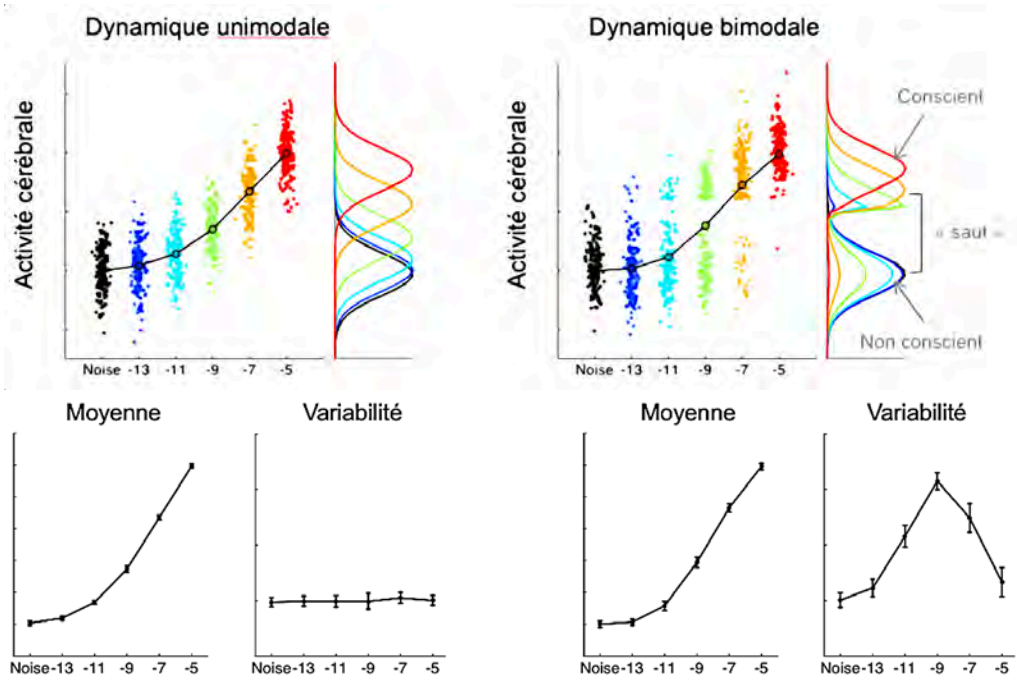



**Figure 14** : Existence de deux types de dynamiques en réaction à une même gamme de stimuli. Image extraite de Sergent, Labouret et al. In Prep.

Dans ce second type de dynamique, certes, l'activation augmente en moyenne avec la force du stimulus, mais, autour du seuil de conscience, dans certains essais l'activité cérébrale est élevée (on a un saut d'activité) alors que dans d'autres essais elle reste faible, le patient n'ayant pas pris conscience du stimulus dont l'intensité était cependant la même. Aux niveaux de stimulation « seuil », la distribution de l'activité à travers les essais est donc « bimodale » : elle oscille autour de deux moyennes ou deux « modes ». D'après nos études précédentes, c'est ce type de dynamique qui devrait constituer une signature de conscience.

Comment pouvons-nous distinguer ces deux dynamiques cérébrales dans des enregistrements EEG sans avoir recours au rapport comportemental du sujet ? On constate que si on traite les enregistrements EEG de la manière classique, c'est-à-dire en moyennant l'activité à travers les essais pour un même niveau de stimulation, les deux modèles donnent des courbes de réponse similaires et sont donc indistinguables. Par contre, si l'on s'intéresse à la variabilité du signal à travers les essais, les deux modèles font des prédictions bien distinctes : dans le cas d'une dynamique unimodale, on prédit une évolution monotone de la variabilité du signal en fonction de la force de stimulation ; dans le cas d'une dynamique bimodale, on prédit un profil très différent avec une explosion de la variabilité à travers les essais pour les stimulations seuil, qui produisent deux types d'essais (conscients et non-conscient). Au-delà et en deca du seuil de détection, cette haute variabilité disparait parce

que (au-delà du seuil) dans tous les essais le stimulus est entendu ou parce que (en deça du seuil) dans tous les essais le stimulus est inaudible.

Cette modélisation simple nous permet donc d'identifier ce profil de variabilité comme étant une bonne signature de dynamique bimodale. Nous allons donc tester si une telle signature se retrouve dans le signal cérébral enregistré chez des participants sains à l'éveil.

### 5.3 Détection des différentes dynamiques d'un signal cérébral

Sommes-nous capables de détecter ces faits en observant l'évolution dans le temps de l'activité cérébrale globale, une fois le stimulus émis. Dans un premier temps nous avons regardé à quels moments l'activité cérébrale globale augmente le plus avec la force du stimulus, reflet uniquement du traitement du stimulus, et dans un second temps regardé les moments où l'activité cérébrale globale montre ce pattern non-monotone très typique de variabilité inter-essais. C'est ce que nous avons fait chez des sujets sains, et le résultat est montré sur la *figure 15a* ; l'origine des abscisses est l'instant où l'on présente le stimulus auditif ; la courbe verte montre la période où le stimulus est traité. Cette période est assez longue, supérieure à une seconde [2]. On voit par contre, sur la courbe rouge, que la période d'accès conscient, elle, est plus restreinte : elle démarre de manière abrupte autour de 0.2 secondes après la présentation du stimulus, atteint son apogée à 0.5 seconde après la présentation du stimulus, mais s'arrête autour de 0.8 seconde. Cette période plus restreinte correspond, au niveau cérébral, à l'activation de tout un réseau de traitement de l'information, avec au centre le cortex auditif, puisque ce sont des stimuli auditifs, et autour, un espace global de travail conscient. Ce qui est intéressant c'est que cette signature peut-être vue chez chaque individu : la figure 15a montre ainsi l'exemple de deux individus identifiés comme « Subject 10 et Subject 14 », où l'on voit cette signature d'accès conscient en rouge.

On est ainsi capable de lire cette dynamique sans trier les essais en fonction de ce qu'a dit la personne ; mais c'est une situation où il avait été demandé préalablement aux sujets de faire un traitement de cette information. Aussi nous nous sommes ensuite placés dans le cas de personnes prenant juste spontanément conscience du stimulus, sans consigne de traitement de l'information apportée : on place les sujets dans une situation où ils reçoivent les mêmes stimuli, mais où ils n'ont rien à faire de spécial sur ces stimuli, ils réalisent d'autres tâches sans rapport, comme par exemple des « quizz ».

Il est évident que ces sujets vont spontanément prendre conscience de ces stimuli, en tout cas les plus audibles. Est-ce que, dans ce cas apparaît toujours sur la courbe rouge la bi-modalité, la variabilité inter-essais autour du seuil de conscience ?

---

[2] par exemple, au bout de 1,4 secondes, il n'y a plus aucun lien entre l'activité cérébrale et l'intensité du stimulus, alors que ce lien atteint sa valeur maximale – pente de régression > 0.2 - autour d'une demi-seconde

La réponse est oui. On constate effectivement toujours une grande période de traitement du stimulus (courbe verte de la *figure 15b*), alors même que l'information apportée n'a pas à être traitée dans un but spécifique, même pas pour être rapportée à l'expérimentateur ; on constate aussi une période d'accès conscient, mais plus restreinte (courbe rouge de la figure 15b) : Le sujet a pris conscience de ce stimulus mais, comme il n'a pas eu à l'utiliser pour une tâche, cette période est plus restreinte.

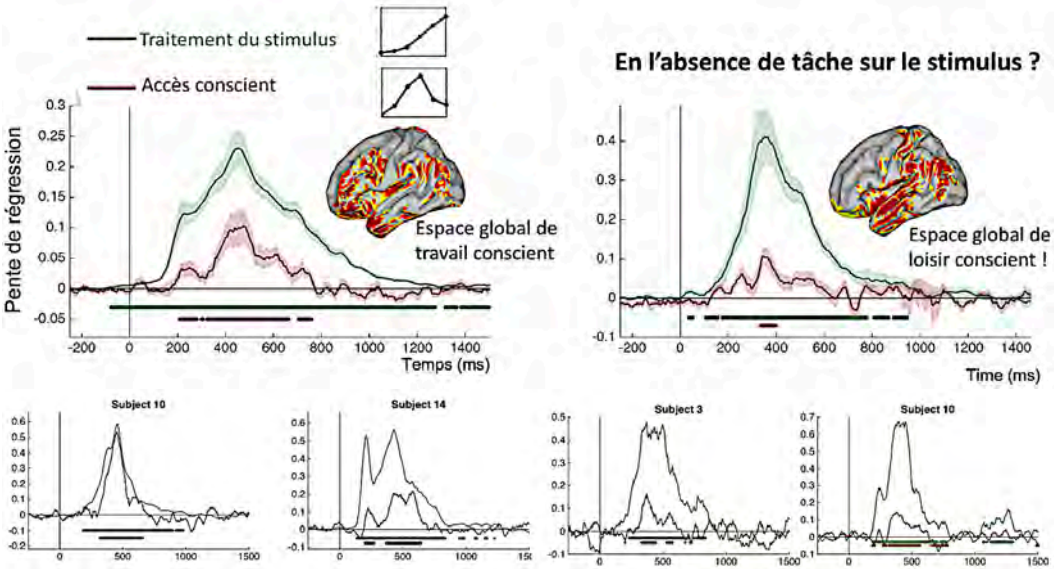



**Figure 15** : Détection des différentes dynamiques dans le signal cérébral. Images extraites de Sergent, Labouret et al. , in Prep

Si maintenant on reconstruit l'activité qui se produit, à ce moment de perception consciente**,** lorsque les sujets n'ont aucune tâche à faire sur le stimulus, on voit encore une fois un réseau apparaître, mais alors les activations fortes préfrontales disparaissent : une espèce de réseau existe bien encore, mais le cortex préfrontal joue un rôle moins important puisqu'il n'y a pas de tâche à effectuer sur le stimulus. Ainsi, si on appelle *espace global de travail conscient* le réseau qui se manifeste dans le cas où un travail d'interprétation du stimulus existe, peut-être pourrions-nous parler ici d'un *espace global de loisir conscient*, un partage de cette information juste pour jouer avec cette information, sans but spécifique.

## 6. En conclusion

Ce tour d'horizon montre que l'on connait déjà pas mal de choses sur les bases neurobiologiques de la conscience. On sait qu'il y a une grande part de notre activité cérébrale même à l'éveil qui est non consciente et que certaines informations semblent être

sélectionnées, probablement par le système attentionnel, pour être traitées consciemment ; que cela entraine une augmentation de l'activité sensorielle associée, des activations frontales et pariétales, les activations frontales pouvant peut-être être modulées par la tâche, une augmentation du dialogue, et donc du couplage à longue distance des différentes aires et enfin une dynamique de saut : il semble y avoir un gap de l'activité cérébrale au moment de la prise de conscience, gap qui apparait après une première phase préconsciente du traitement, et qui produit un effet relativement stable pendant quelques centaines de millisecondes, une phase d'accès conscient.

Ces 40 dernières années de recherche ont fait faire des progrès considérables en partant de l'expérience subjective et en allant vers l'activité cérébrale. Maintenant, nous sommes à un moment charnière où l'on va être capable de partir de l'activité cérébrale pour aller vers l'expérience subjective.

**Références : pour en connaître plus et mieux comprendre**

Stanislas Dehaene, Lionel Naccache et al. ; Imaging unconscious semantic priming ; Nature 395 ; 1998 ; 597-600

Stanislas Dehaene, Lionel Naccache et al. ; Cerebral mechanism of word masking and unconscious repetition priming ; Nature Neuroscience 4 ; 2001 ; 752-758

S. Sandaghiana, G. Hesselmann et al. ; Distributed and antagonist distribution of ongoing activity fluctuations to auditory stimulus detect ; Journal of Neuroscience 29 ( 42) ; 2009 ; 13410-13417

G. Rees, G. Kreiman et al. ; Neural correlate of conciousness in humans ; Nature Reviews Neuroscience 3 ; 2002 ; 261-270

in Research and Perspectives in Neurosciences, S. Dehaene et Y. Christen (Eds): Characterizing consciousness : from cognition to the clinic ? pp.55 à 84, Springer-Verlag 2011

in Consciousness and the Brain, Stanislas Dehaene (Ed) : C. Sergent and S. Dehaene ; The signature of a conscious thought ; pages 115 à 160, Penguin Books 2014

Sergent C, Baillet S, Dehaene S, ; timing of the brain events underlying access to consciousness during the attentional blink ; Nat. Neurosc. 10 ; 2005 ; 13391 – 13400

A cognitive theory of consciousness, Berard S Baass ( Ed) 1998, Cambridge University Press

# 4

# Quelles données subjectives pour l'étude du flux de conscience[1] ?

**Jérôme Sackur**

Laboratoire
des Sciences Cognitives et Psycholinguistique
(ENS/CNRS/EHESS)

## Abstract

Jérôme Sackur stresses that the study of consciousness cannot be limited to snapshots of conscious states. According to him, consciousness is a permanent flow, a "flowing river". This flow of consciousness varies constantly both in intensity and in speed. He first presents so-called "experience sampling" techniques for exploring in children, at all times, the reaction to stimuli involved in a task that these children are performing. He thus highlights, in the flow of consciousness, periods of attention deficit disorder, such as "waking daydreaming" and proposes models of such a dynamic. It is, he says, an essential characteristic of the human mind and of its consciousness, not to be systematically and permanently focused on a task but that there may exist great variations between individuals in this regard. Discussing the weaknesses and strengths of these sampling techniques, he offers others that allow him more direct access to the dynamics of attention and concentration. Applying his investigations to children with attention deficit disorder, Jérôme Sackur shows that they can have practical applications, in particular in the diagnosis of children who are inattentive at school or in life.

[1] Ce chapitre est la transcription, effectuée par Pierre Nabet et Jean Pierre Treuil, membres de l'AEIS, de la conférence de Jérôme Sackur faite au colloque organisé par l'AEIS, à l'Institut Henri Poincarré, le 15 mars 2018 ; le texte a été relu et corrigé par le conférencier. Il est publié avec son accord.

# 1. Introduction

Je vais vous parler du flux de conscience, des moyens de l'étudier, à travers deux catégories d'expériences, et de résultats obtenus. Mais auparavant, il me semble utile de rappeler quelques points.

## 1.1. Un peu d'histoire : la conscience en psychologie expérimentale

Reprenons rapidement l'histoire de la conscience en psychologie. Comme vous avez peut être dû le voir au cours de ce colloque, cela a commencé par une association naturelle entre la conscience et la psychologie, au point que la psychologie fondamentale était considérée comme la *science de la conscience* par les pères fondateurs de cette discipline (James, Ribot, ou Wundt par exemple) : la psychologie était équivalente à l'étude des états conscients. Existaient bien aussi une psychologie animale et une psychologie de l'enfant, concernant des « sujets » qui ne pouvaient pas rapporter leurs consciences, et aussi une psychologie sociale, mais ce n'était pas la Psychologie Fondamentale, la psychologie première, dirais-je.

Ensuite, il y a eu une période pendant laquelle on a tout simplement rejeté l'idée même que la conscience existe. Donc, on ne pouvait faire une science de quelque chose qui n'existait pas. *C'est la période du béhaviorisme,* pour lequel la psychologie était la science du comportement ; la conscience était une illusion, une invention politiquement fondée peut-être, mais qui n'avait pas de justification scientifique, position à l'encontre de tout ce qu'on allait faire par la suite

*Puis vint la révolution cognitive*. La révolution cognitive, vers les années 1950, définissait la psychologie, non plus comme l'étude du seul comportement, non plus comme l'étude exclusive de la conscience, mais plutôt comme l'étude du *mental*. A partir du moment où la psychologie devient une étude du mental, la conscience devient un problème parmi d'autres qu'on peut étudier, ou ne pas étudier.

Ce qui est intéressant à savoir, c'est que pendant des années, en fait, la psychologie cognitive a évité de parler de conscience sans que ce soit, pour autant, une interdiction. Ce n'était pas une nécessité comme à l'époque des pères fondateurs. Ce n'était pas non plus interdit comme à l'époque du behaviorisme. Il aurait été possible d'en parler mais pendant un certain temps on a évité de le faire parce qu'on gardait le souvenir du behaviorisme qui nous avait appris que parler de subjectivité, c'est dangereux. Personne ne parlait de conscience, sauf peut-être quelques Prix Nobel à la retraite, à un moment où on peut s'autoriser plus de choses. Quand on avait le Prix Nobel en physiologie et médecine en particulier ou en physique, on pouvait se mettre à parler de conscience parce que là, de toute façon, on n'avait rien à perdre.

De nos jours, la conscience est devenue un domaine d'étude « normal ». C'est une *question parmi d'autres*, comme celles du langage, de la mémoire, etc, au sein d'une science du mental, une science normale au sens de Kühn.

- Dans le cognitivisme, la conscience n'est ni
  - Exclue (Behaviorisme)
  - Requise (Introspectionnisme)
- "The language of information processing is neutral with respect to the conscious quality of mental events" (Baars, 1989)
- Le *contraste* entre une situation consciente et une situation inconsciente devient une variable expérimentale

Cette nouvelle orientation a donné lieu à des avancées tout à fait magnifiques sur la question de la *conscience perceptive* grâce à l'application de la méthode contrastive.

### 1.2. La méthode contrastive

La méthode contrastive consiste à trouver deux situations expérimentales jumelles, très proches l'une de l'autre, mais qui diffèrent par un élément qui fera que dans une des deux situations le ou la participante à l'expérience est consciente du stimulus, alors que dans l'autre situation, qui est presque identique, le ou la participante n'en n'a pas conscience. Du point de vue physique, du point de vue physiologique, la différence entre les deux situations est tout à fait minime, mais dans l'une de ces situations, il y a perception consciente, dans l'autre, non.

Qu'est que cela a permis ? ce sont par exemple, ces études de Stanislas Dehaene que Claire Sergent a poursuivies ensuite, avec cette idée qu'on va contraster des situations où un symbole est masqué et pourra être démasqué. Dans Dehaene et collab. 1998[2] on présente à un ou une personne participant à l'expérience (Figure 1) un premier masque qui contient une série de caractères sans signification pendant 71 ms (millisecondes), puis arrive un stimulus qui est présenté pendant 43 ms (ici NINE, le nombre 9), un autre masque est présenté avec une autre série de caractères sans signification pendant 71 ms et enfin est présenté un stimulus clairement identifiable (ici le nombre 6) pendant 200 ms. Si le premier stimulus (9) est présenté dans les conditions que je viens de décrire, il n'est pas consciemment perçu ; mais il suffit d'écarter légèrement les masques, de faire en sorte de mettre un écran blanc entre le premier masque et le chiffre 9 pour que ce dernier devienne clairement perçu. Que le sujet prenne conscience ou que le sujet ne prenne pas conscience de ce nombre, quasiment toutes les choses sont égales par ailleurs. La stimulation physique, le nombre lui-même est présenté pendant la même durée, mais la prise de conscience a été modifiée. Qu'est-ce qui se passe dans le cerveau à ce moment-là ?

Si l'étude précédemment citée se concentre sur l'impact de la perception inconsciente du nombre masqué, d'autres études de Dehaene, toujours basées sur le masquage/démasquage et que Claire Sergent a présentées, analysent bien les différences.

[2] Imaging unconscious semantic priming, Dehaene et al. Nature Vol. 395. Oct. 1998

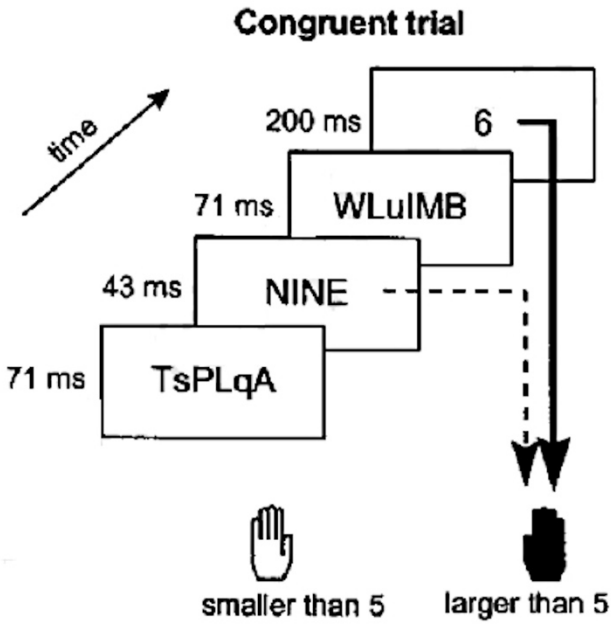



**Figure 1** montrant le dispositif de masquage (Stanislas Dehaene et al. 1998)

Ainsi dans Dehaene 2001[3] on montre qu'il y a effectivement des différences d'activation cérébrale dans les potentiels évoqués donc des modifications de l'EEG (Electro-Encéphalo-Gramme), lorsque le mot est masqué ou, au contraire, lorsqu'il est reconnu, lorsque le sujet en a conscience. Ces différences d'activation cérébrale apparaissent également en IRM Fonctionnelle (Figure 2).

### 1.3. Wilhem ou William ?

Tous ces résultats sont magnifiques et ont fait grandement avancer les connaissances dans le domaine de la conscience perceptive. Mais, dans un certain sens, ils posent la conscience comme étant un phénomène instantané. Ou plutôt comme *une succession d'instantanés* (au sens photographique du terme) qu'on examine séparément.

Il y a certainement dans la conscience un aspect qui va au-delà de cette succession d'instantanés. La conscience est aussi quelque chose comme un flux. Ces deux manières de voir sont bien illustrées respectivement chez ces deux pères fondateurs de la psychologie que sont *Wilhem* Wundt et *William* James.

[3] Cerebral Mechanisms of word masking and unconscious repetition priming, Dehaene et al. *Nature Neuroscience* Aug. 2001

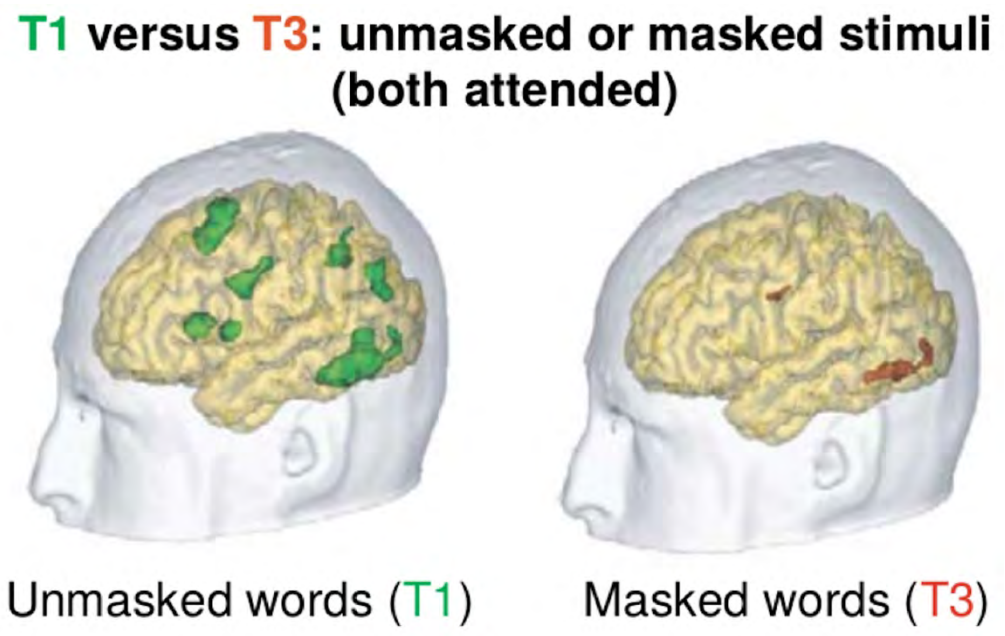



**Figure 2** montrant les modifications d'activation cérébrale mises en évidence dans *(Stanislas Dehaene et al. 2001)*

Pour Wundt la conscience est d'abord un *champ* avec un contour représentationnel, contour qui peut changer d'un instant à l'autre : quelque chose qui est conscient à un certain moment peut ne plus l'être l'instant d'après, où quelque chose d'autre peut parvenir à la conscience ; l'opposition est nette avec James pour qui étudier la conscience, c'est étudier un *flux* qui a une dynamique particulière.

Il y a un aspect particulier dans l'évolution de ces représentations, qui est la structure même de cette évolution ; structure qui est intrinsèquement importante pour comprendre la conscience. Ce n'est pas simplement le moment de perception, mais c'est aussi *la manière dont le contenu se transforme.* Pour faire comprendre cela, je vais reprendre ici quelques citations de James. Ça fait vingt ans que je réfléchis à cette question, mais c'est seulement cette année que je m'aperçois que j'arrive à répondre aux interrogations que ces citations de James suscitent.

En effet, James écrit dans *Les principes de la psychologie (1890)*, que la conscience n'est pas divisée en petits morceaux, des termes tels que chaine ou train ne décrivent pas de manière adéquate, ils décrivent seulement la manière dont elle se présente à elle-même en premier lieu. La conscience n'est pas une chose faite de moments à *jointer*, c'est un flux, un flot, une rivière, un *stream*, toutes métaphores bien plus adaptées pour la décrire. Dans le même chapitre sur le flux de conscience, James continue et il insiste sur un point important pour ce que je vais vous montrer plus loin ; c'est cette idée qu'en fait le flux de

conscience *est plus ou moins rapide* selon les moments et *qu'il a un rythme.* Et donc que si on considère la conscience seulement comme une succession d'instantanés, on perd l'idée qu'elle possède un rythme intrinsèque. Il faut aussi prendre ce rythme en compte.

À la fin de la citation James introduit deux termes qui sont fondamentaux à mon avis, ce sont les idées *d'état transitif et d'état substantif.* James dit : « Il y a des moments dans la conscience où elle s'établit dans un certain contenu ». Ce sont des moments de repos et il les appelle : « des moments substantifs ». Puis, il y a des moments de transition lorsque la conscience passe à un autre contenu et c'est ce qu'il appelle : « des moments transitifs » et en fait notre conscience est une succession de moments substantifs et de moments transitifs : si on veut comprendre ce que c'est la conscience, il ne faut pas simplement prendre des instantanés dans ce flux mais il faut comprendre aussi la dynamique de ces alternances de moments substantifs et de moments transitifs.

La question est maintenant de savoir comment on va étudier ça. C'est compliqué et je pense qu'on est très, très loin d'avoir une bonne technique pour de telles études. Je vais vous présenter d'abord une approche que nous avons mise en œuvre dans un premier temps, basée sur des techniques qui se sont révélées inadéquates, car leurs résultats peuvent être interprétés de plusieurs façons et j'expliquerai précisément pourquoi. Je vous présenterai ensuite d'autres techniques appliquées à des données toutes nouvelles, qu'on vient juste d'acquérir et qui, peut-être, permettront de faire quelque chose de plus adéquat.

- La conscience n'est pas seulement un *champ* (analogue au champ visuel) — Wundt (1878)
- C'est aussi un *flux* — James (1890) :

*Les citations de William James 1890*

« Consciousness, then, does not appear to itself chopped up in bits. Such words as 'chain' or 'train' do not describe it fitly as it presents itself in the first instance. It is nothing jointed; if flows. A 'river' or a 'stream' are the metaphors by which it is most naturally described. In talking of it hereafter, let us call it the stream of thought, of consciousness, or of subjective life. » (James, *Principles of Psychology*, 1890)

"[...] the successive psychoses [*représentations*] shade gradually into each other, although their rate of change may be much faster at one moment than at the next.

[…] When the rate is slow we are aware of the object of our thought in a comparatively restful and stable way. When rapid, we are aware of a passage, a relation, a transition from it, or between it and something else. […] Like a bird's life, [our stream of consciousness] seems to be made of an alternation of flights and perchings. The rhythm of language expresses this, where every thought is expressed in a sentence, and every sentence closed by a period. The resting-places are usually occupied by sensorial imaginations of some sort, whose peculiarity is that they can be held before the mind for an indefinite time, and contemplated without changing; the places of flight are filled with thoughts of relations, static or dynamic, that for the most part obtain between the matters contemplated in the periods of comparative rest.

Let us call the resting-places the 'substantive parts,' and the places of flight the 'transitive parts,' of the stream of thought. It then appears that the main end of our thinking is at all times the attainment of some other substantive part than the one from which we have just been dislodged." (W. James, *Principles of Psychology*, 1890)

## 2. Approche du Flux de Conscience par échantillonnage d'expériences

Je vais parler ici d'études qui utilisent le paradigme de la « *rêverie éveillée* », à savoir le fait que quelle que soit la chose qu'on est en train de faire, il y a des moments où on est, brusquement, en train de penser à quelque chose *qui n'est pas* ce à quoi on devrait être en train de penser à cet instant. Quand je dis « on devrait être en train de penser », c'est, bien sûr, toujours relativement à une tâche particulière. On peut discuter du fait de savoir s'il n'y a pas toujours quelque chose à quoi on devrait être en train de penser ; mais disons qu'en laboratoire, on peut opérationnaliser en demandant aux gens de faire quelque chose de précis. A vrai dire, cette situation-là nous arrive souvent ; j'aime beaucoup ce tableau du 19ème siècle où on voit une dame qui devrait peut-être jouer du piano ou qui devrait peut-être lire un livre, mais qui finalement est en train de faire tout à fait autre chose.

## 2.1 Etudes sur la rêverie éveillée

Commençons par introduire la question de la dynamique du flux. Pour l'étudier, la technique la plus simple est de revenir à l'interrogation de base en psychologie, c'est-à-dire de demander aux gens : à quoi penses-tu ? dans la pratique, on fait exécuter une tâche par des personnes, tâche qui peut durer pendant trente, voire quarante-cinq minutes. En général, on choisit des tâches assez ennuyeuses, des tâches rébarbatives et on interrompt ces personnes de manière aléatoire. On essaie de bien répartir les sondes par lesquelles on les interrompt. On les interrompt donc de manière aléatoire, et on leur demande, au moment où on les interrompt, de rapporter le contenu mental juste à l'instant où on les a sondées

Vous voyez bien ici la différence par rapport aux expériences sur la conscience évoquées dans l'introduction, utilisant la méthode contrastive. Dans la plupart de ces expériences on soumet les personnes à un stimulus, et on leur demande s'ils en ont pris conscience. On leur demande par exemple de se concentrer sur une tâche, puis on leur montre, ou on leur fait entendre quelque chose, puis on les interroge sur la prise de conscience de cette chose-là.

Dans ce nouveau type d'expérience, c'est différent. On place les personnes dans une tâche, mais on ne les interroge pas sur la conscience qu'ils ont eu de tel ou tel évènement. On leur demande, à des instants aléatoires, de rapporter *leur contenu de pensée* pour savoir dans quel état ils se trouvent à cet instant ; avec l'idée qu'on pourra ainsi essayer de reconstruire la dynamique de leur conscience au cours de la réalisation de la tâche. Cette technique s'appelle « *de l'échantillonnage d'expérience* » parce qu'en fait, on va échantillonner le contenu mental de manière plus ou moins libre, soit en laboratoire, soit dans la vie quotidienne en donnant aux gens un *beeper* qui peut sonner à des moments donnés sur leur téléphone portable, par exemple.

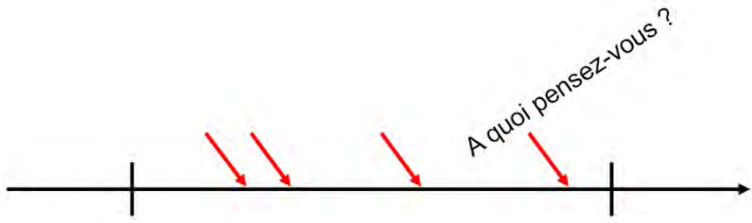



**Figure 3** : Comment étudier le flux de conscience ?

On a fait une telle expérience récemment et on a recueilli les réponses des personnes. La « rêverie éveillée » étudiée ainsi, montre que dans à peu près 30 % du temps, les gens vous disent qu'ils ne pensaient pas à ce qu'ils étaient censés être en train de faire. Ils lisent, par exemple, ou encore ils regardent un film, ils écoutent une conférence, mais en fait, ils sont en train de penser, assez souvent donc, à tout à fait autre chose.

On ne peut pas ne pas considérer que ce phénomène puisse avoir des conséquences pratiques. De fait, depuis 15 ans, une série d'études sur la rêverie éveillée montre l'existence de ces conséquences pratiques, parfois dramatiques. Par exemple si on regarde les accidents de voiture, on s'aperçoit que dans le nombre de victimes d'accidents, la proportion de personnes impliquées qui étaient en état de rêverie éveillée est très importante[4]. Le phénomène de rêverie éveillée a également un impact important dans le domaine de l'aéronautique.

Cette question est illustrée, en creux pourrait-on dire, par l'exemple des voitures autonomes. Les techniques en jeu ne cherchent pas à simuler l'esprit humain, car justement, si vous concevez une voiture autonome vous n'avez pas envie d'implanter la rêverie éveillée, vous avez envie que l'algorithme se focalise sur sa tâche de manière permanente, sans se laisser distraire. Or, c'est un caractère essentiel de l'esprit humain de faire le contraire : *un trait fondamental de la conscience humaine est bien qu'elle ne peut pas se concentrer de manière ininterrompue.*

### 2.2 Etudes sur les troubles de l'attention

Je vais maintenant parler d'une étude que nous avons faite sur des enfants qui ont un trouble de l'attention. Ce, avec l'objectif de savoir comment est organisé le flux de conscience chez de telles personnes, et plus généralement, s'il existe des différences interindividuelles dans l'organisation du flux de conscience. Dans cet exposé, mon but est également d'illustrer à nouveau le principe d'échantillonnage et d'introduire un problème intéressant dans la conception « Jamesienne » du flux de conscience.

Cette étude porte sur des enfants qui ont un déficit de l'attention avec hyperactivité, en anglais « Attention Deficit / Hyperactivity Disorder » (ADHD) ou en français TDAH « Trouble Déficit de l'Attention avec Hyperactivité » . Ce sont des enfants qu'on appelle des hyperactifs, qui ont des problèmes à l'école notamment parce qu'ils ne peuvent pas se concentrer, ils ne peuvent pas rester en place, etc. Tous les enfants ont, dans une certaine mesure, ce trait de caractère par rapport aux adultes, mais chez certains enfants ce trait peut devenir pathologique ; ça devient alors véritablement un problème, se manifestant sur le plan scolaire.

Avant de présenter quelques résultats (Figure 4), décrivons brièvement la méthode employée.

[4] Cf : Mind wandering and driving : responsibility case-control study, Cédric Galéra et al. *British Medical Journal*, dec. 2012

### *2.2.1 Méthode de l'étude*

Dans cette étude[5], réalisée avec des médecins, on a pris deux populations, à l'hôpital et au laboratoire, mais ici je vais ne parler que des enfants, de 7 à 12 ans ; on fait exécuter par ces enfants une tâche de rêverie éveillée : Les enfants sont devant un écran où apparaissent des stimuli, toutes les deux secondes environ. Ces stimuli sont des chiffres, la tâche consiste à appuyer sur une touche dès qu'un chiffre apparaît, sauf si le chiffre est 3. Dans ce dernier cas, il ne faut pas appuyer sur la touche. C'est une tâche classique qu'on appelle en psychologie tâche de *Go/no-Go*. Elle est particulièrement facile mais rébarbative au sens où elle provoque quasi inévitablement de la rêverie éveillée : c'est en effet difficile de rester totalement focalisé sur elle pendant 35 minutes, ce n'est pas possible même pour des enfants très volontaristes et en plus, on avait là des enfants, qui, pour certains d'entre eux, avaient des problèmes d'attention.

Les enfants étaient donc installés devant les écrans et sondés de temps en temps à autre ; nous adressant à des enfants, nous leur avions montré des petits dessins d'oursons illustrant différents types d'états mentaux dans lesquels nous leur demandions de savoir s'ils s'y trouvaient. Cinq catégories d'états mentaux possibles associées chacune à une image d'ourson, étaient imposées a priori. Nous leur demandions de catégoriser l'état mental dans lequel ils se trouvaient à l'instant où nous les sondions, par rapport à ces cinq possibilités. Pour qu'ils puissent répondre - désigner l'image adéquate - en connaissance de cause, nous leur avions expliqué ce à quoi elles correspondaient, à savoir dans quel état mental ils pouvaient se trouver ; nous leur disions ainsi :

- Première possibilité (*On-Task Focus*, couleur bleue claire de la figure 4):
  *Vous pouvez être concentrés sur la tâche, complètement en train de l'exécuter, sans être perturbés par le monde extérieur, ni par des idées incidentes*. Dans notre jargon, l'enfant est complètement dans le *flot* de la tâche
- Seconde possibilité (*Distraction*, couleur jaune de la figure 4) :
  *Vous n'êtes plus vraiment, complètement dans la tâche, car ici c'est l'hôpital, par exemple, il y a du bruit, ce bruit vous déconcentre, vous pensez à quelque chose qui se passe autour de vous.*
- Troisième possibilité (Task-Related Interferences, couleur verte de la figure 4) :
  *Vous êtes en train de penser à la tâche mais vous n'y êtes plus vraiment. Vous vous dites par exemple « Ah, il faut que je me concentre davantage » ou encore « c'est vraiment rébarbatif, plus qu'ennuyeux » , etc.* On n'est pas dans la tâche, mais on y réfléchit.
- Quatrième possibilité (Mind Wandering, couleur bleue foncé de la figure 4):
  *Vous êtes en train de penser à vos vacances, au fait que vous serez bientôt ailleurs, ce genre de choses.* C'est la pure rêverie éveillée où la pensée n'a pas de rapport avec l'environnement ni avec la tâche.

[5] Attentional Lapses in Attention-Deficit/Hyperactivity Disorder : Blank Rather Than Wandering Thoughts. Charlotte Van de Driessche et al. *Psychological Science 28(10)* Aug. 2017

- Cinquième possibilité (Mind Blanking, couleur rouge de la figure 4):
  *Vous ne pensez à rien, il n'y a rien dans votre esprit.* C'est en effet tout à fait possible que les gens ne pensent à rien.

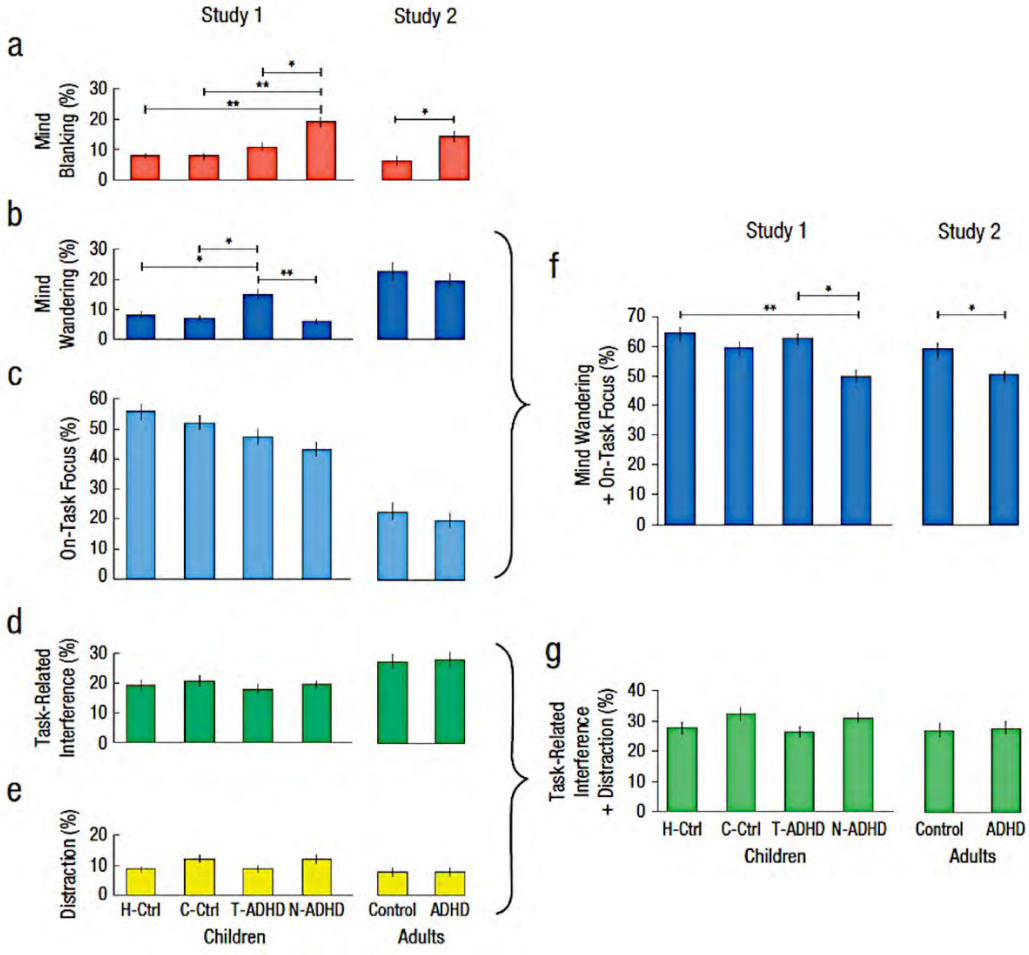



**Figure 4** Etude des troubles de l'attention. Résultats par groupe d'enfants et par possibilités de réponses, en pourcentage moyen de ces réponses dans chaque groupe ; l'étude 2 concerne les adultes, dont il ne sera pas question ici. (Van den Driessche et al, 2017)

Ces réponses possibles ont été comparées sur quatre groupes d'enfants, groupes décrits ci-dessous avec le code qui leur a été attribué, et dont les deux premiers sont des groupes de contrôle.

- Groupe « C-Crl » enfants sains, n'ayant aucun problème particulier

- Groupe « H-Crl » enfants à l'hôpital, mais qui n'y sont pas pour un trouble de l'attention ou pour hyperactivité
- Groupe « T-ADHD » enfants avec trouble de l'attention sous traitement médicamenteux (Methylphenidate)
- Groupe « N-ADHD » enfants avec un trouble de l'attention sans traitement médicamenteux.

### *2.2.2 Résultats*

La figure 4 donne les résultats obtenus pour les quatre groupes d'enfants en fonction des possibilités de réponse imposées. On y observe qu'en fait les ADHD, les enfants qui ont le trouble de l'attention se différencient des autres essentiellement par le taux de Mind-blanking, le fait qu'ils disent : « Je ne pense à rien à ce moment-là ». Cette constatation est particulièrement marquée dans le groupe des N-ADHD, sans traitement, où le taux de Mind blanking est significativement plus élevé que dans les deux groupes de contrôle. Lorsqu'on leur donne le traitement, ce taux redescend vers celui des groupes de contrôle ; par contre le taux de Mind Wandering augmente : ces enfants sous traitement ont plus de rêverie éveillée que tous les autres.

Il faut donc retenir une chose sur la symptomatologie du trouble de l'attention : lorsque les enfants avec un TDAH ne sont pas présents dans la tâche ou qu'ils ne sont pas présents à l'école, ça n'est sans doute pas parce qu'ils sont en rêverie éveillée mais peut-être parce qu'ils ont des moments de vide mental où ils ne pensent à rien et que ça leur pose un problème ensuite pour se reconcentrer. Le traitement par Méthylphenidate restaure, réduit cette proportion de blancs mais encourage la rêverie éveillée. Il y a là quelque chose de paradoxal mais qui ne l'est pas vraiment… c'est sur cet aspect que je voudrais maintenant insister, en revenant à la question du flux de conscience.

### *2.2.3. Discussion*

Le problème qui se pose en effet est : qu'est-ce que c'est que ces blancs ? En fait, c'est un peu ambigu. Nous avions pensé à cette catégorie du Mind Blanking en nous disant : « C'est possible que les enfants veuillent nous dire ça » mais nous n'avions pas vraiment théorisé, à l'origine, ce que cela pouvait être. On s'est aperçu ensuite qu'en fait, il y avait au moins trois manières différentes de concevoir une réponse à la question de ce qui relèverait de cette catégorie.

La première option, c'est que ça pourrait être des vrais épisodes de vide mental. Après tout, il pourrait se faire qu'il y ait des moments dans notre « flux de conscience » où nous n'avons pas de conscience. Il y aurait des moments du temps physique en gros, pendant lesquels il n'y a pas de conscience. Et évidemment, a posteriori, on ne les voit pas ces moments là. Rétrospectivement, ils ne sont jamais présents, puisqu'il n'y a rien, et la seule manière de s'en apercevoir, c'est lorsqu'on est sondé de manière externe. Mais si on

regarde après coup, alors on ne peut pas voir ces blancs parce qu'ils seront raboutés, agencés avec les autres moments où il y a une conscience. Bon, c'est possible qu'il y ait ainsi des moments de vide… et que l'idée de vide mental paraisse paradoxale, tout simplement parce que de l'intérieur, par introspection, ils sont forcément invisibles.

La seconde option, c'est que ça peut être aussi un déficit de métacognition justement, un déficit d'introspection. Peut-être que ces enfants, n'ont pas plus de vide mental que les autres, que ces moments de vide existent ou pas. Peut-être simplement ont-ils plus de difficulté à rapporter leurs contenus mentaux, parce qu'ils n'arrivent pas à expliciter le contenu de leur conscience. Ce n'est pas un trouble attentionnel, dans un certain sens, ou encore c'est le fait que, de manière conjointe, les troubles attentionnels soient associés à des déficits d'introspection.

La troisième option est celle qui nous intéresse le plus avec mes étudiants et étudiantes. C'est en fait l'idée que ça serait lié à la variabilité de la vitesse du flux de conscience. Peut-être que ces enfants-là, ce qui leur arrive, ce n'est pas qu'ils ont plus fréquemment des vides mentaux, ce n'est pas qu'ils ont un problème d'introspection, mais c'est qu'en fait, *ils sont plus souvent en transition*. C'est-à-dire que leur flux de conscience est plus instable et que les *moments substantifs* - au sens de James - dans lesquels ils ont un contenu de pensée relativement accessible, sont plus courts, et il y en a moins. Alors ils vont se mettre à « transitionner » plus souvent et pendant la durée de la transition ils ne pourront rien rapporter parce que rien n'est complètement stabilisé. James d'ailleurs fait allusion à cette possibilité : il y aurait en quelque sorte une plus grande probabilité que le sondage des enfants ait lieu pendant ces moments de transition.

Le problème, c'est que notre méthodologie est complètement insuffisante pour vraiment aborder cette question-là. Parce que ce qu'on fait, c'est de sonder les personnes toutes les 15/ 20 secondes, mais on ne connaît pas la fréquence des moments transitifs et des moments substantifs. On ne peut donc pas reconstruire à partir de ces sondages très, très irréguliers, très espacés, *very sparse* on dirait en anglais, on ne peut pas reconstruire l'ensemble du flux avec des données si pauvres. Nous avons bien donc là un problème d'adéquation entre la méthodologie et notre question, qui est tout à fait important.

### 2.3 Modélisation markovienne versus modélisation continue du flux de conscience

Je voudrais, à ce stade, faire une parenthèse théorique sur les modèles sous-jacents, pour montrer en quoi cette méthodologie de *l'experience sampling* est insuffisante. Ce que je dis ici, c'est que cette méthodologie est pauvre. Elle ne nous permet pas en effet de distinguer ce que j'appellerai une vision *markovienne* du train de pensée et une vision du flux de conscience comme quelque chose de continu avec cette dynamique proprement « Jamessienne ».

La version markovienne du train de pensée serait l'idée que les enfants sont, à un instant dans un état mental précis et qu'ils ont, à chaque instant, une certaine probabilité d'y rester, ou au contraire de faire une transition vers un autre état mental. Dans la figure 5, toutes les

flèches ne sont pas représentées mais on pourrait concevoir assez facilement une matrice avec des probabilités de transition d'un état mental à un autre. Ça expliquerait un train de pensée tel que celui figuré sur le même schéma où l'enfant passe d'un des cinq états mentaux possibles, - ici de rêverie éveillée, je pense à mon dernier anniversaire - à un autre, savoir l'état de concentration sur la tâche.

MW (Mind Wandering) ----------$P_{MW/F}$----------> F (Focus, On-task

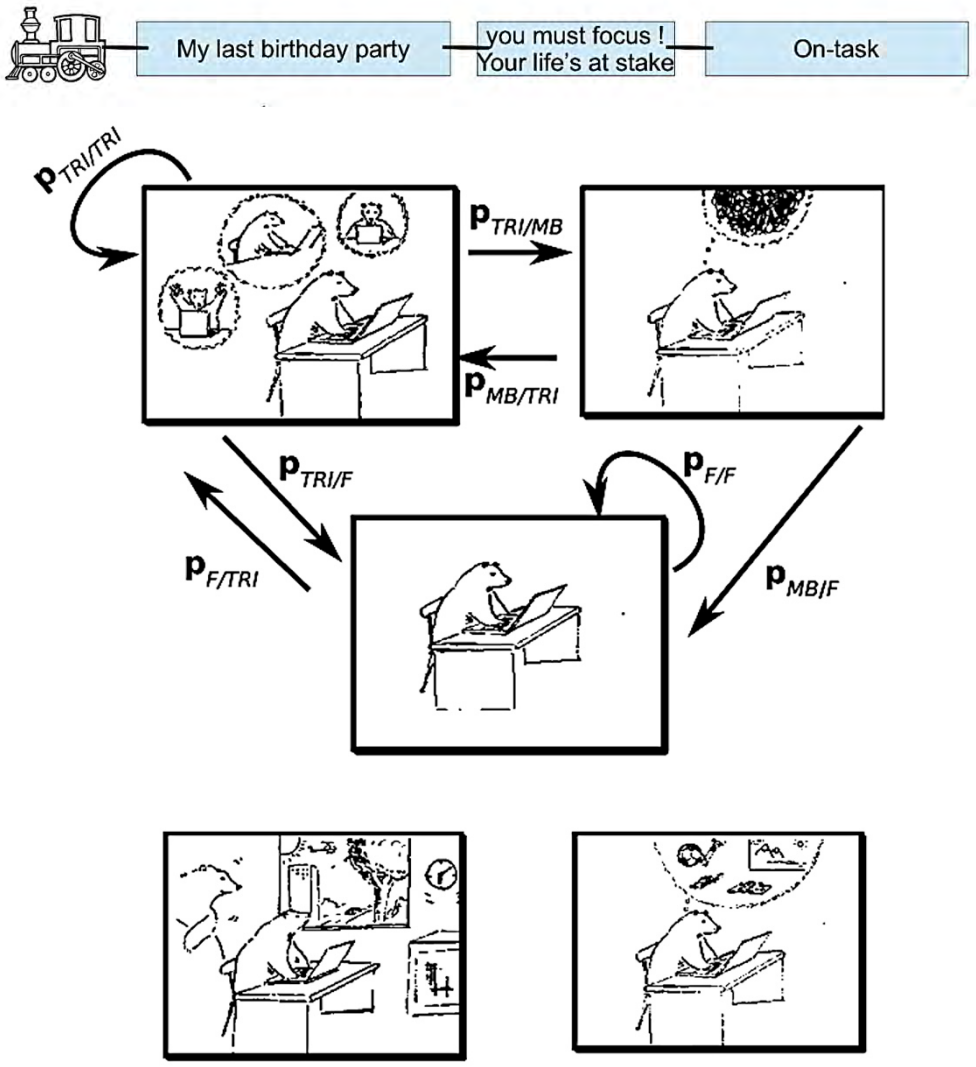



**Figure 5** Le train de pensée dans sa version Markovienne (toutes les transitions possibles ne sont pas représentées sur la figure)

Avec un étudiant, nous avions essayé de mettre en œuvre un modèle de ce genre[6] ; pour déterminer la succession des états mentaux d'un sujet impliqué dans une tâche de Go/No-go, nous nous étions basés sur le fait que, dans ces tâches, exploitées depuis des années, existe une certaine variabilité des temps de réponse aux stimulus, temps observés aux instants où ils sont interrogés sur leurs états mentaux. Cette variabilité est relativement importante et avec Mikaël Bastian, on s'était rendu compte qu'elle corrèle avec les rapports subjectifs, les états mentaux : Les périodes pendant lesquelles les temps de réponse (appui ou non appui sur la touche dans la tâche de Go/No-go) des sujets sont les plus variables, sont aussi les périodes pendant lesquelles ils ont le plus souvent répondu qu'ils sont off-task, qu'ils sont déconcentrés. Alors que les périodes pendant lesquelles ils sont réguliers dans leur temps de réponse, sont celles où ils vont répondre qu'ils sont relativement concentrés.

Nous nous étions alors basés sur cette constatation pour concevoir des modèles Markoviens avec, de manière sous-jacente, deux états possibles : On-task (OT), concentré, ou en rêverie éveillée (MW) (cf Figure 6), chacun de ces états étant associé à une distribution différente des temps de réponse, grande variabilité pour l'état MW, faible variabilité pour l'état OT.

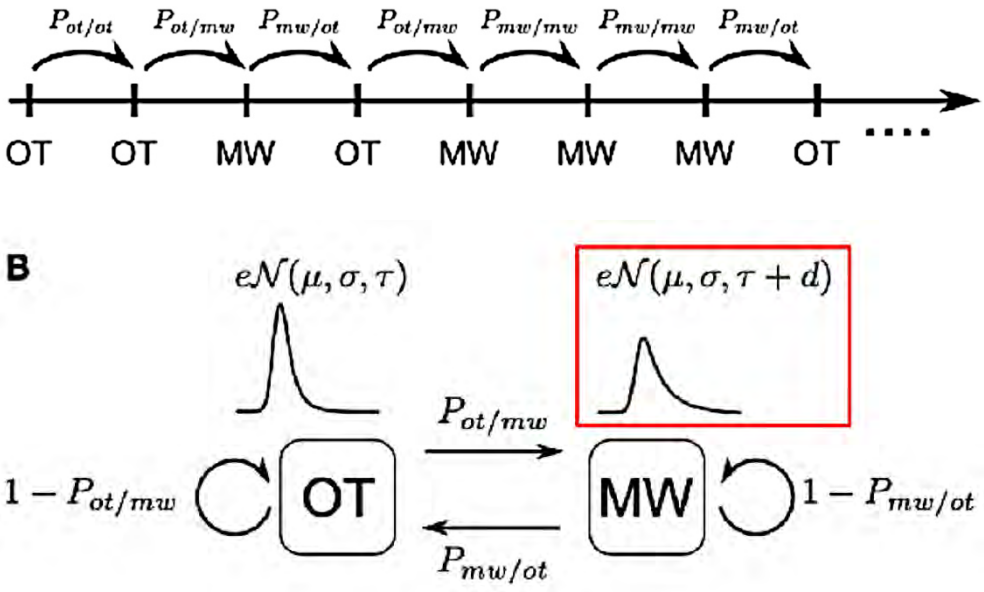



**Figure 6** illustrant un modèle Markovien binaire de transition entre deux états OT et MW (Bastian et Sackur 2013)

[6] Mind wandering at the fingertips : automatic parsing of subjective states based on response time variability. Mikael Bastian et Jérôme Sackur. *Frontiers in Psychology*, sept. 2013

Nous avions alors montré que ça marchait à peu près. En gros, les états dans lesquels on prédisait que les gens se trouvaient d'après l'analyse du modèle de Markov latent sur leur temps de réponse, c'était aussi ceux qui, lorsqu'on les sondait, correspondaient à leur réponse ON ou OFF. Mais évidemment, c'est extrêmement pauvre comme approche. C'est extrêmement pauvre parce que d'une part on se base sur les temps de réponse qui sont les résultats les plus extérieurs du travail de la pensée et d'autre part, pour valider ces modèles de Markov latents, avec une interrogation des sujets toutes les 15 secondes, on a des données qui sont totalement sous-dimensionnées. Puis on s'est rendu compte qu'on pouvait très bien construire un modèle psycho-physique expliquant exactement la même chose, basé non pas sur des états mentaux *discrets* entre lesquels il y aurait des transitions possibles, mais sur une sorte d'état mental sous-jacent animé de variations *continues* de concentration et qu'au moment où l'on demande aux personnes de répondre : êtes-vous concentrées ou déconcentrées ?, elles mobilisent un seuil de décision fixé de façon interne (Cf Figure 7). Les personnes discrétisent en quelque sorte les valeurs de concentration de leur état mental au moyen de ce seuil de décision. C'est un modèle alternatif de ce type que nous avons tenté de proposer avec Mikael Bastian et Valentin Wyart[7],

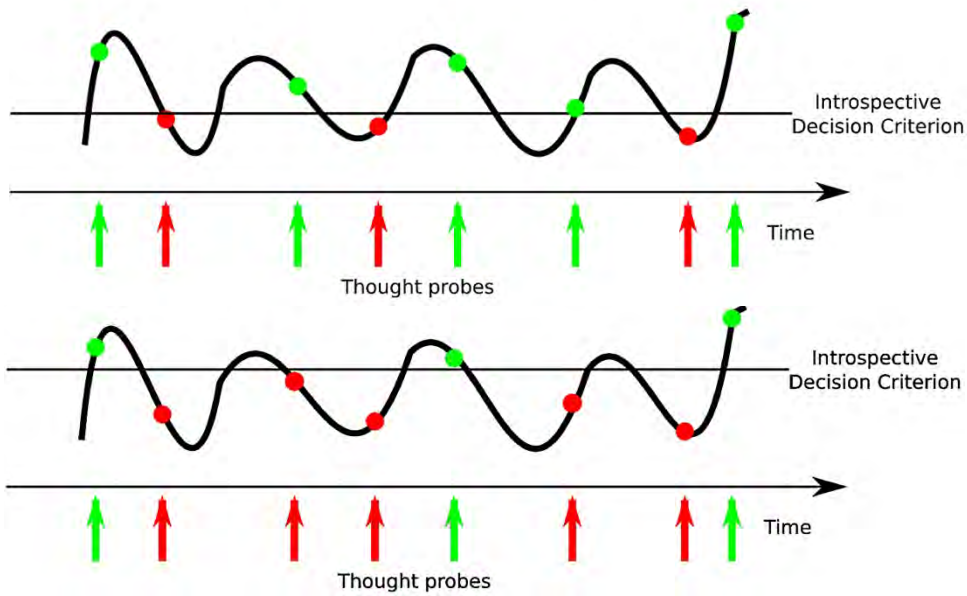



**Figure 7** : Le flux de la conscience dans sa version psycho-physique : les réponses des sujets (concentré vs. Déconcentré) dépendent d'un seuil interne de décision. (Bastian, Wyart & Jérôme Sackur, 2014)

Mais en réalité la faiblesse des données disponibles fait qu'on n'a aucun moyen de faire la distinction entre les deux interprétations. William James disait déjà cela : qu'on n'a pas un

[7] Mind wandering reports and the stream of consciousness. Bastian, Wyart et Sackur, *Psychonomics 2014 (Posters)*, Long Beach, CA

train de pensée, discret et Markovien, mais qu'en même temps, manquent les études exactes pour le démontrer.

Pour donner une image de cette situation, c'est comme si l'on essayait de deviner, de reconstruire des discussions qui ont lieu dans une maison la nuit, alors qu'on se trouve à l'extérieur et qu'on observe seulement l'allumage ou l'extinction des lumières dans différentes pièces. Face à cette espèce d'impossibilité, on s'est dit qu'il fallait aborder la question avec une toute autre méthodologie et c'est ça que je voudrais maintenant présenter.

## 3. Un accès direct au flux de conscience sans experience sampling ?

Avec Charlotte Van den Driessche[8], nous nous sommes dit : « Ce qu'il faut faire, c'est trouver des tâches où à chaque instant l'exécution de la tâche par le sujet nous renseigne sur son état mental ». On ne va plus avoir besoin de sonder le sujet de l'extérieur, on va prendre une tâche qui par sa nature nous dit quelque chose de l'état mental dans lequel le sujet se trouve à chaque sondage. Pour ça, on a pris à nouveau des enfants de 8-9 ans, une soixantaine d'enfants parce qu'il y a de grandes variabilités d'attention chez les enfants à cet âge-là, et on leur a fait faire deux tâches, une tâche essentiellement visuelle, de « barrages de cloches » parmi un ensemble de dessins, et une tâche verbale dite de « fluence sémantique », en l'occurrence nommer le plus grand nombre d'animaux. Par ailleurs, avec l'aide de leurs parents et leurs enseignants, les enfants ont été notés sur leur capacité attentionnelle. On a ainsi les enfants dont on sait qu'ils sont très bons pour se concentrer et d'autres enfants qui ont des problèmes de concentration. On utilise la même échelle que celle utilisée pour étudier l'ADHD. On dispose donc d'une sorte de gradation de la capacité attentionnelle de tous ces enfants ; l'idée sous-jacente est que ces capacités vont corréler avec le style de recherche que les enfants vont déployer dans le domaine visuel comme dans le domaine sémantique.

### 3.1 Expériences sur une tâche visuelle de « barrage de cloches »

Dans ces expériences, on présente aux enfants une feuille comme celle de la figure 8, où des cloches sont réparties parmi d'autres dessins. On demande aux enfants de barrer toutes les cloches. C'est là une des tâches classiques en neuropsychologie. Mais, ce qui n'est pas classique, c'est qu'on a enregistré l'ordre dans lequel les enfants ont barré les cloches. La figure 9 donne des exemples de « barrage de cloches » pour six enfants, avec des parcours très différents les uns des autres : on constate que certains enfants font une exploration

[8] Cf Lower Attentional Skills predict increased exploratory foraging patterns. C. Van de Driessche et al. www.nature.com/scientificreports, *Scientific Reports 9:10948,* Jun.2019

systématique du voisinage d'une cloche donnée, d'autres font parfois des « sauts » importants. Ces différences s'avèrent liées aux différences de capacité attentionnelle.

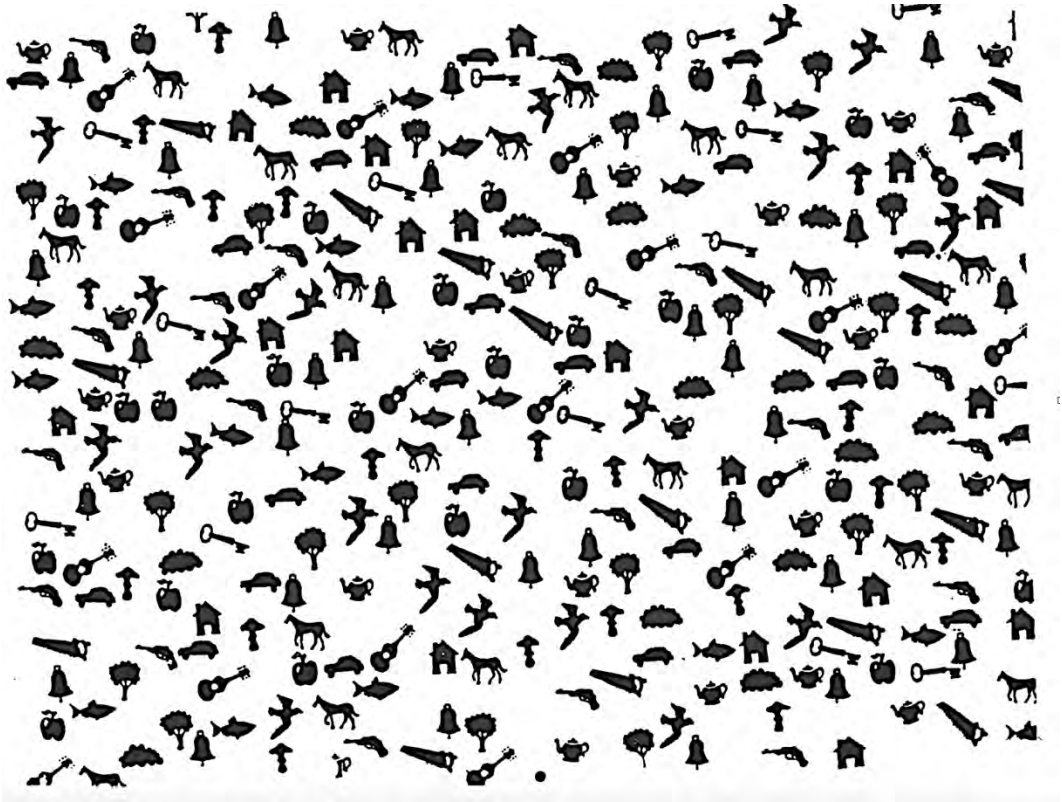


**Figure 8** : trouver les dessins de cloches dans cet ensemble de petits dessins ! (Van den Driessche et al, 2019 ; figure fournie par J. Sackur, co-auteur.)

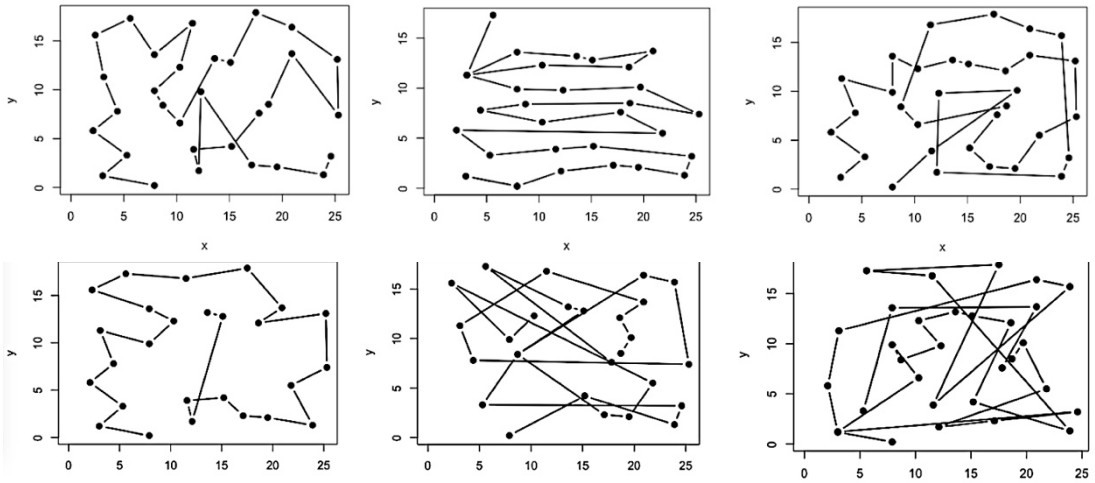



**Figure 9** Parcours de barrages de cloches pour six enfants

### 3.2 Expériences sur une tâche verbale de « fluence sémantique »

Dans ces expériences, on demande aux enfants de nommer en deux minutes autant d'animaux qu'ils ou elles le peuvent. Les enfants connaissent beaucoup d'animaux et sont très contents d'effectuer cette tâche. On a ensuite pris les productions de ces enfants (c'est-à-dire les listes des noms d'animaux) en les traduisant chacune en anglais et en appliquant sur ces traductions des modèles sémantiques tel que le modèle Word2vec de Google.

Chaque nom d'animal devient un point dans un espace de 350 dimensions (dans lequel tous les noms d'animaux courants sont déjà placés) ; chaque liste donnée par un enfant devient ainsi un parcours dans cet espace. Ce parcours une fois obtenu, on peut calculer la distance entre deux noms d'animaux successifs dans la liste (par exemple, la distance entre les deux points de l'espace respectivement associé à chien et chat est assez faible, comparativement avec celle entre chien et alligator). On a ensuite vérifié que ce modèle était cohérent, correspondant bien à une réalité sous-jacente de ces explorations : pour ce faire, nous avons regardé s'il existait une corrélation entre l'intervalle de temps séparant la production de deux noms successifs, que nous appellerons *temps de réponse* (RT), et la *proximité sémantique* (semantic similarity) de ces deux noms ; cette proximité sémantique est un score compris entre 0 (les deux noms sont très éloignés sémantiquement) et 1 (les deux noms donnés sont identiques, ou désignent le même animal)[9]. On s'attend à ce que plus le temps de réponse est long, plus la proximité sémantique des deux noms concernés sera proche de zéro. Et, à l'inverse, plus le temps est court, plus cette proximité sémantique sera proche de 1. Et c'est bien ce qui se passe, comme le montre la figure 10. La droite de régression du nuage descend de gauche à droite : plus les noms sont proches sémantiquement, plus les enfants ont été rapides pour passer de l'un à l'autre.

Pour illustrer les parcours sémantiques ainsi produits par les enfants, on ne peut faire autrement que de projeter ces parcours sur un espace réduit à deux dimensions. C'est ce qui est fait sur la figure 11, qui représente deux exemples de parcours : le parcours rouge est celui d'un enfant disposant d'une plus faible capacité attentionnelle, et le parcours bleu celui d'un enfant ayant plus de ressources attentionnelles.

Il ne s'agit ci-dessus que de deux exemples. Dans les deux types d'expériences que nous venons d'évoquer (visuelles et sémantiques), nous avons effectué des analyses statistiques en séparant en deux le groupe des 60 enfants – ceux qui savent bien se concentrer d'une part, et ceux qui ont plus de difficulté à le faire. La figure 12 donne les histogrammes des distances entre deux cloches successives, dans l'approche visuelle et les histogrammes des proximités entre deux noms successifs, dans l'approche sémantique ; dans les deux cas les

[9] Le calcul de la proximité sémantique est techniquement lié à la métrique de l'espace où les noms sont projetés par Word2vec : il s'agit en fait d'un *espace vectoriel,* et la distance entre deux noms est l'*angle* séparant les deux vecteurs qui les représentent respectivement : la proximité se calcule alors par le *cosinus* de cet angle. Elle est nulle lorsque les vecteurs sont orthogonaux, et égale à 1 lorsque ces vecteurs se superposent exactement.

histogrammes noirs sont ceux relatifs aux enfants du premier sous-groupe (bonne capacité de concentration) et les histogrammes rouges sont ceux relatifs au second sous-groupe.

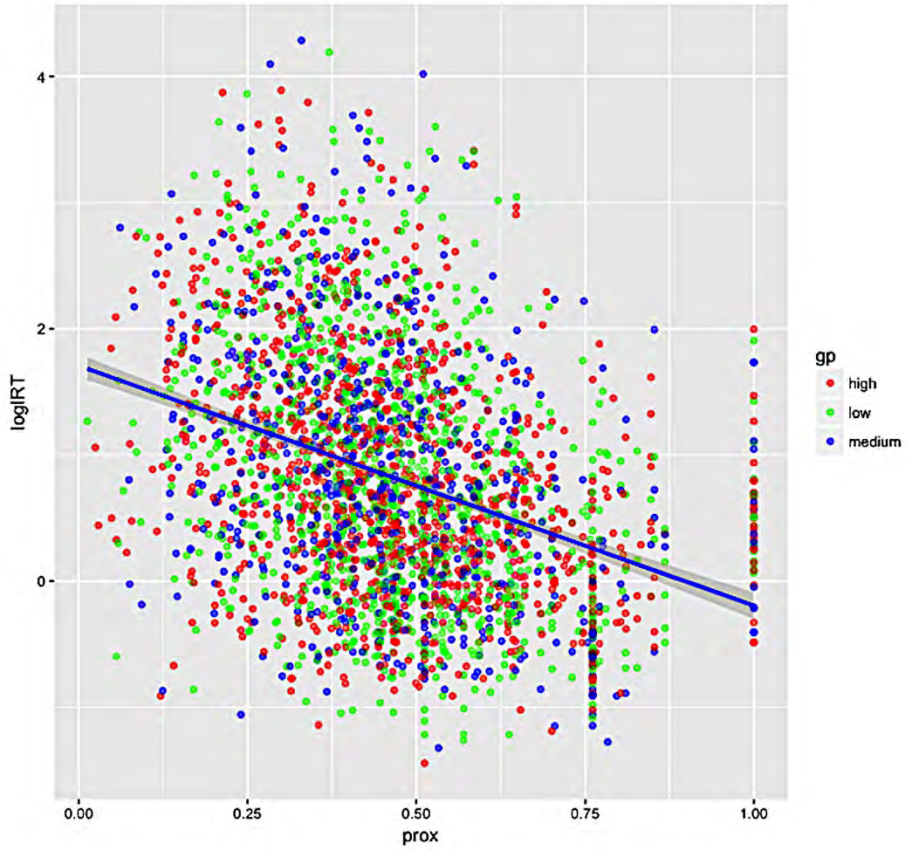



**Figure 10** : corrélation entre temps de réponses (en ordonnées logarithmiques) et proximité sémantique (en abscisse), pour trois catégories d'enfants classés selon leur capacité attentionnelle

L'analyse de ces résultats est complexe car chez les enfants ne disposant que d'une faible capacité de concentration, deux tendances inverses se manifestent[10] : d'une part ils ont tendance à faire des sauts plus importants (saut vers une cloche éloignée, saut vers un animal vraiment différent) d'autre part – et c'est particulièrement visible dans l'approche

[10] Les enfants explore le voisinage d'une façon assez précise et de temps en temps, ils manifeste une *indécision*, introduisent de plus grands sauts, que ce soit dans l'espace visuel ou sémantique. Les parcours dans les deux approches évoquent ce qu'on appelle, dans la théorie des processus aléatoires dans l'espace, des processus ou *vols de Lévy*, manifestant en pratique une combinaison de sauts dont la distribution en amplitude est contrôlée par trois paramètres. Les parcours des enfants pourraient se rapprocher de tels modèles théoriques, qui pourraient ainsi aider à en caractériser les différences.

sémantique, ils ont tendance à se répéter, ou bien à effectuer de « petits » sauts demandant peu d'efforts cognitifs, par exemple sauts du type *chien-chat*.

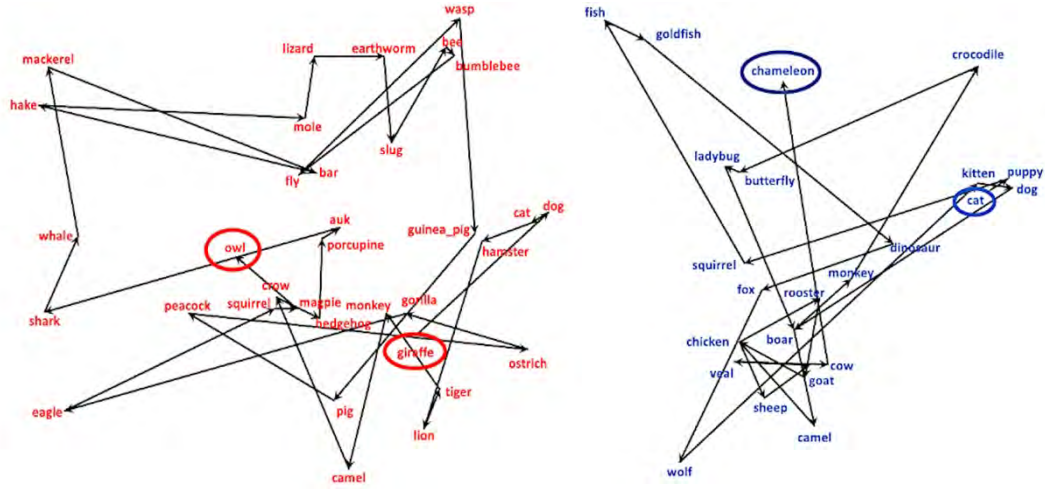



**Figure 11** : Illustration en deux dimensions du parcours sémantique de deux enfants dont l'un (à gauche) est affecté d'un degré d'ADHD élevé (Van den Driessche et al, 2019)

### 3.3 En brève conclusion de ces expériences

Les stratégies d'explorations sont peut-être un moyen d'accéder directement à la structure du flux mental. Aussi les dernières expériences présentées nous semblent intéressantes : elles nous paraissent suivre le flux de conscience de manière plus proche, plus précise que lors des expériences précédentes – même si ce n'est pas encore aussi exact qu'on le souhaiterait. Le moment en effet où l'enfant fait ce grand saut dans une exploration est précisément un moment de transition, un moment transitif au sens de James, une *syncope* du flux mental ; ce qu'on arrive à montrer, c'est que les capacités attentionnelles, telles qu'elles transparaissent dans plusieurs des tâches psychologiques, sont liées à la fréquence des moments de longue transition à l'intérieur du flux de conscience : plus ces capacités sont faibles, plus cette fréquence est élevée.

## 4. En guise de conclusion générale

Pour résumer, j'ai essayé de suggérer que la recherche sur la conscience devrait désormais inclure l'étude non pas simplement des moments, ou des instantanés, de la conscience, mais aussi la dynamique du flux conscient. J'ai souligné à quel point ce travail se heurte à un double défi : d'une part il nous faut concevoir des tâches expérimentales qui apporteraient plus d'information que les tâches classiques d'échantillonnage d'expérience; mais d'autre part et peut-être surtout, il faut concevoir les modèles nous permettant d'interpréter ces données. Je n'ai pas évoqué la question de l'apport que les techniques d'imagerie cérébrale (IRMf, EEG, MEG) pourraient avoir dans ce domaine: il me semble que cet apport sera

décisif une fois que nous aurons réussi à valider des tâches et des modèles comportementaux qui rendront compte fidèlement de la dynamique subjective du flux de conscience. J'ai évoqué les premiers résultats que nous avons obtenus grâce à l'utilisation d'outils d'analyse du langage naturel.

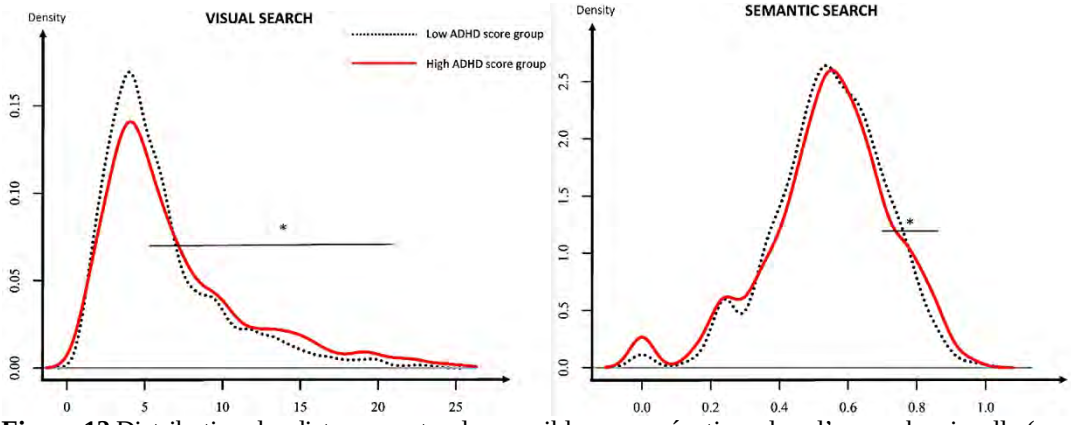



**Figure 12** Distribution des distances entre deux « cibles » consécutives dans l'approche visuelle (en abscisse, les distances entre les cloches) et dans l'approche sémantique (en abscisse, la proximité entre les noms d'animaux) (Van de Driessche et al 2019)

Enfin, je voudrais conclure en remerciant Mikaël Bastian, Charlotte Van Den Driessche et Valentin Wyart, sans qui ces idées n'auraient pas vu le jour.

# 5

# Bases cérébrales de la spécialisation hémisphérique de l'attention visuo-spatiale et des relations complémentaires entre l'attention spatiale et le langage

**Laure Zago**

Groupe d'Imagerie Neurofonctionnelle
Institut des Maladies Neurodégénératives, UMR 5293
Université de Bordeaux

## Abstract

Laure Zago studies the functional difference and the complementarity of the functions of the two hemispheres of the brain. This asymmetry has an impact on the phenomenology of consciousness, as shown for example by the non-awareness of certain stimuli in patients affected by certain unilateral brain lesions. It is moreover by questions concerning a concept linked to consciousness, namely attention, that Laure Zago approaches her point. She thus deals with the *attentional biases* assumed to be linked to lateralization; behavioral biases, of which she lists different types and describes their characteristics in patients with brain damage affecting the right hemisphere, in comparison with healthy subjects. She shows some results around the correlation between the importance of these biases and the intensity of the lateralization of the concerned processes. A final section takes up the question of the origin of the specialization entrusting visuospatial processes to the right hemisphere and those of language to the left hemisphere. The results seem to show that the population of left-handers (less than 10% of the general population) is a population to be favored for "understanding the rules for implementing hemispheric specialization of lateralized cognitive functions".

La spécialisation hémisphérique (SH) ou dominance hémisphérique fait référence à la relation entre un ensemble de structures cérébrales d'un hémisphère donné et une fonction cognitive donnée. Cette latéralisation des fonctions est observée chez l'animal (*1*) mais c'est chez l'Homme que l'on observe un biais d'espèce. Alors que chez l'animal la SH peut varier d'un individu à l'autre, chez l'Homme il existe une faible variabilité entre les individus. Ainsi, la principale expression comportementale de la latéralisation est la préférence manuelle, avec, pour plus de 90% de la population humaine, une préférence de l'utilisation de la main droite pour effectuer des activités uni-manuelles quotidiennes.

Au niveau cérébral, les premiers travaux de M. Dax et ceux de P. Broca (*2*) au 19ème siècle ont permis d'associer des troubles aphasiques (déficits de compréhension et/ou de production du langage) à la présence de lésions cérébrales de l'hémisphère gauche. Similairement, une lésion de l'hémisphère droit induit un syndrome d'hémi-négligence, défini comme un trouble attentionnel caractérisé par une perte (totale ou partielle) de conscience des stimuli localisés dans l'hémi-espace gauche de l'individu (*3*, *4*). De nombreux travaux de neuropsychologie (*5*) et de neuroimagerie chez l'individu non cérébro-lésé (*6*) ont confirmé l'existence de ce biais de latéralisation cérébrale dans la population, avec l'hémisphère gauche hébergeant le langage et l'hémisphère droit intégrant les fonctions attentionnelles visuo-spatiales. Toutefois, l'origine et les mécanismes de la mise en place de cette SH complémentaire des fonctions latéralisées restent méconnus (*7*–*9*). En ce qui concerne la latéralisation du langage, celle-ci est cruciale pour un fonctionnement cognitif optimal, des formes atypiques de latéralisation pour le langage ayant été associées à une moindre efficience cognitive (*10*, *11*) et à certaines pathologies telles que les troubles développementaux du langage (*12*).
La SH constitue donc un principe fondamental de l'organisation cérébrale de l'Homme, dont la complexité vient d'être récemment reconsidérée à l'aide des nouvelles techniques de neuroimagerie (*13*–*15*).

Dans ce chapitre, nous décrirons nos contributions à la compréhension des réseaux latéralisés, et plus particulièrement sur les études des bases neurales de la SH de l'attention visuo-spatiale et de la SH complémentaire des fonctions du langage et de l'attention spatiale.

## 1. Latéralisation cérébrale de l'attention spatiale

Chaque jour, nous sommes confrontés à des quantités d'informations de nature différente, provenant de notre environnement extérieur et intérieur. Afin de réaliser des tâches et d'atteindre des buts, certaines de ces informations sont sélectionnées par nos capacités attentionnelles et traitées par les processus cognitifs, d'autres sont, au contraire, ignorées.
Bien que nous puissions mobiliser et diriger avec précision nos ressources attentionnelles, notre capacité à contrôler notre attention n'est pas parfaite. D'une part, de nombreux

distracteurs peuvent interférer avec notre capacité à exécuter des tâches avec succès et d'autre part, la manière dont nous déployons notre attention dans l'environnement ne se fait pas de manière homogène. C'est cette dernière caractéristique que nous avons particulièrement étudiée, avec notamment l'étude des biais attentionnels.

## 2. Les biais attentionnels comportementaux chez l'adulte sain

Les biais attentionnels sont des biais comportementaux dont il a été proposé qu'ils reflèteraient la dominance de l'hémisphère droit pour les processus visuo-spatiaux attentionnels. Cette dominance hémisphérique biaiserait le déplacement de l'attention vers l'hémi-espace controlatéral à l'hémisphère recruté (*16–18*).

Ces biais comportementaux sont aisément observables et quantifiables chez le patient hémi-négligent présentant des lésions cérébrales postérieures droites qui déplace alors son attention uniquement dans l'hémi-espace droit et ignore l'hémi-espace gauche. Par exemple, *le biais de bissection* est évalué au cours du test de bissection de ligne qui consiste à barrer avec un crayon une ligne horizontale en son milieu (*19*). Ce biais se caractérise chez le patient par une évaluation du milieu largement décalée à droite. *Le biais de cochage* évalué au cours du test de cochage consiste à cocher une cible parmi des distracteurs (*20*). Chez le patient hémi-négligent, il se caractérise par un cochage des cibles sur la moitié droite de la feuille (*21*, *22*).

Chez l'adulte non cérébro-lésé, ces biais sont aussi observables, mais de manière plus ténue, et se caractérisent à l'inverse des patients hémi-négligents par une exploration attentionnelle légèrement décalée vers l'hémi-espace gauche (*23*). Ces biais illustrent le phénomène de « pseudo-négligence », intitulé en référence au syndrome d'hémi-négligence. Des études comportementales indiquent que cette pseudo-négligence se développe au cours de l'enfance (*24*, *25*) suggérant un lien avec la maturation cérébrale. Il a été également montré que ces biais sont modifiés dans des pathologies développementales telles que celles présentant des déficits de l'attention et dans une moindre mesure dans les troubles « dys ». A l'heure actuelle, les biais attentionnels tant au niveau comportemental que cérébral restent peu explorés chez l'adulte sain.

### 2.1 Les biais de pseudo-négligence

Une première étude comportementale effectuée sur un groupe de 50 témoins droitiers, nous a permis de mettre en évidence l'existence de biais de pseudo-négligence comportementale au cours d'une condition de jugement de bissection de ligne et d'une condition de cochage de cibles (*26*). De manière intéressante, nous avons montré que ces deux biais, bien que chacun d'eux en faveur d'une pseudo-négligence, n'étaient pas corrélés l'un à l'autre, indiquant qu'un témoin pouvait être biaisé dans une des conditions et pas dans l'autre. Ce

résultat suggère que ces deux biais ne reflètent pas les mêmes mécanismes associés à la pseudo-négligence, et supportent la mise en place d'études spécifiques à chacun des biais. De plus, cette absence d'association peut être mise en rapport avec l'existence de patients dissociés présentant, par exemple des déficits lors d'une tâche de bissection et une préservation des performances lors de la tâche de cochage (*27*).

Une deuxième étude comportementale effectuée sur un plus large échantillon incluant des droitiers et des gauchers, nous a permis de mettre en évidence que le biais de bissection est un biais multifactoriel, dont l'intensité et la direction résultent de l'intégration des processus attentionnels, mais également des traitements perceptifs et moteurs reliés à la réalisation de la tâche et enfin, des facteurs individuels de latéralisation tels que la préférence manuelle et la préférence oculaire. Ainsi, nous avons pu démontrer que le biais comportemental est majoré dans les conditions où ces différents facteurs produisant une potentielle activation cérébrale asymétrique coïncident (*28*). Notamment, le biais de bissection était maximisé lorsque le témoin était gaucher avec œil directeur droit, et utilisait sa main dominante gauche pour effectuer la bissection de stimuli localisés dans l'hémi-champ gauche. Notre hypothèse est qu'au niveau cérébral, ces facteurs se conjuguent pour produire les conditions d'une asymétrie d'activation de différentes régions cérébrales en faveur de l'hémisphère droit.

### 2.2 Relation entre les biais attentionnels et la latéralisation cérébrale

Suite à la mise en évidence de biais comportementaux chez l'adulte, une deuxième étape a été d'identifier les bases cérébrales à l'origine de cette pseudo-négligence.

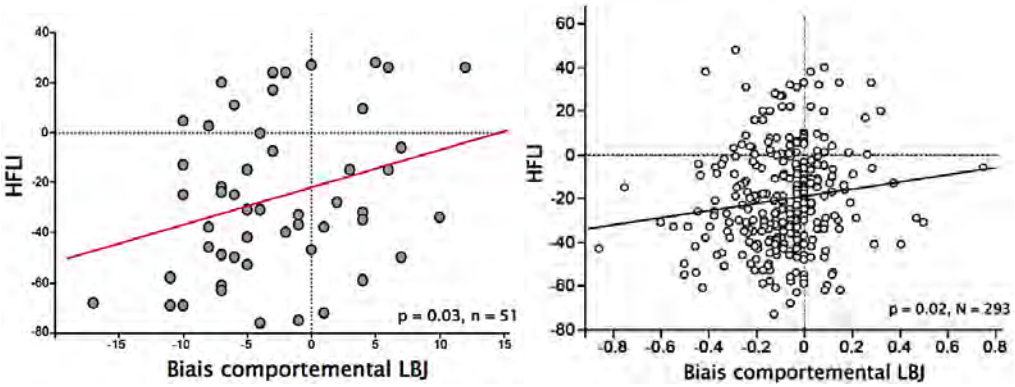



**Figure 1**. Corrélations entre les indices de latéralisation hémisphérique (HFLI) au cours de la tâche de jugement de bissection de ligne (LBJ, LBJ en anglais) et les valeurs individuelles de biais comportemental. Dans les deux études effectuées sur deux échantillons différents, cette association démontre que la latéralisation cérébrale droite (indices de HFLI négatifs) est associée à l'intensité de la pseudo-négligence (valeurs de biais négatives).

En utilisant une tâche de jugement de bissection de ligne (JLB), nous avons montré dans deux études différentes incluant un grand nombre de témoins que cette tâche recrute des

réseaux cérébraux asymétriques droits dont l'asymétrie fonctionnelle est prédictive du biais de pseudo-négligence comportemental (*26, 29*). Comme illustré dans la Figure 1, les témoins ayant une forte latéralisation droite sont ceux ayant un biais de pseudo-négligence important. Ainsi, le biais attentionnel est un marqueur de la latéralisation hémisphérique de l'attention visuo-spatiale.

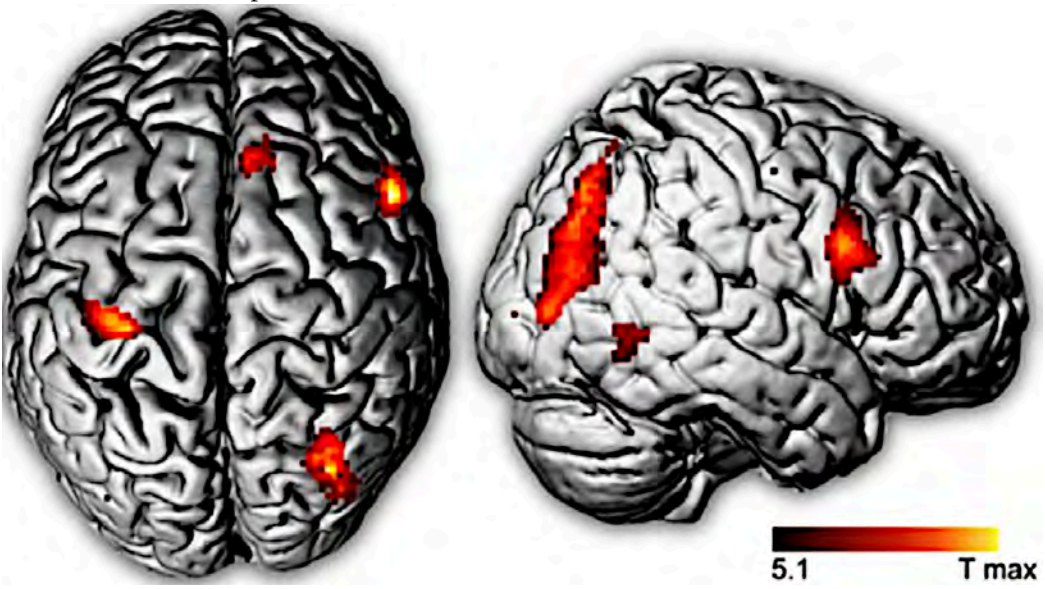



**Figure 2**. Régions cérébrales activées et asymétriques droites au cours de la tache de LBJ comparé à une condition visuo-motrice de saccades visuellement guidées.

Concernant les régions cérébrales, nous avons mis en évidence l'implication asymétrique droite des régions postérieures occipito-pariétales, de la jonction occipito-temporale (JOT), et des régions antérieures comprenant le gyrus frontal inférieur, le gyrus frontal médian et l'insula antérieure (Figure 2).

Afin d'évaluer si les régions présentant une asymétrie droite chez les adultes sains étaient similaires à celles dont la lésion provoque une hémi-négligence, nous avons comparé les réseaux asymétriques avec les cartes de probabilité de lésions de patients hémi-négligents. Cette comparaison a démontré que certaines régions ventrales temporo-pariétales et frontales inférieures droites sont identiques à celles lésées de patients ayant un déficit spécifique à la bissection de ligne comparés à ceux ayant un déficit spécifique au cochage (*29, 29*) (Figure 3 en jaune)

Ces différents résultats permettent de conclure que ce réseau de régions latéralisé à droite pendant la tâche de bissection, sous-tend le biais comportemental de bissection. Les travaux sur les origines neurales des biais se poursuivent, ainsi que ceux visant à étudier la variabilité de la latéralisation des réseaux impliqués dans l'orientation de l'attention.

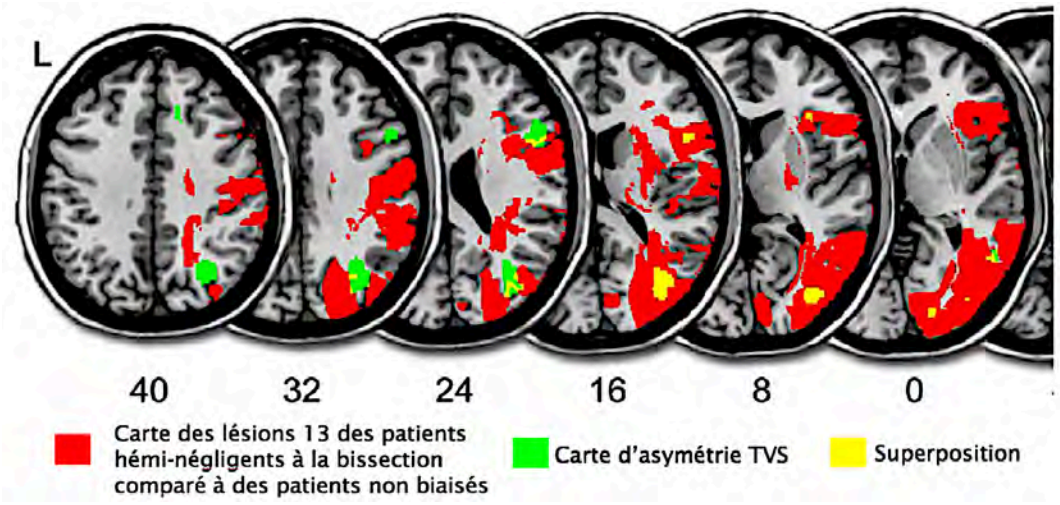



**Figure 3**. Superposition de la carte statistique d'asymétrie fonctionnelle de témoins sains au cours du jugement de bissection de ligne (en vert) et de la carte de lésions (> 20 %) de patients hémi-négligents présentant un déficit spécifique au test de bissection de ligne (en rouge, détails dans la figure 3 de Rorden et al., 2006). La superposition est en jaune.

Ces résultats sont complémentaires de ceux obtenus sur les bases neurales des saccades oculaires et de leur latéralisation. En effet, la spécialisation de l'hémisphère droit dans les processus d'attention spatiale est déjà observable dès le déplacement du regard permettant le déplacement du foyer attentionnel (*30*).

### 2.3 Latéralisation du réseau de déplacement du regard

Le plus souvent, le déplacement de l'attention s'effectue avec le déplacement du regard. Cette activité oculomotrice est donc la brique de base sur laquelle vont s'implémenter les processus attentionnels. Parmi le large réseau de régions pariéto-frontales bilatérales sous-tendant le déplacement du regard illustré dans la Figure 3, on remarque une latéralisation droite des activations dans les régions dorsales et ventrales du réseau du déplacement attentionnel (Figure 4, régions bleues).

Dans une deuxième étude, nous avons démontré que la force de la préférence manuelle et la préférence oculaire sont deux facteurs explicatifs de la variabilité de la latéralisation du réseau dorsal du déplacement attentionnel associé au déplacement du regard. Ainsi, les individus gauchers ayant une forte préférence manuelle gauche présentent une plus forte latéralisation droite du réseau fronto-pariétal dorsal que les droitiers. De plus, les gauchers avec un œil directeur droit sont plus latéralisés que les autres participants.

Nous avons fait l'hypothèse que ces sujets pourraient bénéficier d'un avantage en termes de traitement des informations plus rapide car l'hémisphère droit contrôle à la fois leur main dominante et les traitements visuo-spatiaux (*31*). Rappelons que c'est dans cette population de gauchers ayant un œil directeur droit que nous avons observé les plus forts biais de pseudo-négligence (*28*).

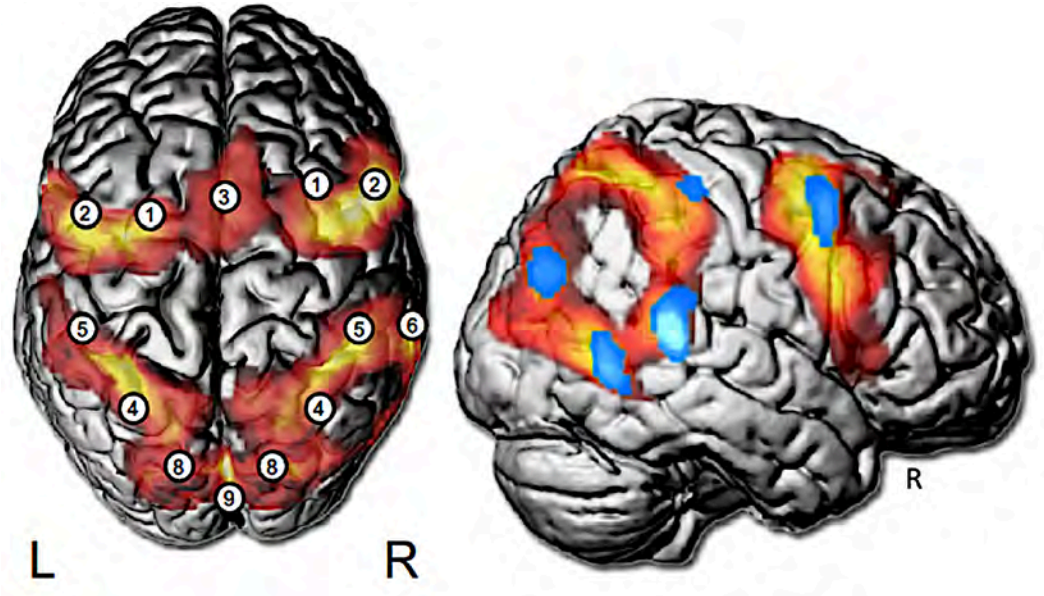



**Figure 4.** A gauche : Vue supérieure du cerveau illustrant les régions fronto-pariétales bilatérales impliquées dans le déplacement de regard au cours de saccades visuellement guidées comparée à une condition de fixation du regard. A droite : Vue latérale de l'hémisphère droit. En bleu, les régions cérébrales montrent une asymétrique droite du réseau de déplacement du regard (Petit, Zago et al. 2009).

## 3. Complémentarité hémisphérique des fonctions latéralisées

Dans la majorité de la population, l'hémisphère gauche est spécialisé pour le langage et l'hémisphère droit pour les processus visuo-spatiaux. Bien que les mécanismes et l'origine de cette mise en place restent largement méconnus, deux hypothèses ont été proposées afin de rendre compte de cette complémentarité hémisphérique (*32*). Une première hypothèse propose que la latéralisation gauche pour le langage est à l'origine de la latéralisation droite pour les fonctions visuo-spatiales. Cette hypothèse causale propose que ces latéralisations seraient associées (*6*, *33*, *34*, *34*). La deuxième hypothèse intitulée l'hypothèse statistique considère que cette complémentarité est une norme statistique et que la latéralisation de ces deux fonctions s'effectue de manière indépendante (*35*).

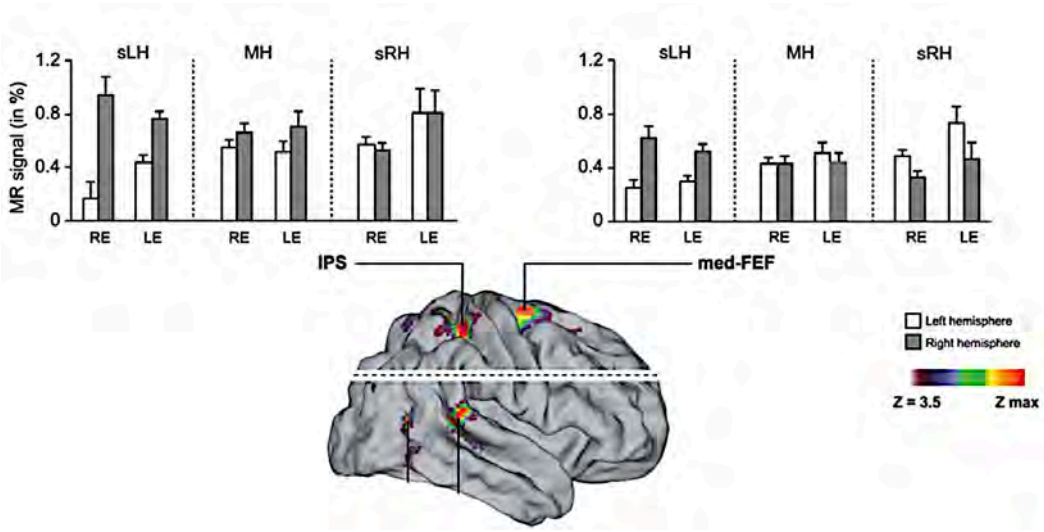



**Figure 5.** Régions cérébrales pour lesquelles il existe une interaction entre la force de la préférence manuelle (sLH = gauchers, MH = faiblement latéralisé, sRH = droitiers) et l'œil directeur (RE = œil directeur droit ; LE = œil directeur gauche) sur l'asymétrie fonctionnelle droite dans le sillon intrapariétal (IPS) et les champs oculomoteurs frontaux médians (med-FEF). Dans ces régions, la plus forte asymétrie droite est observée pour les témoins gauchers avec un œil directeur droit.

En analysant les résultats des différentes études en faveur de l'une ou de l'autre des hypothèses, nous avons identifié que l'inclusion de gauchers dans la population d'étude était un facteur de différence.

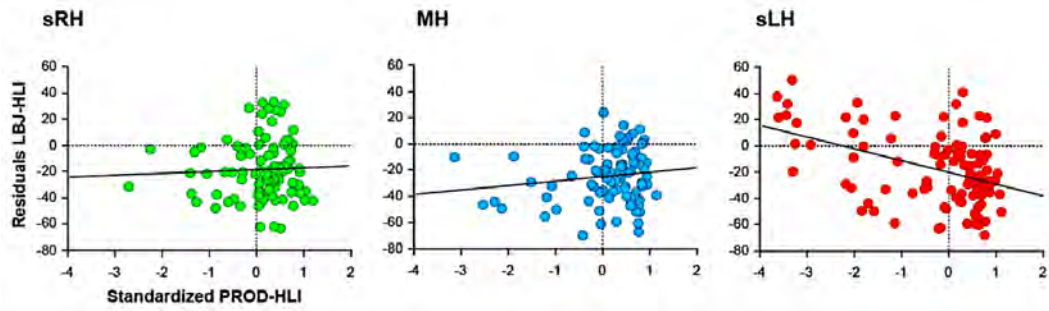



**Figure 6**. Corrélations entre les indices de latéralisation hémisphérique pour l'attention spatiale (LBJ-HLI) et le langage (PROD-HLI) en fonction de la force de la préférence manuelle (sRH = droitiers avec une forte préférence manuelle (vert) ; MH : témoins ne présentant pas une préférence manuelle marquée (bleu) ; sLH = gauchers avec une forte préférence manuelle (rouge). Les résultats montrent que la corrélation est significative uniquement chez les gauchers.

Nous avons donc proposé que la préférence manuelle pourrait être un facteur explicatif de la variabilité de la complémentarité. Afin de tester cette hypothèse, nous avons étudié une population de témoins volontaires sains incluant le même nombre de droitiers et de gauchers, testé la force de leur préférence manuelle et étudié les latéralisations cérébrales en IRMf au cours d'une condition de production du langage et d'une condition d'attention spatiale (jugement de bissection de ligne). Les résultats indiquent que la nature de la relation entre les latéralisations du langage et de l'attention spatiale est différente suivant la préférence manuelle. Ainsi, dans la majorité de la population incluant des droitiers et des sujets faiblement latéralisés pour leur main, il n'existe pas d'association entre la latéralisation hémisphérique du langage et celle de l'attention spatiale, résultat en faveur de l'hypothèse d'une indépendance de la complémentarité des fonctions. Par contre, nous avons découvert chez les gauchers fortement latéralisés de la main gauche une association entre les latéralisations cérébrales : plus un gaucher est latéralisé dans un hémisphère pour le langage et plus il sera latéralisé pour l'attention spatiale dans l'autre hémisphère.

Ce résultat très important indique que les règles d'organisation cérébrale diffèrent chez les gauchers. Cette population, qui représente moins de 10 % de la population générale, semble être la population « cible » pour la compréhension des règles de mise en place de la spécialisation hémisphérique des fonctions cognitives latéralisées. C'est ainsi avec cette population que l'on observe la plus grande variabilité de latéralisation des fonctions cognitives. Nos travaux se poursuivent et nous cherchons à comprendre les liens qu'entretiennent d'autres fonctions latéralisées telles que le calcul, le langage et l'attention.

**Références**


1. W. D. Hopkins, C. Cantalupo, *Current Directions in Psychological Science* **17**, 233 (2008).
2. L. Manning, C. Thomas-Antérion, *Rev Neurol (Paris)* **167**, 868 (2011).
3. M. Corbetta, G. L. Shulman, *Annu Rev Neurosci* **34**, 569 (2011).
4. H.-O. Karnath, *Neuropsychologia* **75**, 61 (2015).
5. M. P. Bryden, H. Hécaen, M. DeAgostini, *Brain Lang* **20**, 249 (1983).
6. G. Badzakova-Trajkov, I. S. Häberling, R. P. Roberts, M. C. Corballis, *PLoS One* **5**, e9682 (2010).
7. M. C. Corballis, *Prog Brain Res* **195**, 103 (2012).
8. M. S. Gazzaniga, *Brain* **123**, 1293 (2000).
9. A. Oleksiak, A. Postma, I. J. M. van der Ham, P. C. Klink, R. J. A. van Wezel, *Brain Res Rev* **67**, 56 (2011).
10. R. Everts *et al.*, *Hum Brain Mapp* **30**, 473 (2009).
11. E. Mellet *et al.*, *Neuropsychologia* **65**, 56 (2014).
12. D. V. M. Bishop, *Science* **340**, 1230531 (2013).
13. M. C. Corballis, *PLoS Biol* **12**, e1001767 (2014).
14. P. Y. Hervé, L. Zago, L. Petit, B. Mazoyer, N. Tzourio-Mazoyer, *Trends Cogn Sci* **17**, 69 (2013).
15. N. Tzourio-Mazoyer, M. Perrone-Bertolotti, G. Jobard, B. Mazoyer, M. Baciu, *Cortex* (2016).
16. D. Bowers, K. M. Heilman, *Neuropsychologia* **18**, 491 (1980).

17. M. Kinsbourne, in *The cerebral basis of lateral asymmetries in attention*, Ed. (North-Holland Publishing Company, Amsterdam, 1970), pp. 193-201.
18. M. M. Mesulam, *Philos Trans R Soc Lond B Biol Sci* **354**, 1325 (1999).
19. K. M. Heilman, T. Van Den Abell, *Neuropsychologia* **17**, 315 (1979).
20. S. Weintraub, M. M. Mesulam, *J Neurol Neurosurg Psychiatry* **51**, 1481 (1988).
21. H. O. Karnath, S. Ferber, M. Himmelbach, *Nature* **411**, 950 (2001).
22. D. J. Mort *et al.*, *Brain* **126**, 1986 (2003).
23. G. Jewell, M. E. McCourt, *Neuropsychologia* **38**, 93 (2000).
24. K. Patro, H.-C. Nuerk, P. Brugger, *Journal of Experimental Child Psychology* **173**, 16 (2018).
25. M. Hausmann, K. E. Waldie, M. C. Corballis, *Neuropsychology* **17**, 155 (2003).
26. L. Zago *et al.*, *Neuropsychologia* **93**, 394 (2016).
27. C. Rorden, M. Fruhmann Berger, H.-O. Karnath, *Brain Res* **1080**, 17 (2006).
28. A. Ochando, L. Zago, *Frontiers in Psychology* **9**, (2018).
29. L. Zago *et al.*, *Neuropsychologia* **94**, 75 (2017).
30. L. Petit *et al.*, *J Neurophysiol* **102**, 2994 (2009).
31. L. Petit *et al.*, *Hum Brain Mapp* **36**, 1151 (2015).
32. M. P. Bryden, *Canadian Psychology* **31**, 297 (1990).
33. Q. Cai, L. Van der Haegen, M. Brysbaert, *Proc Natl Acad Sci U S A* **110**, 322 (2013).
34. J. L. Powell, G. J. Kemp, M. García-Finaña, *Neuroimage* **59**, 1818 (2012).
35. A. J. O. Whitehouse, D. V. M. Bishop, *Neuropsychologia* **47**, 1938 (2009).

# 6 | Représentation et manipulation des concepts mathématiques par le cerveau humain

**Marie Amalric**
Research group The Concepts, Actions, and Objects (CAOs) Lab/The Kid Neuro Lab
Carnegie Mellon University Pittsburgh, USA

## Abstract

Marie Amalric studies the localization of cognitive processes linked to mathematical activities. According to certain hypotheses, these processes are close to those concerned with language, while other hypotheses make them processes related to visuospatial processes. Marie Amalric first recalls previous work showing that we have from birth a minimal nucleus giving an intuitive approach to quantity, number, and positioning in space. These studies suggest that there is an innate proto-mathematical capacity, which then increases by learning. Then she presents experiments carried out comparatively with mathematician and non-mathematician subjects, submitting to them exercises in understanding statements of different types, mathematical and non-mathematical, and observing what is happening in the brain using fMRI imaging. The fMRI tests show that the areas involved in solving mathematical questions are different from those involved in the syntactic and semantic understanding of ordinary language. This is what is observed, as much with professional mathematicians as with non-mathematicians. They also show that at the level at which they can be observed, the same areas of the brain are called upon by mathematical questions, whatever their level and whatever the mathematical discipline concerned. Marie Amalric finally exposes additional investigations and their conclusions, concerning the localization of logical reasoning and the links with visuo-spatial processes. All of the experiments carried out would seem to indicate that the initial nucleus, which hosts proto-mathematical capacities from birth, then remains involved in the exercise of mathematics within a large sector which would grow around this initial nucleus.

# 1. Des Intuitions « proto-mathématiques » fondamentales

Des études comportementales ont révélé que les humains disposent, dès leur naissance, d'intuitions fondamentales relatives au nombre et à l'espace, intuitions qu'ils partagent avec de nombreuses autres espèces animales. En particulier, des bébés âgés de quelques heures sont déjà capables d'extraire des informations numériques de leur environnement. Izard et collaborateurs (2009) ont ainsi montré qu'après avoir été exposés à des séries de sons contenant un certain nombre de syllabes, ces nouveau-nés manifestaient plus d'intérêt pour une image montrant le même nombre d'objets que pour une image montrant un nombre d'objets différent. Très rapidement au cours du développement, les bébés se révèlent dotés de la capacité de percevoir de manière exacte des quantités inférieures à 3 (c'est la subitisation), ainsi que de comparer et ajouter des quantités de manière approximative (Wynn, 1992). La présence de capacités similaires a pu être identifiée chez de très nombreuses espèces animales, à commencer par les singes et plus généralement les mammifères tels que les félins ou les chevaux, mais aussi chez les poissons, les oiseaux, les amphibiens et même les insectes (Benson-Amram et al., 2011; Rugani et al., 2009). De façon remarquable, la perception numérique chez les différentes espèces animales partage une caractéristique commune : un effet de distance mesuré sur une échelle logarithmique, qui traduit le fait que plus des nombres sont grands et proches, plus il est difficile de les comparer (Cantlon et al., 2016). Très récemment, il a été montré que cet effet de distance s'étend également au traitement non-symbolique des ratios (Matthews et al., 2016). La perception spontanée que les humains semblent avoir des ratios pourrait d'ailleurs sous- tendre une autre compétence mathématique qui émerge très précocement chez l'homme : une sensibilité certaine aux probabilités et aux régularités statistiques (Marcus et al., 1999; Teglas et al., 2011). Enfin, l'homme s'avère être muni, comme beaucoup d'autres espèces animales, d'intuitions spontanées des relations spatiales, mais aussi des formes et de leurs propriétés. Dans des tâches de réorientation spatiale où l'objectif est de localiser un objet, les très jeunes enfants et de nombreuses espèces animales se repèrent et s'orientent principalement grâce à des indices géométriques de distance et de direction (Chiandetti and Vallortigara, 2007; Lee and Spelke, 2008). De plus, malgré l'absence d'éducation formelle, les jeunes enfants américains et les adultes Mundurukus se

En Amazonie, les indiens Mundurukus disposent d'un vocabulaire mathématique pauvre et d'un accès limité à l'éducation.

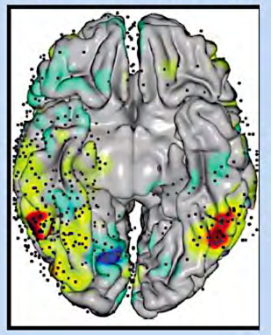
**Théorie des « Core knowledge »**: Cette théorie développée par Elizabeth Spelke suggère que l'homme est muni de façon innée de 6 noyaux de connaissances fondamentales sur les objets, les actions, les personnes, les nombres, les formes et l'espace navigable. Ces systèmes sont encapsulés et initialement indépendants les uns des autres.

montrent spontanément capables d'extraire et d'utiliser les informations géométriques abstraites contenues dans une carte pour localiser un objet, même lorsqu'ils y sont confrontés pour la première fois (Dillon et al., 2013). Par ailleurs, des tâches de détection d'intrus ont été utilisées pour montrer que l'absence d'éducation ou de vocabulaire numérique n'empêche en rien les Mundurukus de reconnaître spontanément de nombreux concepts géométriques tels que des formes et des propriétés euclidiennes, topologiques ou métriques (Dehaene et al., 2006). Ces intuitions forment des systèmes de connaissances fondamentales, ou « core knowledge » selon la théorie du même nom développée par Elizabeth Spelke.

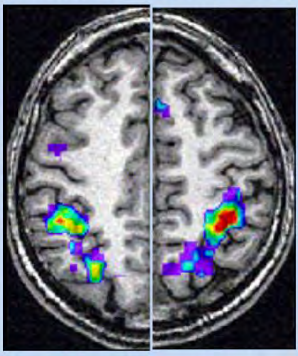
## 2. La « bosse des maths » revisitée

Le récent développement des techniques d'imagerie cérébrale a offert un éclairage nouveau sur les bases cérébrales de ces intuitions fondamentales. L'utilisation de la technique d'imagerie par résonance magnétique fonctionnelle (IRMf) a notamment mené à l'hypothèse que le sillon intrapariétal puisse jouer un rôle décisif dans la représentation des nombres. Il semble en effet systématiquement impliqué dans la réalisation de tous types de tâches numériques (calcul, comparaison, approximation, avec des symboles ou non,…) chez les adultes mais aussi chez de très jeunes enfants (Dehaene et al., 2003; Hyde et al., 2010; Piazza et al., 2004). Une étude très récente a même suggéré l'existence d'une certaine « numérotopie » dans le lobule pariétal supérieur (Harvey et al., 2013). Cela signifie que les réponses corticales à des nombres voisins se situent aussi dans des régions voisines, rangées dans le même ordre que les nombres auxquels ces régions répondent. De plus, des études intracrâniennes ont révélé que des régions pariétales équivalentes semblent également être impliquées dans la réalisation de tâches numériques chez le singe (Nieder and Miller, 2004). Chez l'homme, les études intracrâniennes sont bien entendu rares, et cantonnées aux cas de patients, le plus souvent épileptiques. Mais notamment deux d'entre elles ont permis de révéler que des électrodes localisées dans le gyrus temporal inférieur répondaient davantage à la présentation de chiffres arabes que de lettres ou de faux symboles (Shum et al., 2013), et davantage au calcul qu'à la lecture de phrases (Hermes et al., 2015). Les régions cérébrales ainsi identifiées ont été nommées « aires de la forme visuelle des nombres ».

Aire de la forme visuelle des nombres (Shum et al 2013)



"Sens du nombre" régions pariétales (Dehaene et al 2003)



*Cependant, les mathématiques ne peuvent se résumer à la compréhension basique des nombres et de l'espace, et les mécanismes par lesquels le cerveau humain parvient à conceptualiser des objets mathématiques plus avancés restent obscurs. Notamment, le rôle du langage dans le développement des connaissances mathématiques est très débattu en sciences cognitives*

## 3. La pensée mathématique peut-elle exister sans langage ?

Voici une question séculaire qui intrigue bien des philosophes et scientifiques. L'activité mathématique et les compétences pour le langage parlé reposent pour certains sur les mêmes mécanismes d'abstraction. Noam Chomsky (2006) prétend même que l'activité mathématique a émergé chez l'Homme comme conséquence de ses capacités de langage. Pourtant, la plupart des mathématiciens et physiciens ne perçoivent pas l'influence du langage dans leur réflexion. C'est ce qu'évoque notamment le mathématicien français Jacques Hadamard en 1945, dans un ouvrage analysant les processus à l'œuvre dans l'invention mathématique. En particulier, il y rapporte sa correspondance avec nombre de ses collègues, dont Albert Einstein, qui lui écrit : « Les mots ou le langage, écrit ou parlé, ne semblent pas jouer le moindre rôle dans le mécanisme de ma pensée. Les entités psychiques qui servent d'éléments à la pensée sont certains signes ou des images plus ou moins claires […] qui sont, dans mon cas, de type visuel et parfois moteur. Les mots […] n'ont à être cherchés avec peine qu'à un stade secondaire. »

En accord avec l'idée de Noam Chomsky, quelques études tendent à montrer que les mots de nombres sous-tendent le développement des compétences numériques exactes. Par exemple, les Mundurukus sont capables de réaliser des additions approximatives aussi bien que les Français, mais ils échouent lorsqu'il s'agit de réaliser des soustractions exactes dont le résultat ne dépasse pourtant pas le seuil de subitisation (Pica et al., 2004).

D'un autre côté, plusieurs études ont suggéré que l'arithmétique et le calcul algébrique simple sont dissociés du langage parlé naturel. C'est le cas chez des patients aphasiques, qui présentent des troubles de la compréhension ou de la production du langage, mais qui peuvent toujours faire des calculs ou de l'algèbre (Varley et al., 2005). A l'inverse, une acalculie acquise, suite à un AVC par exemple, peut laisser les capacités de langage intactes (Dehaene and Cohen, 1995). Quelques études d'IRM ont également révélé des différences nettes d'activations cérébrales pour les activités algébriques versus linguistiques (Monti et al., 2012). Une alternative à l'hypothèse de Chomsky est donc que l'arithmétique se construit directement sur les intuitions proto-mathématiques du nombre, et qu'au niveau cérébral, elle recycle certaines régions associées au traitement numérique de base. Certaines études plaident en faveur de cette hypothèse, en montrant qu'au niveau individuel, l'acuité du "sens du nombre" prédit les compétences mathématiques ultérieures (Halberda et al., 2008).

*Alors que les études passées se sont principalement intéressées à l'arithmétique élémentaire, mon travail de thèse a, pour la première fois, permis de mettre en place une série d'expériences visant à déterminer quelles aires cérébrales sont impliquées dans la réflexion mathématique de haut niveau.*

## 4. Origine cérébrale des concepts mathématiques de haut niveau (Amalric and Dehaene, 2016)

Dans une première étude en IRMf à haute résolution, 15 mathématiciens professionnels (chercheurs ou professeurs), et 15 sujets contrôles ayant le même niveau universitaire mais n'ayant pas fait d'études mathématiques au-delà du lycée (chercheurs ou professeurs en littérature, histoire, linguistique, etc…), ont écouté 72 affirmations mathématiques (en analyse, algèbre, topologie et géométrie) et non-mathématiques de haut niveau et ont dû déterminer en quelques secondes si elles étaient vraies, fausses ou dépourvues de sens.

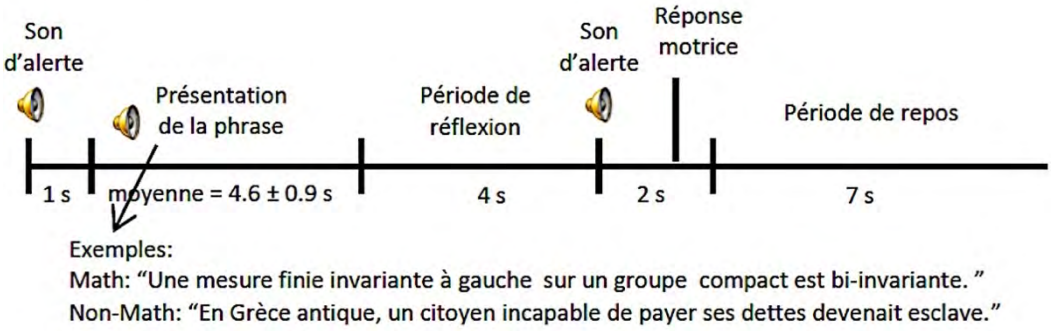



**Figure 1** Schéma d'essai d'un protocole auditif de l'expérience 1

Les résultats comportementaux ont permis de vérifier que les affirmations mathématiques et non-mathématiques avaient des difficultés équivalentes. En effet, le taux de bonnes réponses des mathématiciens aux affirmations mathématiques et non-mathématiques étaient identiques. Dans un second temps, des images de maisons, outils, visages, corps humains, mots, nombres, fragments d'équations et damiers circulaires (servant de contrôle pour l'activation des aires visuelles primaires) ont été présentées par blocs de 8 images d'une même catégorie. Les participants ont dû appuyer sur un bouton lorsqu'ils détectaient la répétition consécutive d'une même image.

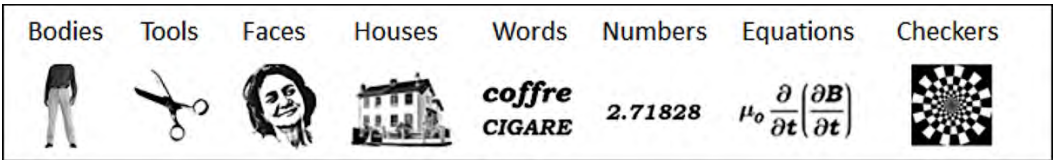



**Figure 2** Catégories visuelles

Enfin, dans les 5 dernières minutes de l'examen IRM, les participants ont, entre autres, fait du calcul mental simple (des soustractions présentées oralement ou visuellement) et lu ou écouté des phrases non-mathématiques simples (voir (Pinel et al., 2007) pour plus de détails).

Le contraste des affirmations mathématiques moins les affirmations non-mathématiques pendant la phase de réflexion a alors permis de mettre au jour un ensemble d'aires cérébrales (en rouge sur la figure 3) impliquées dans la réflexion mathématique de haut niveau chez les mathématiciens professionnels. Cet ensemble est constitué, bilatéralement, des sillons intrapariétaux, de régions temporales inférieures et de régions frontales. L'analyse du décours temporel de l'activation au sein de chacune de ces régions a révélé qu'elles étaient systématiquement activées par chacun des domaines mathématiques testés, analyse, algèbre, topologie et géométrie. Au contraire, aucune activation n'y était visible en réponse aux affirmations non-mathématiques.

De plus, cet ensemble de régions cérébrales n'a été identifié dans cette expérience que chez les mathématiciens. En effet, les affirmations mathématiques ont été traitées de manière similaire aux affirmations non-mathématiques dépourvues de sens par les sujets contrôles. Cela n'est pas surprenant, mais tend à montrer qu'il ne suffit pas de savoir que le contenu présenté est un contenu mathématique pour entraîner une activation des régions cérébrales ci-dessus.

En revanche, aucune différence n'a été observée entre nos deux groupes de participants lorsqu'il s'agissait de faire du calcul mental simple. En accord avec de précédentes études, nous avons confirmé que cette tâche entraîne notamment une activation des sillons intrapariétaux bilatéraux et de régions préfrontales.

De plus, nous avons pu observer, pour la première fois en IRMf, une activation des régions temporales inférieures désignées comme « aires de la forme visuelle des nombres », en réponse à cette tâche de calcul mental (cf. la carte cérébrale bleue de la figure 3). Par ailleurs, la simple exposition visuelle à des images de nombres versus toutes les autres catégories visuelles présentées entraîne également des activations cérébrales similaires chez les deux groupes de participants.

Ces activations se situent, à nouveau, dans les sillons intrapariétaux bilatéraux, des régions préfrontales et des régions temporales inférieures bilatérales (cf. la carte verte de la figure 3). Comme le révèle la figure 3, toutes les tâches portant sur un contenu mathématique dans cette expérience activent des régions dont le recouvrement et l'intersection sont remarquables. En d'autres termes, la réflexion mathématique sur des concepts de haut niveau semble activer les mêmes régions cérébrales que des tâches mathématiques beaucoup plus basiques. Nous avons par ailleurs vérifié que ce recouvrement n'était pas simplement dû à la présence de nombres au sein des affirmations mathématiques. Bien que nous ayons été très vigilants à ce que nos phrases ne contiennent aucune mention directe de nombres, certaines présentaient une référence indirecte aux nombres ou aux fractions

(ex: ℝ2, sphère unité, demi grand axe, etc). Après avoir éliminé toutes ces affirmations, nous avons donc ré-analysé le contraste math > non- math et avons obtenu des résultats tout à fait comparables à ceux précédemment obtenus. De plus, nous avons réalisé une analyse dite « representational similarity » (RSA) à l'échelle individuelle. Au sein d'une région donnée, identifiée comme répondant aux mathématiques à l'échelle du groupe, nous avons vérifié que l'activation entraînée par la réflexion mathématique était plus semblable à l'activation entraînée par les autres stimuli mathématiques (nombres, formules et calcul mental) qu'à l'activation entraînée par les stimuli (phrases et images) non-mathématiques.

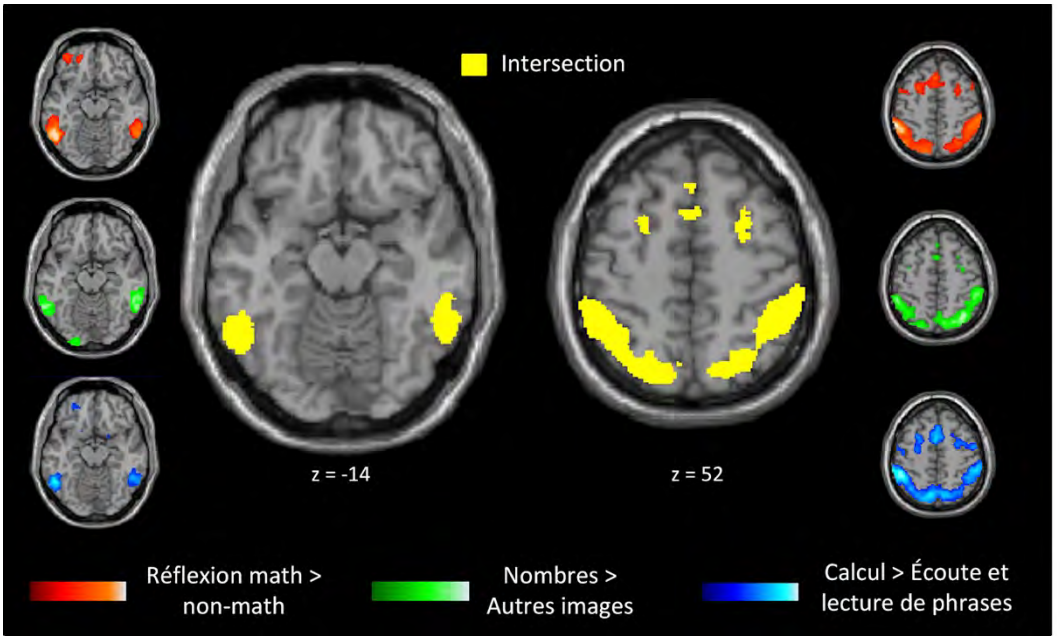



**Figure 3** Cartes montrant les contrastes principaux de l'expérience 1 et leur intersection

*Ces observations coïncident avec la théorie du recyclage neuronal, développée par Stanislas Dehaene, et qui stipule que les activités culturelles de haut niveau, telles que les mathématiques, recyclent des fondations cérébrales très anciennes dans l'évolution, telles que le sens du nombre, de l'espace ou du temps.*

## 5. Organisation de la voie visuelle ventrale chez les mathématiciens professionnels (Amalric and Dehaene, 2016)

L'analyse des contrastes de chaque catégorie d'images versus toutes les autres a révélé une mosaïque typique de réponses préférentielles à chaque catégorie dans le cortex occipito-temporal ventral chez nos deux groupes de participants. Nous avons en particulier observé

la zone de reconnaissance des visages du gyrus fusiforme droit (FFA, en rouge sur la figure 4), les zones parahippocampiques bilatérales de reconnaissance des lieux (PPA, en jaune), les zones de reconnaissance des corps humains du cortex extrastrié (EBA, en orange), les cortex occipitaux latéraux bilatéraux de reconnaissance des outils (LOC, en violet) et l'aire de la forme visuelle des mots dans l'hémisphérique gauche (VWFA, en rose). Comme cela a déjà été évoqué plus haut, nous avons également trouvé, grâce à l'IRMf à haute résolution, une forte activation liée au nombre dans les régions bilatérales du gyrus temporal inférieur (en bleu), aux sites correspondants aux zones de la forme visuelle des nombres gauche et droite (VNFA). Nous avons également observé des réponses préférentielles aux formules (en vert) chez nos deux groupes, à des sites bilatéraux chevauchant partiellement la VNFA.

Étant donné que le traitement des mathématiques avancées recrute des aires ventrales du gyrus temporal inférieur, une question était de savoir dans quelle mesure l'activation de ces régions varie avec l'expertise mathématique. Chez les mathématiciens, par rapport aux sujets contrôles, nous avons observé : (1) une augmentation de l'intensité et de l'étendue de la réponse aux formules mathématiques dans le gyrus temporal inférieur gauche ; (2) dans une moindre mesure une augmentation de l'intensité de la réponse aux nombres au pic de la VNFA gauche; (3) une diminution de la réponse aux visages dans l'hémisphère droit.

Ces résultats soutiennent à nouveau l'idée d'une forme de recyclage neuronal de régions qui préexistent. Les représentations des nombres s'inscrivent ainsi, chez tous, dans des sous-régions occipito-temporales ventrales, remarquablement dissociées de l'expertise pour la lecture. Puis l'expertise mathématique entraîne l'acquisition de symboles nouveaux dont la représentation vient augmenter l'activité de ces sous-régions.

## 6. Mathématiques et langage sont sémantiquement distincts dans le cerveau (Amalric and Dehaene,2018, 2016, submitted)

Cette première expérience a également montré que, bien que les affirmations mathématiques présentées prennent la forme de phrases, les aires cérébrales activées par la réflexion sur des problèmes mathématiques (en bleu sur la figure 4) ne présentaient aucun recouvrement avec les aires du langage (identifiées indépendamment par l'activation engendrée par la lecture et l'écoute de phrases non-mathématiques simples dans la dernière partie de l'expérience 1). A l'inverse, lorsque la réflexion des mathématiciens portaient sur des problèmes d'histoire ou de géographie, le réseau d'aires cérébrales qui s'activaient (en vert sur la figure 4) était complètement différent des régions mathématiques et impliquait certaines aires du langage. Cette observation a ensuite été confirmée par la réalisation d'une analyse fine, à l'échelle individuelle, de l'activation entraînée par les affirmations mathématiques et non-mathématiques au sein de 7 régions d'intérêt présentant une réponse à la syntaxe de la langue d'après l'étude de Pallier et collaborateurs (2011). Dans chacune de ces régions, nous avons pu vérifier que les affirmations non-mathématiques entraînaient plus d'activation que les affirmations mathématiques.

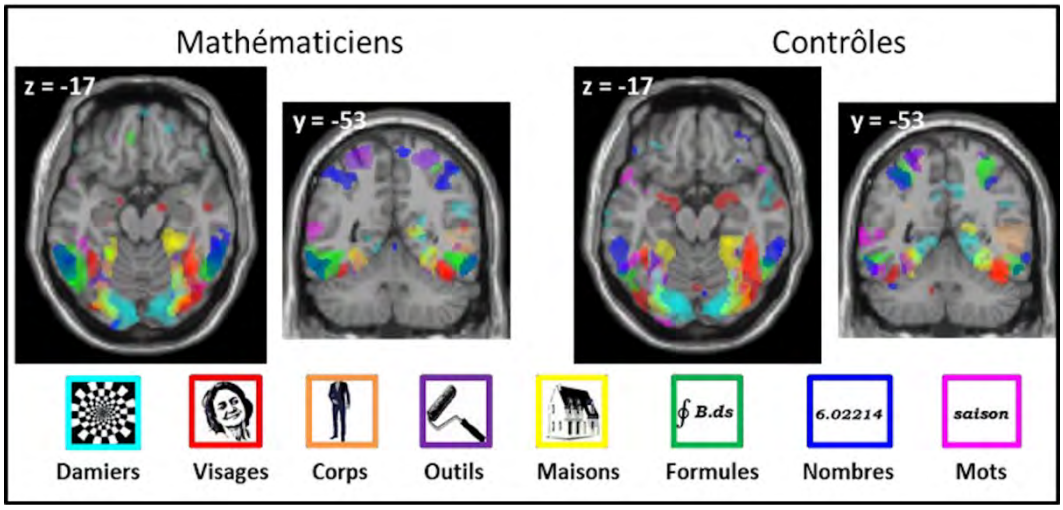



**Figure 4** Mosaïque ventrale de réponses préférentielles aux catégories d'images présentées dans l'expérience 1

*Nous avons donc observé une séparation complète entre le traitement cérébral des affirmations mathématiques et le traitement cérébral des affirmations non-mathématiques, et nous avons dans un deuxième temps voulu tester si cette séparation pouvait émaner du seul contenu sémantique des affirmations. Dans deux nouvelles expériences, nous avons testé plusieurs alternatives possibles à cette hypothèse. D'abord, après avoir constaté que le réseau d'aires qui répond aux mathématiques coïncide en partie avec le « système de demande multiple » décrit par Duncan (2010), une question légitime était de savoir si les problèmes mathématiques requièrent intrinsèquement plus d'attention et de ressources cognitives que les problèmes non-mathématiques. A l'inverse, se pourrait-il que certains problèmes mathématiques connus par cœur ou reposant sur la seule visualisation de la solution court-circuitent le réseau d'aires répondant aux mathématiques ? Par ailleurs, les différences syntaxiques entre les problèmes mathématiques et non-mathématiques de notre première expérience pourraient-elles induire des stratégies de résolution intrinsèquement différentes? Enfin, quels facteurs déterminent ce qui active les réseaux du langage et des mathématiques? L'algèbre est-elle intrinsèquement reliée au traitement du langage? La présence d'opérateurs logiques ou numériques suffit-elle à activer le réseau répondant aux mathématiques même dans des phrases non-mathématiques*

Un nouveau groupe de 14 mathématiciens professionnels a donc participé à deux expériences contrôles en IRMf, basées sur un protocole tout à fait similaire à l'expérience 1 : les participants devaient écouter des affirmations mathématiques et non-mathématiques et décider si elles étaient vraies ou fausses. Cette fois, ils ne disposaient que de 2.5 secondes à l'issue de la présentation de la phrase pour répondre, et plus aucune affirmation n'était dépourvue de sens. De plus, les affirmations ne consistaient qu'en des faits mathématiques

et non-mathématiques simples. Dans l'expérience 2, nous avons proposé aux participants de réfléchir à des faits algébriques connus par cœur (telles que les identités remarquables), du calcul algébrique simple, des problèmes de trigonométrie exprimés avec des cosinus et sinus ou avec des nombres complexes et nécessitant la visualisation du cercle trigonométrique, des problèmes de géométrie non métrique, et des faits non-mathématiques simples portant sur les arts. Dans l'expérience 3, nous avons réduit les affirmations à leur plus simple expression en utilisant seulement l'auxiliaire être, et avons été très vigilants à ce que la syntaxe des affirmations mathématiques et non-mathématiques soit strictement identique. Nous avons inclus de simples déclaratives affirmatives (« La fonction sinus est périodique » ; « Les bus londoniens sont rouges »), quantifiées (« Certaines matrices sont diagonalisables » ; « Certains romans sont autobiographiques »), négatives (« La fonction exponentielle n'est pas constante » ; « La forêt amazonienne n'est pas désertique »), négatives et quantifiées (« Certains ensembles finis ne sont pas dénombrables » ; « Certaines plantes vertes ne sont pas grimpantes »).

La *figure 5* montre que dans ces deux expériences, la dissociation de traitement cérébral des affirmations mathématiques et non-mathématiques est répliquée, même lorsque celles-ci sont très simples, et même lorsque leur syntaxe est parfaitement identique. Autre fait remarquable, les activations préfrontales observées dans l'expérience 1 pour le contraste math > non-math semblent disparaître à mesure que les affirmations deviennent plus simples. Les sillons intrapariétaux et les régions inférieures temporales semblent, eux, constituer un noyau d'aires systématiquement activées par les mathématiques. Un nouveau type d'analyse, consistant à moyenner sur plusieurs régions d'intérêt l'intensité d'activation pour chacune des affirmations, a révélé qu'aucune affirmation non-mathématique n'entraîne l'activation du sus-désigné noyau d'aires cérébrales. À l'inverse, dans les régions d'intérêt liées à la syntaxe du langage, certains domaines mathématiques tels que l'algèbre ou la trigonométrie n'entraînent que très peu voire pas du tout d'activation.

Par ailleurs, nous avons pu vérifier que même les problèmes mathématiques connus par cœur ou nécessitant une certaine imagerie visuelle ne court-circuitent pas pour autant le réseau d'aires cérébrales activé par les mathématiques. Au contraire, toutes les catégories testées dans l'expérience 2 entraînent des activations similaires dans les sillons intrapariétaux et les régions temporales inférieures.

Enfin, les opérateurs logiques minimaux tels que les quantifieurs et la négation n'entraînent pas à eux seuls l'activation des régions répondant aux mathématiques : l'effet principal des quantificateurs se situe dans le gyrus angulaire droit, et l'effet principal de la négation se situe dans une région frontale inférieure gauche, deux régions qui sont en dehors du réseau répondant aux mathématiques.

*Tous ensembles, ces résultats soutiennent l'idée que la sémantique mathématique est fondamentalement dissociée du reste de la sémantique dans le cerveau, et repose sur un*

*noyau d'aires cérébrales constitué des sillons intrapariétaux et de régions temporales inférieures.*

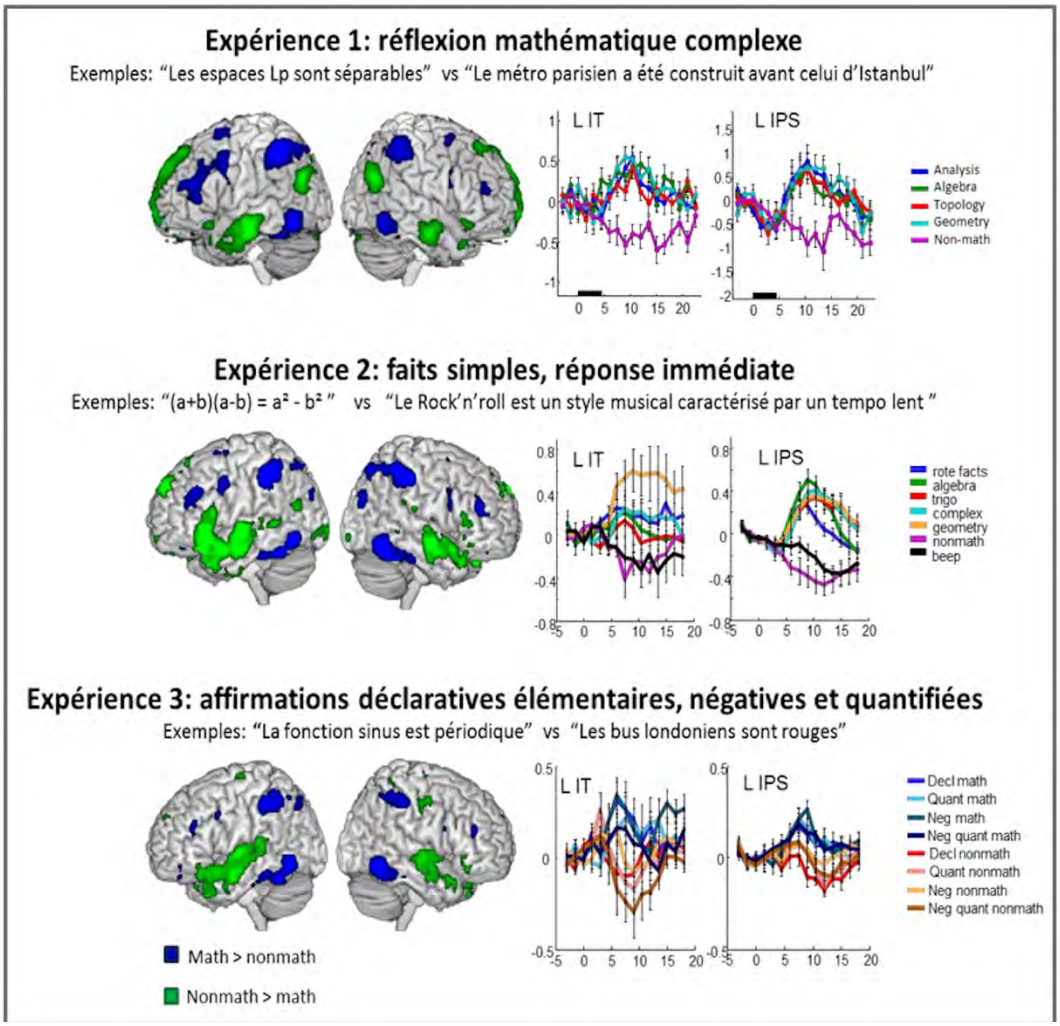



**Figure 5** Comparaison des contraste math-> non-math et non-math>math dans les trois expériences d'IRMf et décours temporels de l'activation dans les principales régions répondant aux mathématiques

## 7. Un langage mathématique sans mots pour le dire (Amalric et al., 2017b ; Wang et al, submitted)

Ce résultat concorde avec d'autres observations. Notamment, disposer d'un vocabulaire mathématique ne se révèle pas être absolument nécessaire à la compréhension de notions

mathématiques. Typiquement, les études menées en Amazonie auprès des indiens Mundurukus, qui disposent d'un vocabulaire numérique et géométrique très pauvre et d'un accès limité à l'éducation, ont révélé qu'ils sont tout à fait capables de réaliser des opérations arithmétiques, ou d'identifier des propriétés géométriques du plan et de l'espace (Dehaene et al., 2008, 2006; Pica et al., 2004). Ces études suggèrent l'existence de primitives des mathématiques en l'absence de langage. On peut ne pas savoir nommer un carré et pourtant posséder le concept de carré.

La question qui se pose alors est de savoir s'il existe également un langage mathématique élémentaire, même lorsque l'on n'a pas été éduqué, même lorsque l'on n'a pas de mots pour le dire. Au cours de ma thèse, j'ai conçu une situation suffisamment simple pour être présentée à de jeunes enfants ou à des indiens Mundurukus, mais qui requiert une sorte de langage de l'esprit (Fodor, 1975), en l'occurrence un « langage de la géométrie ». Cette situation demande de regarder une séquence spatiale présentant un certain nombre de positions successives, d'en prédire la suite et d'en mémoriser l'ensemble. La prédiction s'appuie sur la détection de régularités géométriques dans la séquence : symétries, rotations, formes… (cf. ***figure 6***) Et la présence de régularités facilite également la mémorisation.

La tâche proposée aux sujets consistaient, après avoir vu les premiers points de la séquence, à désigner les suivants. À chaque erreur, la séquence recommençait au début jusqu'à corriger le point erroné. Afin de tester l'existence d'un langage géométrique spontané, il était important que les participants soient non seulement des adultes français (n=23), mais aussi des enfants de 5 ans qui n'ont pas encore eu d'éducation mathématique (n=47), ainsi que des adultes Mundurukus d'Amazonie qui n'ont pas eu d'éducation mathématique du tout (n=14).

En se servant de la séquence « irregular », qui ne présente aucune régularité apparente, comme référence du degré d'apprentissage par cœur des séquences, l'analyse des erreurs commises à des points clés des séquences « repeat », « alternate », « repeat +2 », « 4segments » et « 4diagonals » a permis de montrer que tous les participants sont capables d'identifier et d'utiliser rapidement des rotations et des symétries axiales afin de compléter correctement les séquences avant même qu'elles n'aient été entièrement présentées. Seule la symétrie centrale a semblé plus difficile à comprendre en l'absence d'éducation (chez les enfants et les Mundurukus). De plus, l'analyse des erreurs commises aux points 5 et 13 des séquences « 2arcs » et « 2squares » (qui correspondent à l'application de la règle indiquant comment changer le point de départ de l'arc ou du carré), ont montré que tous les participants sont également capables de détecter rapidement des structures enchâssées. Les séquences « 2rectangles » et « 2crosses » qui contiennent un niveau d'enchâssement supplémentaire se sont révélées plus difficiles pour les adultes français et Mundurukus, voire complètement inintelligibles pour les jeunes enfants. Nous avons ainsi pu montrer que le langage requis pour décrire les séquences observées, contient à la fois des primitives géométriques (symétries, rotations) et la capacité de combiner ces primitives sous forme de règles de répétition.

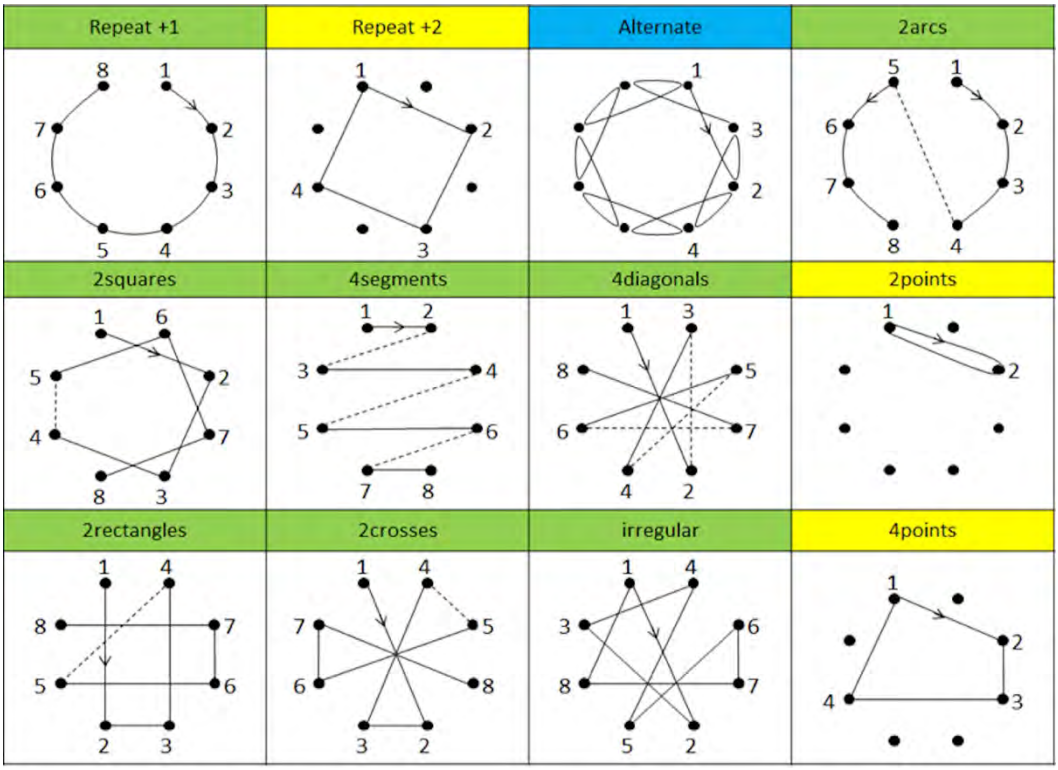



**Figure 6** Séquences présentées aux adultes (en vert et en bleu) et aux enfants (en vert et en jaune)

Dans ce langage formel, chaque séquence a une certaine complexité, associée à la longueur de sa description minimale (Romano et al., 2013). Nous avons pu vérifier que notre langage géométrique et les complexités des séquences prédisent efficacement les erreurs commises par les participants dans les tâches de prédiction et de mémorisation des séquences (cf. *figure 7*).

À partir du langage géométrique ainsi créé, nous avons ensuite développé un modèle computationnel de l'apprentissage des séquences géométriques. Comme représenté sur la figure 8, l'algorithme prend en entrée les premiers points de la séquence présentés au sujet, puis crée la liste de toutes les séquences possibles commençant par ces points et leur associe leur expression (appelée programme sur la *figure 8*) dans le langage prédéterminé. Pour refléter le fait que plus la longueur de l'expression augmente, plus la probabilité que les sujets se trompent est grande, notre algorithme évalue la complexité de chaque programme avec un bruit gaussien. L'algorithme choisit ensuite le programme qui minimise la complexité. Afin d'éviter que la performance au dernier point soit toujours parfaite, on suppose que le modèle ne peut identifier des expressions qu'en deçà d'une complexité seuil. Au-delà, le point suivant est choisi aléatoirement. Sinon, le point suivant est défini par le programme choisi.

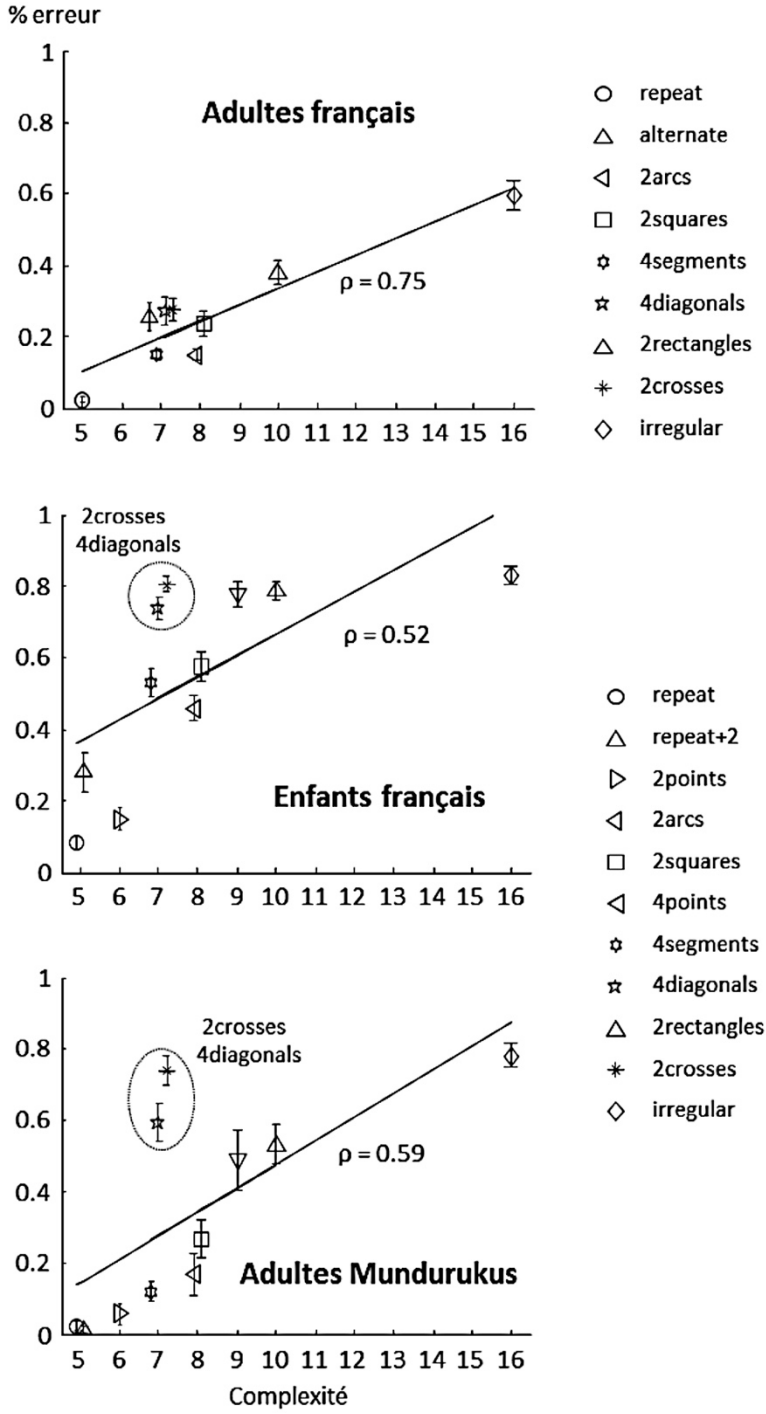



**Figure 7** Corrélation du taux d'erreur avec la complexité des séquences

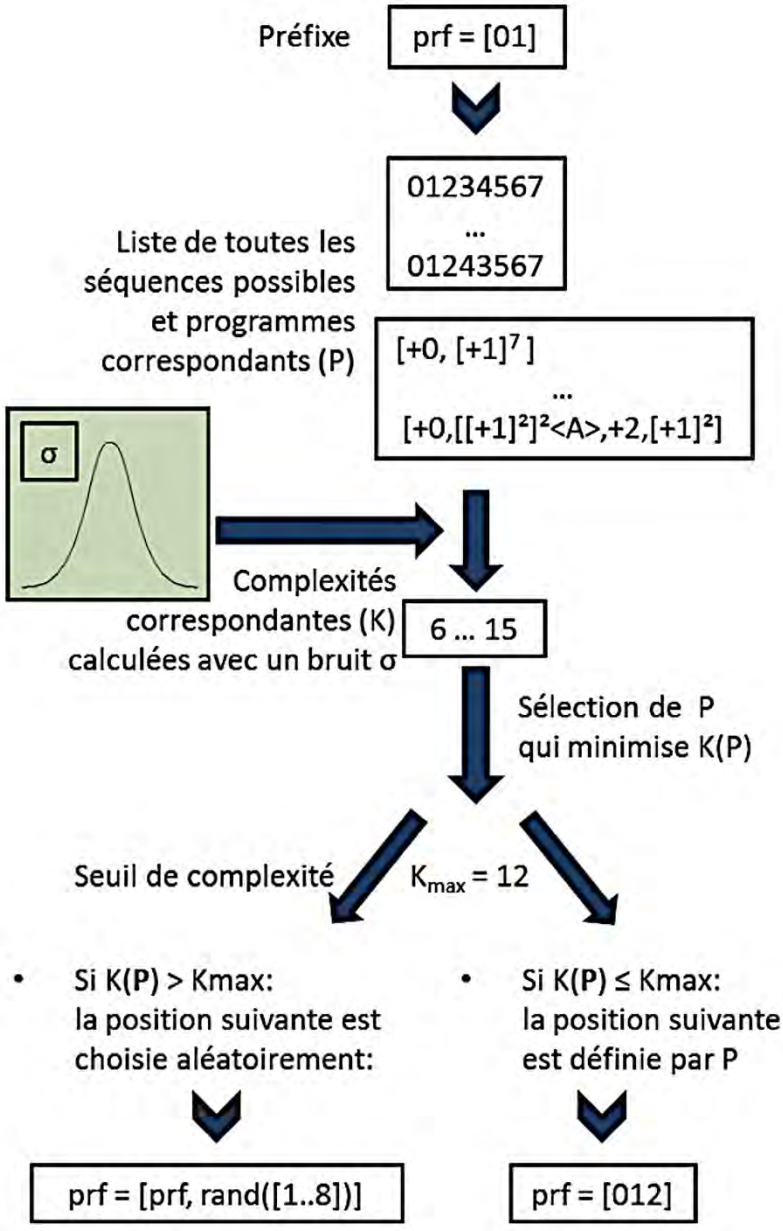



**Figure 8** Schéma de l'algorithme utilisé

Pour évaluer ce modèle, nous avons seulement considéré les 8 séquences présentées dans tous les groupes (en vert sur la figure 6) et qui ne présentent pas de répétition interne. La détermination du seul paramètre σ a suffi à modéliser les données des adultes de manière tout à fait remarquable. En revanche, pour reproduire le comportement des enfants et des Mundurukus de façon satisfaisante, il a fallu considérer un langage restreint, sans symétrie centrale et sans la capacité d'encoder plus d'un niveau de répétition.

Ces résultats semblent indiquer d'une part que le cerveau humain possède une forme de langage de la pensée géométrique indépendant de tout langage parlé naturel, présent spontanément mais qui se raffine avec l'âge et l'éducation, et d'autre part que le cerveau cherche dans ce langage l'expression la plus courte possible qui permet de rendre compte de ce qui a été observé.

Afin de tester quelles régions cérébrales s'activent au cours de l'apprentissage des séquences, Liping Wang et moi avons ensuite adapté ce protocole comportemental à l'expérimentation en IRMf. Nous avons proposé à 20 adultes français non plus de désigner les positions successives, mais de suivre un point mobile du regard en essayant d'anticiper sa position. Grâce à un dispositif de suivi oculaire, nous avons pu calculer un indice d'anticipation du regard, et avons vérifié que les effets d'anticipation étaient un bon marqueur de la compréhension implicite des séquences. En effet, les données oculaires étaient tout à fait similaires aux données issues du pointage, et l'indice d'anticipation global était fortement corrélé à la complexité des séquences.

Nous avons par la suite étudié dans quelles régions du cerveau l'activité corrélait avec la complexité des séquences. Après avoir éliminé l'effet de la mémoire de travail, nous avons pu observer une activation dans la partie supérieure du gyrus frontal inférieur, région qui appartient au réseau impliqué dans le traitement mathématique identifié dans mes précédentes expériences.

Nous avons ensuite quantifié l'anticipation des structures enchâssées, en faisant la différence entre les points 5, qui correspondent à la règle de second niveau, et les points 3 et 7, qui correspondent à la règle de premier niveau. Les régions cérébrales dont l'activité corrélait avec cet indice sont représentées en jaune sur la ***figure 9*** et incluent des régions pariétales bilatérales, préfrontales et temporales inférieures droites. En comparant l'intensité de cette activation à l'intérieur de régions d'intérêt du langage (en rouge) et des mathématiques (en bleu), nous avons finalement observé que l'anticipation de l'enchâssement repose davantage sur l'activation du réseau de traitement mathématique que du réseau de traitement linguistique.

*Ces résultats suggèrent l'existence d'un « langage de la pensée » de nature géométrique, indépendant du langage parlé naturel.*

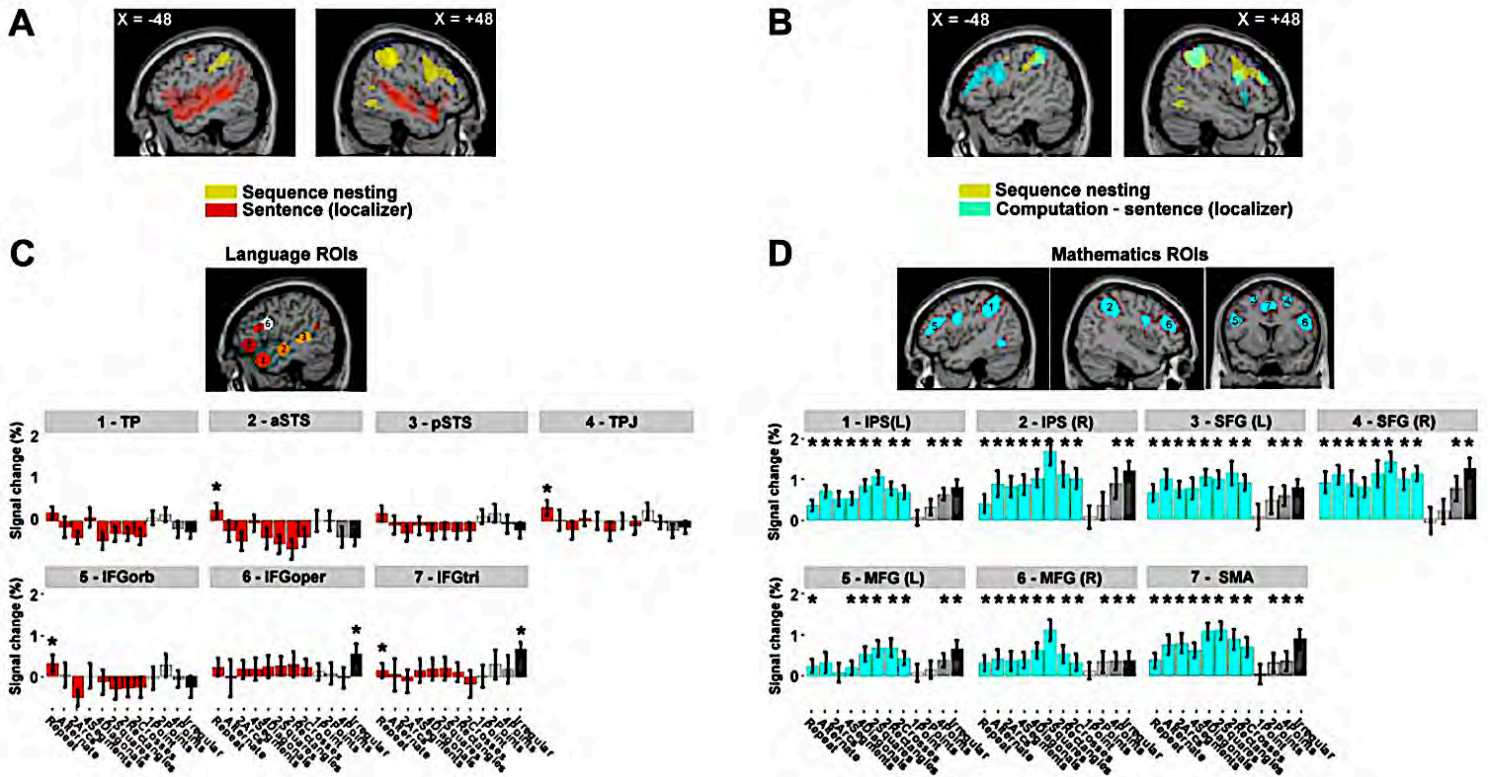



**Figure 9** Cartes corticales de l' activité évoquée par l' anticipation de l' enchâssement (en jaune), superposées aux régions d' intérêt du langage (A) et des mathématiques (B) par l' écoute des phrases et le calcul mental. (C) et (D) montrent les résultats de l' analyse individuelle effectuée dans les régions d' intérêt sus-mentionnés

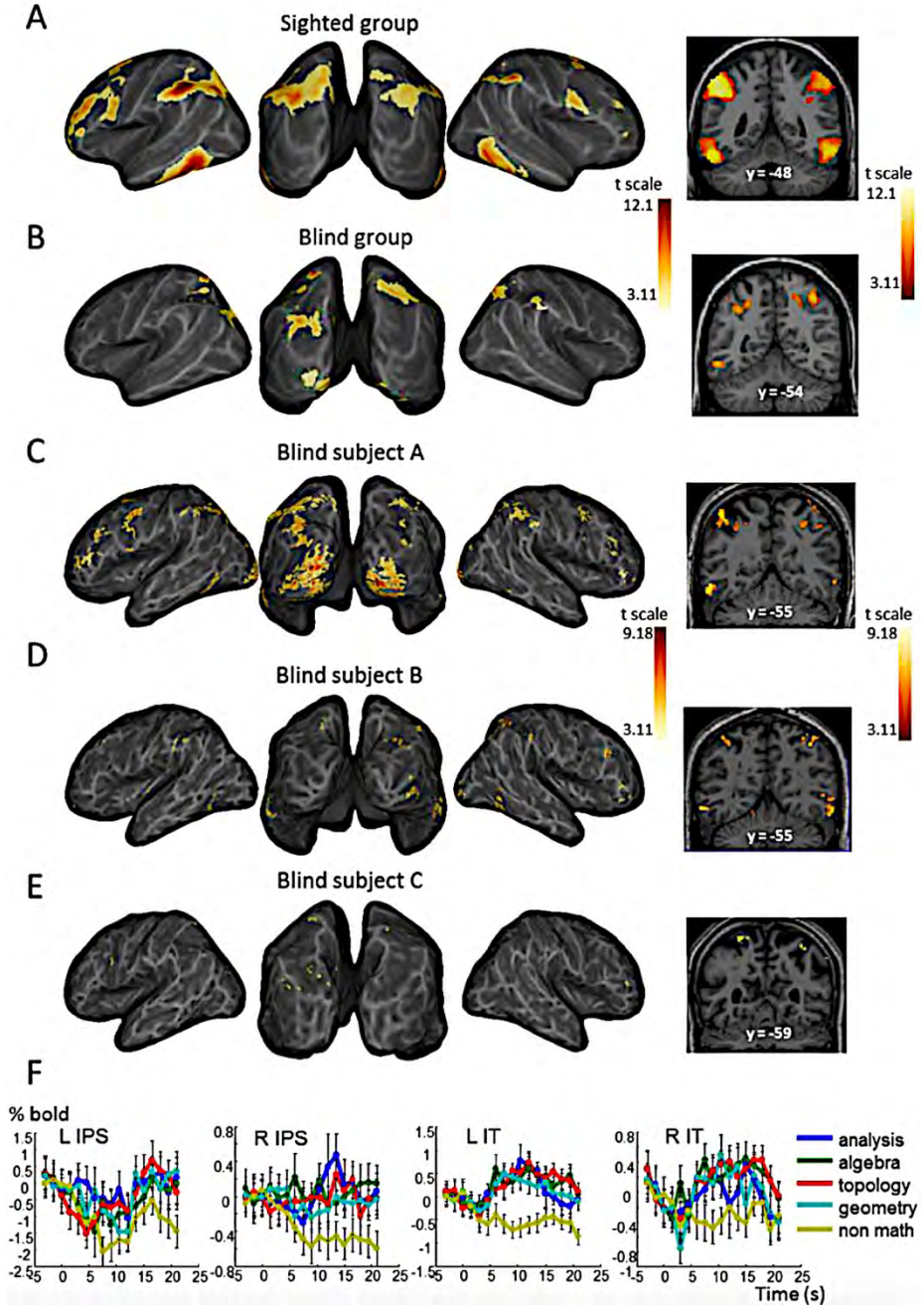
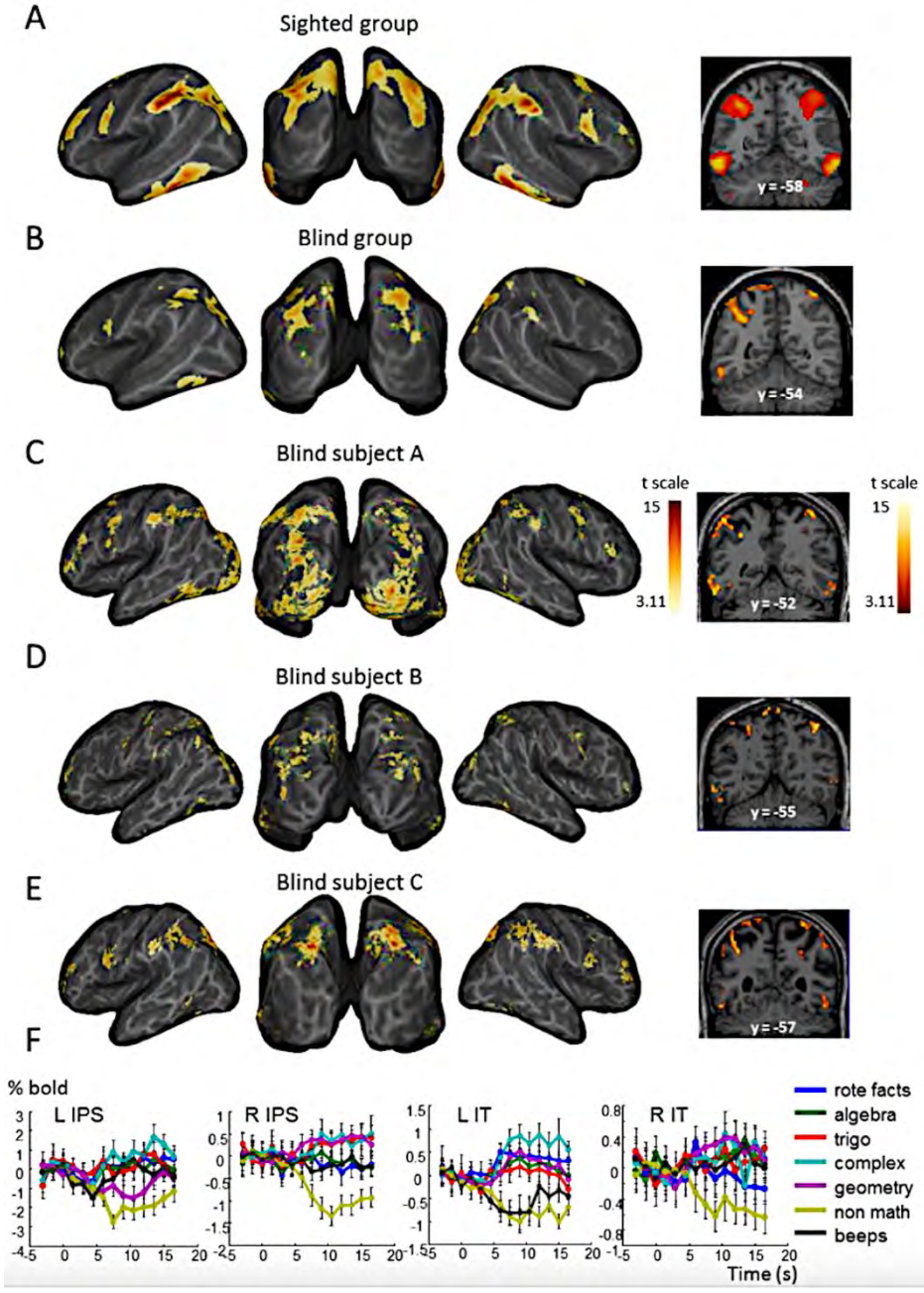



**Figure 10** Activations cérébrales pour le contraste math>non-math chez les voyants (première ligne) et les aveugles (quatre dernières lignes). Décours temporel de l'activation chez le sujet aveugle A (troisième ligne) dans les principales régions répondant aux mathématiques, pour différentes disciplines (analyse, algèbre, topologie, géométrie, énoncés non mathématiques).
Page de gauche : Expérience 1. Page de droite : Expérience 2.


A
Sighted group
y = -58
B
Blind group
y = -54
C
Blind subject A
t scale
15
3.11
y = -52
t scale
15
3.11
D
Blind subject B
y = -55
E
Blind subject C
y = -57
F
% bold
L IPS
R IPS
L IT
R IT
rote facts
algebra
trigo
complex
geometry
non math
beeps
Time (s)

## 8. L'expérience visuelle est-elle nécessaire pour développer des connaissances mathématiques (Amalric et al., 2017a)

Les résultats décrits jusqu'à présent semblent donner raison à l'introspection d'Albert Einstein selon laquelle les mots et le langage n'interviennent pas dans la réflexion mathématique. Avait-il également raison de penser que les images mentales constituent le support privilégié du traitement du contenu mathématique ?

Tout d'abord, afin d'évaluer si l'expérience visuelle est requise pour développer des concepts mathématiques avancés, nous avons proposé à trois mathématiciens professionnels non-voyants (les sujets A et B étant devenus aveugles respectivement aux âges de 3 ans et 10 ans, et le sujet C étant aveugle de naissance) de réaliser un examen IRMf utilisant les protocoles des expériences 1 et 2.

Comme le révèle la *figure 10*, des activations similaires à celles observées chez les mathématiciens voyants ont été observées chez les mathématiciens non-voyants pour le contraste math > non-math. En particulier dans la seconde expérience, portant sur des concepts plus simples que la première, nous avons pu observer l'activation des sillons intrapariétaux bilatéraux et des régions temporales inférieures bilatérales chez chacun des sujets.

*Ces résultats réfutent l'hypothèse d'un lien entre expertise mathématique et expérience visuelle. Au contraire, ils suggèrent que la représentation corticale des mathématiques de haut niveau, qui implique de manière fondamentale les sillons intrapariétaux et les régions temporales inférieures, peut se développer indépendamment de toute expérience visuelle.*

Toutefois, nos résultats n'excluent pas la possibilité d'une forme d'imagerie mentale, aussi bien chez les voyants que chez les non-voyants, nécessaire au traitement de certaines notions mathématiques. Tout d'abord, à l'issue de l'expérience 1, chaque participant a été invité à revoir l'ensemble des affirmations présentées au cours de l'examen IRMf, et à évaluer notamment le niveau d'imagerie visuelle suscitée par chacune d'entre elles. Au sein du groupe des mathématiciens voyants, une région temporale inférieure gauche et une zone du sillon intra-occipital gauche (i.e. deux zones du cortex visuel) ont montré une activation positivement corrélée à la mesure subjective d'imagerie mentale. De plus, dans l'expérience 2, toujours chez les mathématiciens voyants, nous avons observé que des aires cérébrales de traitement visuel telles que la scissure calcarine s'activent lorsque les notions mathématiques abordées impliquent une certaine visualisation de la solution sur le cercle trigonométrique. Enfin, chez les mathématiciens non-voyants, une activation additionnelle par rapport aux voyants a été observée dans le cortex occipital (cf. figure 11). Cette activation est particulièrement intense chez le sujet A qui a également rapporté le plus fort taux d'imagerie mentale à l'issue de l'expérience 1. Même s'il est impossible d'établir une quelconque conclusion sur la base de l'étude d'un groupe de seulement trois individus dont les causes de la cécité sont qui plus est tout à fait différentes, il est tentant de supposer que cette activation occipitale puisse refléter une certaine forme d'imagerie mentale.

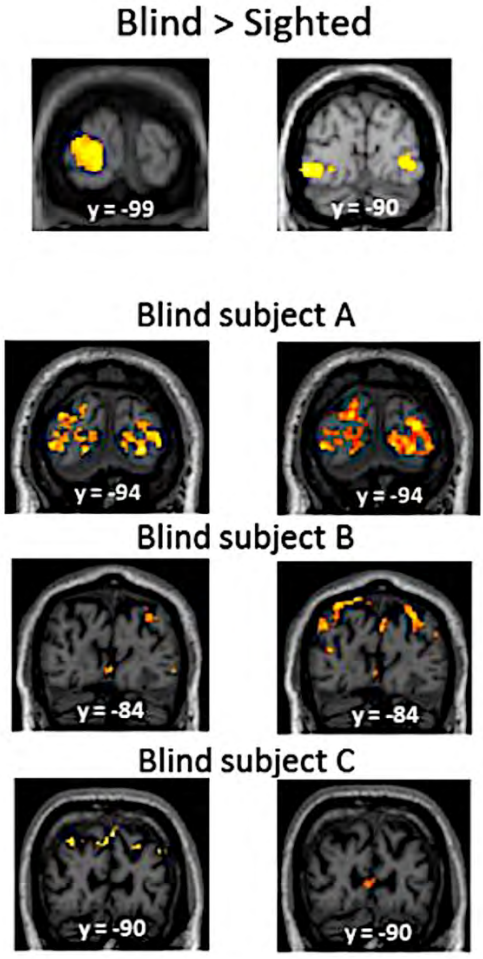



**Figure 11** Activation additionnelle du cortex occipital chez les mathématiciens non voyants

## 9. Conclusions et perspectives

*Quels supports corticaux pour les mathématiques ?*
Au cours de ma thèse, j'ai pu montrer l'existence d'un noyau d'aires cérébrales, bilatéralement composé des sillons intrapariétaux et de régions temporales inférieures, répondant systématiquement aux mathématiques, quels qu'en soient le domaine et le niveau de difficulté, y compris en l'absence d'expérience visuelle. L'ensemble des résultats reportés ici suggère qu'une forme de recyclage neuronal intervient dans le traitement cortical des concepts mathématiques avancés qui semblent se construire sur des fondements proto-mathématiques intuitifs. Les résultats de ma thèse révèlent également que le réseau d'aires répondant aux mathématiques est dissocié des régions cérébrales classiquement impliquées dans le traitement de la syntaxe et de la sémantique du langage.

Cette dissociation semble intervenir aussi bien au niveau sémantique qu'au niveau syntaxique puisque nous avons pu mettre en évidence l'existence d'une forme de langage de la pensée de nature géométrique, indépendant du langage parlé naturel.

L'idée que cette dissociation puisse refléter l'existence de deux systèmes sémantiques distincts dans le cerveau est même soutenue par les résultats obtenus par Huth et collaborateurs en 2016. En effet, en appliquant de nouvelles méthodes d'analyse de données IRMf enregistrées alors que des participants écoutaient des histoires à contenu très varié, les auteurs ont identifié les deux premières composantes principales de la variation de l'activité cérébrale liée à la sémantique des mots. Or, les deux réseaux sémantiques distincts ainsi identifiés correspondent presque parfaitement aux deux ensembles d'aires cérébrales distincts obtenus en réponse respectivement aux stimuli mathématiques et non-mathématiques dans nos expériences.

*Le langage dans l'apprentissage des mathématiques.*
Il est important de noter que les mathématiciens ayant participé à nos études d'imagerie cérébrale avaient tous bénéficié de nombreuses années d'étude préalables des mathématiques. Nous ne pouvons donc que conclure que l'utilisation et le traitement de notions mathématiques bien connues se passent du langage parlé naturel. Toutefois, si on peut supposer qu'une fois acquis, les concepts mathématiques sont encodés de manière abstraite, symbolique, et ne font plus appel au langage, celui-ci pourrait au contraire jouer un rôle important dans leur apprentissage. C'est donc vers l'étude des relations entre mathématiques et langage dans le contexte de l'apprentissage des mathématiques à l'école, que mon travail se tourne désormais.

*Quelle est la nature des activités mathématiques ?*
Enfin, mes résultats d'IRMf soulèvent de nombreuses questions quant à la définition exacte des processus linguistiques et mathématiques dans le cerveau. Tout d'abord que signifie exactement faire des mathématiques pour le cerveau humain ? Si les sillons intrapariétaux et les régions temporales inférieures s'activent systématiquement en réponse à tout stimulus mathématique, ils ne sont pour autant pas spécifiques des mathématiques et s'activent également dans diverses tâches impliquant un raisonnement logique (Goel and Dolan, 2001; Monti et al., 2009), dans des tests de QI (Duncan, 2010) ou encore dans la représentation de concepts de physique (Mason and Just, 2016). La question reste ouverte de déterminer quel point commun à ces tâches entraîne l'activation de structures cérébrales similaires. Enfin, il est important de remarquer que d'un côté, on ne peut se passer de langage pour communiquer les résultats mathématiques, et de l'autre, le langage courant est envahi de termes mathématiques (nombres, unités, ensembles, quantificateurs, prépositions spatiales, etc…). Où se trouve donc la frontière entre les processus linguistiques et mathématiques dans le cerveau ?

**Références**

Amalric, M., Dehaene, S., 2018. Cortical circuits for mathematical knowledge: evidence for a major subdivision within the brain's semantic networks. Phil Trans R Soc B 373, 20160515. https://doi.org/10.1098/rstb.2016.0515

Amalric, M., Dehaene, S., 2016. Origins of the brain networks for advanced mathematics in expert mathematicians. Proc. Natl. Acad. Sci. 201603205. https://doi.org/10.1073/pnas.1603205113

Amalric, M., Dehaene, S., submitted. A distinct cortical network for mathematical knowledge in the human brain. NeuroImage.

Amalric, M., Denghien, I., Dehaene, S., 2017a. On the role of visual experience in mathematical development: Evidence from blind mathematicians. Dev. Cogn. Neurosci. https://doi.org/10.1016/j.dcn.2017.09.007

Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., Dehaene, S., 2017b. The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. PLOS Comput. Biol. 13, e1005273. https://doi.org/10.1371/journal.pcbi.1005273

Benson-Amram, S., Heinen, V.K., Dryer, S.L., Holekamp, K.E., 2011. Numerical assessment and individual call discrimination by wild spotted hyaenas, Crocuta crocuta. Anim. Behav. 82, 743–752. https://doi.org/10.1016/j.anbehav.2011.07.004

Cantlon, J.F., Merritt, D.J., Brannon, E.M., 2016. Monkeys display classic signatures of human symbolic arithmetic. Anim. Cogn. 19, 405–415. https://doi.org/10.1007/s10071-015-0942-5

Chiandetti, C., Vallortigara, G., 2007. Is there an innate geometric module? Effects of experience with angular geometric cues on spatial re-orientation based on the shape of the environment. Anim. Cogn. 11, 139–146. https://doi.org/10.1007/s10071-007-0099-y

Chomsky, N., 2006. Language and mind, 3rd ed. ed. Cambridge University Press, Cambridge ; New York.

Dehaene, S., Cohen, L., 1995. Towards an Anatomical and Functional Model of Number Processing.

Math. Cogn. 1, 83–120.

Dehaene, S., Izard, V., Pica, P., Spelke, E., 2006. Core Knowledge of Geometry in an Amazonian Indigene Group. Science 311, 381–384. https://doi.org/10.1126/science.1121739

Dehaene, S., Izard, V., Spelke, E., Pica, P., 2008. Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures. Science 320, 1217–1220. https://doi.org/10.1126/science.1156540

Dehaene, S., Piazza, M., Pinel, P., Cohen, L., 2003. THREE PARIETAL CIRCUITS FOR NUMBER PROCESSING.

Cogn. Neuropsychol. 20, 487–506. https://doi.org/10.1080/02643290244000239

Dillon, M.R., Huang, Y., Spelke, E.S., 2013. Core foundations of abstract geometry. Proc. Natl. Acad. Sci.

110, 14191–14195. https://doi.org/10.1073/pnas.1312640110

Duncan, J., 2010. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn. Sci. 14, 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Fodor, J.A., 1975. The Language of Thought. Harvard University Press.

Goel, V., Dolan, R.J., 2001. Functional neuroanatomy of three-term relational reasoning. Neuropsychologia 39, 901–909. https://doi.org/10.1016/S0028-3932(01)00024-0

Hadamard, J., 1975. Essai sur la psychologie de l'invention dans le domaine mathématique.

Halberda, J., Mazzocco, M.M.M., Feigenson, L., 2008. Individual differences in non-verbal number acuity correlate with maths achievement. Nature 455, 665–668. https://doi.org/10.1038/nature07246

Harvey, B.M., Klein, B.P., Petridou, N., Dumoulin, S.O., 2013. Topographic Representation of Numerosity in the Human Parietal Cortex. Science 341, 1123–1126. https://doi.org/10.1126/science.1239052
Hermes, D., Rangarajan, V., Foster, B.L., King, J.-R., Kasikci, I., Miller, K.J., Parvizi, J., 2015. Electrophysiological Responses in the Ventral Temporal Cortex During Reading of Numerals and Calculation. Cereb. Cortex bhv250. https://doi.org/10.1093/cercor/bhv250
Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://doi.org/10.1038/nature17637
Hyde, D.C., Boas, D.A., Blair, C., Carey, S., 2010. Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants. NeuroImage 53, 647–652. https://doi.org/10.1016/j.neuroimage.2010.06.030
Izard, V., Sann, C., Spelke, E.S., Streri, A., 2009. Newborn infants perceive abstract numbers. Proc. Natl.
Acad. Sci. 106, 10382–10385. https://doi.org/10.1073/pnas.0812142106
Lee, S.A., Spelke, E.S., 2008. Children's use of geometry for reorientation. Dev. Sci. 11, 743–749. https://doi.org/10.1111/j.1467-7687.2008.00724.x
Marcus, G.F., Vijayan, S., Rao, S.B., Vishton, P.M., 1999. Rule learning by seven-month-old infants.
Science 283, 77–80.
Mason, R.A., Just, M.A., 2016. Neural Representations of Physics Concepts. Psychol. Sci. 27, 904–913. https://doi.org/10.1177/0956797616641941
Matthews, P.G., Lewis, M.R., Hubbard, E.M., 2016. Individual Differences in Nonsymbolic Ratio Processing Predict Symbolic Math Performance. Psychol. Sci. 27, 191–202. https://doi.org/10.1177/0956797615617799
Monti, M.M., Parsons, L.M., Osherson, D.N., 2012. Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. Psychol. Sci. 23, 914–922. https://doi.org/10.1177/0956797612437427
Monti, M.M., Parsons, L.M., Osherson, D.N., 2009. The boundaries of language and thought in deductive inference. Proc. Natl. Acad. Sci. 106, 12554–12559. https://doi.org/10.1073/pnas.0902422106
Nieder, A., Miller, E.K., 2004. A parieto-frontal network for visual numerical information in the monkey.
Proc. Natl. Acad. Sci. U. S. A. 101, 7457–7462. https://doi.org/10.1073/pnas.0402239101 Pallier, C., Devauchelle, A.-D., Dehaene, S., 2011. Cortical representation of the constituent structure of sentences. Proc. Natl. Acad. Sci. 108, 2522–2527. https://doi.org/10.1073/pnas.1018711108

# DEUXIÈME PARTIE

## SCIENCES COGNITIVES ET INTELLIGENCE ARTIFICIELLE

## Deuxième Partie

## Sciences cognitives et Intelligence Artificielle

# Présentation

*La seconde partie* – Sciences cognitives et Intelligence Artificielle - rassemble quatre textes en rapport avec *l'Intelligence Artificielle (IA)*. Cette discipline fait partie des sciences cognitives. Il est clair en effet en premier lieu, que dès son origine, l'IA s'est posée en « aiguillon » des sciences visant à comprendre les processus cognitifs en œuvre dans le cerveau. La programmation des algorithmes d'IA exige en effet de s'appuyer sur des modèles, de nature logique ou mathématique, des processus cognitifs que l'on veut simuler. Ces modèles peuvent alors être proposés à ces sciences du cerveau et de la cognition. Et la réciproque est également vraie, l'IA ayant, également dès son origine, cherché à intégrer les avancées dans la compréhension et le fonctionnement des structures cérébrales, à différentes échelles. Le premier et le dernier chapitre de cette partie traitent, sous des accents un peu différents mais complémentaires, de l'histoire de l'IA, de ses performances actuelles, de ses perspectives et de sa place dans la société. Les deux chapitres centraux traitent de rapports de l'IA avec le langage humain, dont on dit souvent qu'il a un lien étroit avec la conscience, qu'il en est un corrélat voire une condition. Le premier de ces deux chapitres expose des travaux visant à donner à une machine la capacité de *comprendre* – en un sens fort - le langage humain ; le second présente des travaux proposant des mécanismes à même, au sein d'une *société artificielle*, composée d'individus dotés de capacités d'IA, de faire émerger un langage commun, avec son vocabulaire, sa sémantique, voire sa grammaire.

## L'IA et son histoire

L'IA est devenue une composante majeure de l'utilisation de l'Informatique, elle bénéficie d'une couverture médiatique importante, entrainant peurs ou espoirs parfois inconsidérés. Mais sa nature de *discipline scientifique*, son histoire, ses tâtonnements, la portée de ses succès et leurs limites actuelles, les directions de recherches en cours, peuvent être assez mal connues. Aussi nous a-t-il paru utile de demander à deux chercheurs éminents de l'IA de tracer un tableau de l'évolution de la discipline depuis sa naissance dans les années 1940, évolution qui montre les liens de la discipline avec les sciences du cerveau et les sciences cognitives, mais aussi avec les mathématiques.

- Ainsi dans le premier chapitre, intitulé « Intelligence Artificelle, des Big Data au « Cerveau », Jean-Gabriel Ganascia rappelle d'abord les racines de l'Intelligence Artificielle, plongeant dans la cybernétique et le début des sciences cognitives avec les « conférences Macy » de l'immédiate « après guerre » 1946-1953. Il dresse un bref historique de l'IA depuis sa naissance « officielle » en 1956 ; sont donnés en parallèle des éléments de la montée en puissance des « Big Data ». Il introduit ensuite, de façon illustrée, l'aventure des réseaux neuronaux, depuis les premiers articles parus dans les années 1940 ; y sont présentés certains des principes de leurs performances classificatoires, ainsi que les algorithmes dit « de rétro-propagation du gradient », qui constituent une des bases de l'apprentissage profond. Avec, en arrière-plan, l'émergence de la disponibilité de grandes masses de données, à une échelle jamais connue dans l'histoire de l'Humanité, grandes masses de données que l'apprentissage profond permet d'exploiter. En conclusion Jean-Gabriel Ganascia aborde la question de la possibilité d'une « Intelligence Artificielle Forte » dépassant les performances de l'actuelle IA faible, qui restent très spécialisées.

- Dans le dernier chapitre de cette partie, intitulé « Les machines pensantes, un panorama de l'Intelligence Artificielle », Jean-Paul Haton retrace d'abord l'historique de l'Intelligence Artificielle ; il énonce les grands types de problèmes auxquels l'IA s'attelle et détaille les différentes pistes de travail qui ont été en œuvre dans cette discipline : approches symboliques, approches neuro-mimétiques, approches probabilistes et statistiques. Les problématiques et méthodes de l'Apprentissage sont ensuite abordées : apprentissage symbolique versus apprentissage numérique, apprentissage supervisé versus apprentissage non supervisé, apprentissage par renforcement. Puis est dressé un panorama des réseaux de neurones profonds : conditions hardware et software qui en ont permis l'émergence et les performances actuelles, différents types de réseaux utilisés, leurs limites et les domaines d'applications. Après un rappel des grands acteurs privés ou institutionnels concernés, Jean-Paul Haton termine en évoquant les tendances actuelles, qui ne se réduisent pas aux réseaux neuronaux profonds, en direction de l'Intelligence Artificielle Forte. Ainsi que les problèmes éthiques, sociétaux, juridiques, que les progrès déjà en place ou attendus ne manquent pas de provoquer.

## L'IA, Conscience et Langage humain

S'il existe donc une Intelligence Artificielle, il n'existe pas – pour le moment – de *Conscience Artificielle.* Les deux tableaux historiques se terminent bien par la constatation qu'« Aucun des système d'IA n'est pour l'instant doté de conscience », même si « des modèles de ce que pourrait être une telle conscience ont été proposés ». On retrouve bien dans ces tentatives récentes une continuité avec les origines de l'IA : concevoir des modèles

formels des fonctions du cerveau, pour pouvoir ensuite simuler ces fonctions sur un support *non vivant.* Dans cet ouvrage dédié aux phénomènes conscients, une partie consacrée à l'IA ne pouvait donc concerner – en l'état actuel des choses - que son apport à des aspects considérés comme plus ou moins liés à la conscience, mais qui n'en constituent pas nécessairement le cœur. Avec la mémoire, la capacité d'anticiper les conséquences d'actes potentiels, l'un de ses aspects est *le langage.* C'est par le langage que, par exemple, un sujet manifeste qu'il est conscient d'avoir perçu tel ou tel stimulus ; comme on l'a vu en première partie, le rapport alors effectué sera pris comme assurant la réalité de cette perception consciente. Il était donc logique de s'intéresser à certaines des recherches menées en IA concernant le langage humain :

- Un chapitre sous la signature d'Antoine Bordes est intitulé « Former les machines à la compréhension du langage naturel ». Les performances atteintes dans l'exploitation des bases de données relationnelles par le biais des langages de requêtes, la facilité avec laquelle on accède aux informations en utilisant des mots clés sur les moteurs de recherche, pourraient laisser croire que l'on est proche d'une situation où les algorithmes « comprendront » le langage « naturel ». Comprendre en un sens opérationnel, c'est-à-dire être capable d'exploiter le corpus de données ou de textes pour répondre de façon pertinente à des questions complexes exprimées par un humain dans son langage. Mais en réalité, doter les machines d'une telle capacité est toujours un travail de recherche. Certes, les capacités d'indexation permettant de rassembler de façon quasiment instantanée les éléments d'information contenant virtuellement la réponse à une question sont en place depuis longtemps mais l'exploitation pertinente de ces éléments d'information une fois rassemblés pour donner la bonne réponse est une tâche encore très difficile pour une machine. Antoine Bordes présente quelques-uns des travaux menés sur ces thèmes, dans la ligne générale des réseaux de neurones et de l'apprentissage profond. Il en expose les méthodes : encodage d'éléments symboliques dans des espaces continus, réseaux de mémoire ; il en discute les résultats obtenus : dans le domaine de l'interrogation de bases de connaissances structurées comme dans celui, plus largement ouvert, de l'interrogation sur des bases de textes, comme Wikipédia.

- Un chapitre sous la signature de Luc Steels est intitulé « L'Origine et l'Evolution du langage ». Comment peut-on expliquer l'apparition, l'évolution et la diversité des langages dans les sociétés humaines ? Pour traiter de ces questions Luc Steels défend l'hypothèse d'une analogie entre les mécanismes de l'évolution des langages et ceux de l'évolution biologique : savoir des processus de réplication/transmission, de mutation et de sélection, auxquels s'ajoutent des processus de hiérarchisation en différents niveaux d'organisation. Après avoir montré des exemples d'intervention de tels processus dans les langages, Luc Steels expose les méthodes ressortant de l'Intelligence Artificielle et de la Robotique qui lui permettent de tester son hypothèse à travers diverses expérimentations. Dans

ces expérimentations il montre d'abord comment des Intelligences Artificielles, interagissant les unes avec les autres, font émerger un vocabulaire commun, et un sens commun donné à chaque mot de ce vocabulaire. Et au-delà, toujours à travers de telles interactions, il montre comment peut également émerger – et pour quelles raisons – des structures grammaticales élémentaires telles que des accords de nombre ou de genre, voire même des structures plus complexes comme les emboitements de phrases. En conclusion, il réaffirme l'idée que les langues sont des systèmes culturels changeant de façon permanente, sous l'effet de dynamiques de nature similaire à celles à l'œuvre dans l'évolution des espèces, tout en reconnaissant que cette idée est loin de faire l'unanimité.

Pour le comité de lecture[1]

[1] Eric Chenin, Jacques Printz, Jean-Pierre Treuil

# 7

# Intelligence Artificielle, Des Big Data au Cerveau[1]

**Jean-Gabriel Ganascia**
Sorbonne Université, ACASA
(Agents Cognitifs et Apprentissage
Symbolique Automatique)
Président du Comité d'Ethique du CNRS

## Abstract

Jean-Gabriel Ganascia recalls the roots of Artificial Intelligence, going down into cybernetics and the beginning of cognitive sciences with the "Macy lectures" of the immediate post-war years 1946-1953. He provides a brief history of AI since its "official" birth in 1956; at the same time he gives elements of the rise of "Big Data". He then introduces, in an illustrated way, the adventure of neural networks, since the first articles published in the 1940s. Some of the principles of their classification performance are presented, as well as the so-called "gradient back propagation" algorithms, which constitute one of the bases of deep learning. With, in the background, the emergence of the availability of large masses of data, on a scale never known in the history of the Humanity, large masses of data which deep learning makes it possible to exploit. In conclusion, Jean-Gabriel Ganascia addresses the question of the possibility of "Strong Artificial Intelligence" exceeding the performance of current weak AI, which remains very specialized.

[1] Ce chapitre est la transcription, effectuée par Eric Chenin et Jean Pierre Treuil, membres de l'AEIS, de la conférence de Jean-Gabriel Ganascia faite devant l'AEIS le 12 septembre 2016. Le texte a été relu et corrigé par le conférencier. Il est publié avec son accord.

Pour mieux appréhender les liens entre l'Intelligence Artificielle (IA) et les recherches sur le fonctionnement du cerveau, il est utile de faire un peu d'histoire. A l'origine de l'IA, une question s'est posée : le cerveau étant supposé être le siège de notre intelligence, peut-on, pour construire une « intelligence artificielle », simuler son fonctionnement ? et cette approche est-elle la plus efficace ? Cette problématique a été reprise de façon récurrente au cours des 75 dernières années. Nous allons ici retracer les méandres de cette histoire, pour arriver ensuite au plus contemporain.

Nous parlerons ainsi de qu'on appelle les *masses de données*, en anglais les « *big data* » ; notons ici cette tendance à prendre un mot employé depuis longtemps, puis le traduire en anglais pour faire plus moderne. Il en est de même pour le terme d'apprentissage, terme utilisé en France depuis des lustres dans le domaine de l'IA (cf. les journées françaises d'apprentissage, JFA, se sont tenues régulièrement en France depuis le milieu des années quatre-vingt), puis remplacé par « machine learning » on ne sait très bien pourquoi…

Nous parlerons également des *réseaux de neurones formels.* Ici encore, il s'agit d'approches très anciennes, présentes même avant le début de l'intelligence artificielle, et qui ont fait l'objet de plusieurs renaissances. Ces approches reposent sur une analogie féconde entre les ordinateurs et le fonctionnement du cerveau, en en donnant une image, certes, un peu trop métaphorique et grossière pour être fidèle.

Nous parlerons enfin de ce que nous appelons l'IApocalypse, en écho à toutes les déclarations entendues dans les dernières années, bien que leur contenu soit lui aussi en substance beaucoup plus ancien, déclarations qui annoncent une sorte de « grand soir » de l'Intelligence Artificielle et de l'Humanité.

## 1. Bref aperçu historique : de la cybernétique à AlphaGo

A l'origine de l'Intelligence Artificielle, on peut citer les conférences interdisciplinaires organisées entre 1946 et 1953 par la fondation Macy en vue de fonder une science générale du fonctionnement de l'esprit humain. Elles furent à l'origine du courant cybernétique et des sciences cognitives, d'où émergea la notion d'Intelligence Artificielle. Jean-Pierre Dupuy, dans « Aux origines des sciences cognitives » relate cette aventure intellectuelle passionnante, initiée notamment par deux articles princeps. Celui de David Rosenblueth, Norbert Wiener, et Julian Bigelow : « Machine téléologique », qui traite des liens entre information, commande et rétroaction ; et celui de Warren McCulloch et Walter Pitts, qui introduit les réseaux de neurones formels, lesquels, longtemps délaissés, sont revenus sur le devant de la scène avec les développements fulgurants que l'Intelligence Artificielle connaît actuellement. A ces articles, il faut ajouter celui d'Alan Turing, paru dans la revue Mind en octobre 1950[2], qui traite de ce que signifie penser pour une machine, et décrit le

[2] Intitulé « *Computing Machinery and Intelligence* ». Cet article fait suite à une première version parue en 1947. Voir figure 1

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

**Figure 1** L'article de Turing de 1950, où il décrit son fameux test

fameux Test de Turing ; savoir le jeu dit de l'imitation, où un interrogateur essaie de distinguer, en échangeant uniquement des questions-réponses écrites, entre d'un côté, une femme, et de l'autre, un ordinateur imitant un homme qui essaie de se faire passer pour une femme : Turing prédit qu'avant 2000, un ordinateur sera capable de tromper l'interrogateur durant cinq minutes dans au moins 30% des cas.

Mais on considère communément que la naissance de l'expression « Intelligence Artificielle » date de l'atelier d'été co-organisé en 1956 au Dartmouth College par John Mac Carthy, Marvin Minsky, Nathaniel Rochester et Claude Shannon[3]. L'idée de cet atelier (figure 2) reposait sur la conjecture selon laquelle tous les aspects de l'apprentissage, et toutes les autres caractéristiques de l'intelligence, peuvent en principe être décrits de façon si précise qu'une machine pourrait être conçue pour les simuler.

Contrairement à ce que l'expression pouvait suggérer, le nouveau champ disciplinaire envisagé n'avait pas pour ambition de construire UNE intelligence artificielle, mais d'étudier l'intelligence, sous ses différentes composantes, par analogie avec les machines. Ces composantes sont les fonctions cognitives classiquement identifiées : les fonctions réceptives, la mémoire et l'apprentissage, le raisonnement et la pensée, les fonctions interprétatives, et les fonctions exécutives. Toutes ces fonctions peuvent faire l'objet de simulations avec les techniques d'intelligence artificielle, et c'est ce qui a été fait, durant les soixante dernières années.

[3] A cette époque, John Mac Carthy et Marvin Minsky, tous deux de formation mathématique, avaient moins de trente ans ; malgré leur jeunesse, ils avaient réussi à convaincre de leur projet Nathaniel Rochester, alors Directeur scientifique d'IBM, et Claude Shannon.

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

**Figure 2** Eté 1956, Darmouth College, naissance de l'Intelligence Artificielle

Et de fait, si l'expression a pu susciter de l'incrédulité, et si sur la période ont alterné des phases d'emballement et des moments où les progrès marquaient le pas, il faut reconnaître que sur les soixante dernières années, il n'y a pas une discipline scientifique qui a autant transformé le monde que l'intelligence artificielle. Et je vais en donner quelques exemples. A commencer par le Web tout simplement, dont la force réside dans le couplage de l'hypertexte et du réseau Internet. Le concept d'hypertexte a été inventé par un philosophe, Ted Nelson, qui l'a implémenté en LISP comme une nouvelle manière d'organiser la mémoire, et qui l'a publié dans une conférence d'intelligence artificielle[4]. Et Tim Berners Lee, au CERN, a associé l'hypertexte à l'Internet pour créer le Web.

On peut multiplier les exemples, avec la voiture autonome, les robots, la biométrie, l'apprentissage, la vision par ordinateur, la reconnaissance de la parole et la compréhension du langage naturel … Sur cet exemple du traitement de la parole et du langage naturel, on peut mesurer l'accélération des progrès, depuis *2001 l'Odyssée de l'espace*, où cela relevait encore de la science-fiction, avec une recherche balbutiante en laboratoire, jusqu'à des applications comme SIRI, qui sont maintenant courantes sur les téléphones portables, au sous-titrage en temps réel des vidéos sur Internet, ou à la traduction automatique. Au niveau des jeux aussi, il y a eu des progrès considérables, depuis la victoire de Deep Blue sur Gary Kasparov, aux échecs, en 1997, jusqu'à celle d'AlphaGo sur Lee Sedol, au jeu de

[4] T.H. Nelson : *A file structure for the complex, the changing and the indeterminate* ; Complex Information Processing, Proceedings of the 1965 20th ACM National Conference, August 1965

go, en 2016 (figure 3), en passant par la machine Watson, qui a gagné en 2011 contre les meilleurs au jeu de Jeopardy.

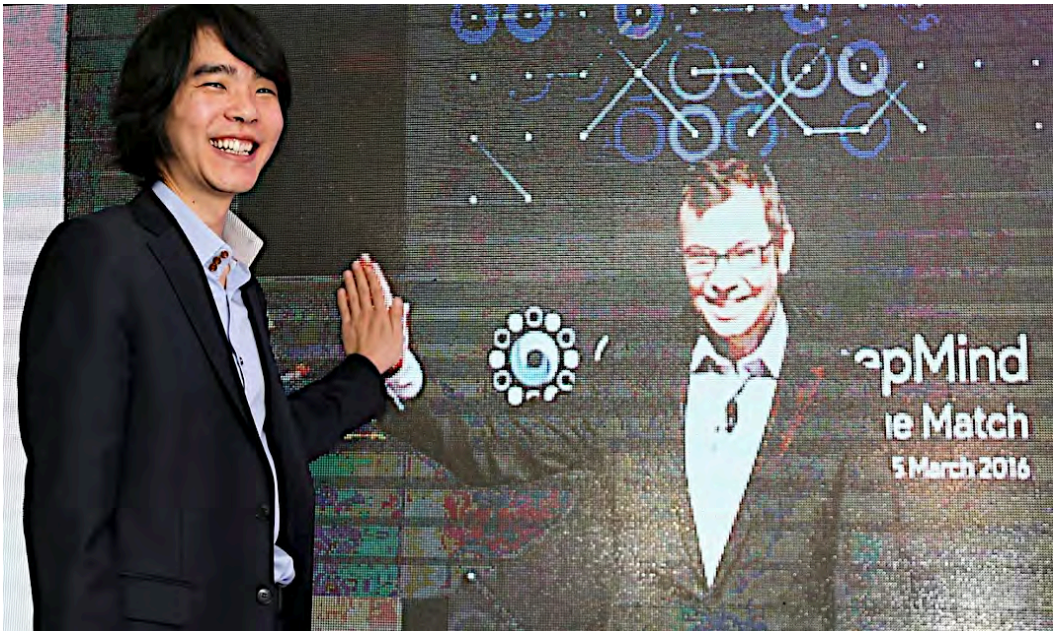



**Figure 3** Lee Sedol, un des meilleurs joueurs mondiaux de go et Demis Hassabis, père d'AlphaGo

## 2. La montée en puissance des « Big data »

Ces progrès fonctionnels sont accompagnés d'une croissance impressionnante des volumes de données traités, ce que l'on désigne en anglais par « big data ». On peut se faire une idée de ces volumes en se référant par exemple à celui de l'information contenue dans la totalité des ouvrages que possède la Bibliothèque Nationale de France (BNF). Celle-ci réunit quatorze millions d'ouvrages dans son catalogue des livres et imprimés : si l'on considère qu'un ouvrage compte en moyenne un million de caractères, cela fait une volume de données de $14 \times 10^{12}$ octets, soit 14 tera-octets. En comparaison, le volume total du Web était, en 2015, d'environ 7 zeta-octets, $7 \times 10^{21}$, soit un demi-milliard de BNF. Et Twitter génère à lui seul 7 téraoctets par jour, c'est-à-dire une demi BNF par jour. Quant à Whatsapp, 900 Millions de personnes l'utilisent. Et FaceBook, 1,5 Milliard, avec 9 Milliards de photos partagées par jour.

On attache à la notion de *big data* trois caractéristiques : le Volume bien sûr, la Variété (textes, images, sons, etc.), et la Vitesse (la rapidité d'évolution au cours du temps), caractéristiques que l'on désigne couramment par « les 3 V ». Parmi ces trois caractéristiques, la vitesse joue un rôle clé. L'information est en effet souvent acquise en

continu : c'est un facteur essentiel dans l'économie contemporaine, et c'est un aspect déterminant dans la logique du traitement de l'information.

Une source majeure pour cette acquisition en continu est le « crowd sourcing », lequel est notamment orchestré au moyen d'outils grand public comme les moteurs de recherche, les réseaux sociaux ou AMT (*Amazon Mechanical Turk*). Ici la source d'information est chacun d'entre nous : il en va ainsi des requêtes que nous soumettons aux moteurs de recherche ; ou bien des données et informations diverses - y compris multimédia : images, sons, vidéos - que nous déposons régulièrement sur les réseaux sociaux ; ou encore des informations que nous fournissons à divers logiciels dont la qualité de service dépend de l'ensemble des informations fournies collectivement : le logiciel de guidage Waze en est un parfait exemple, qui évalue le trafic en temps réel à partir des vitesses des véhicules de l'ensemble des utilisateurs connectés.

Un autre aspect de ces exemples de *crowd sourcing* est que l'information recueillie couvre des aspects très divers et concerne une part très importante de la population, à laquelle on a accès sans avoir besoin de sélectionner un échantillon, comme on le fait dans les sondages traditionnels. Bien sûr, sans les récents progrès de l'intelligence artificielle, on ne serait pas capable de traiter une telle diversité et une telle quantité de données.

Cette approche paraît puissante et prometteuse. Elle pourrait permettre de détecter des signaux faibles, voire indirects, par corrélation d'éléments divers. Selon Google, l'analyse des requêtes sur un moteur de recherche peut permettre, par exemple, de détecter une épidémie. Mais il ne faut peut-être pas accorder à cette approche une confiance excessive : par exemple, lorsque Edward Snowden dénonce l'aspiration de toutes les télécommunications par les États-Unis, rien ne garantit pour autant que des signaux très faibles leur suffisent pour détecter la préparation d'un attentat.

L'Intelligence Artificielle revêt aujourd'hui une importance d'autant plus grande que toute la vie sociale est maintenant numérisée. Les moteurs de recherche, les réseaux sociaux, les robots, les véhicules autonomes, les agents conversationnels, les objets connectés : tout cela génère d'un côté une énorme quantité de données d'une grande diversité, et requiert d'un autre côté des capacités nouvelles de traitement. Ce sont là deux défis imbriqués que seuls les progrès les plus récents de l'IA permettent de relever.

Si l'on considère le cas de la voiture autonome, les nouvelles capacités de traitement relèvent de la perception, et notamment de la reconnaissance d'images : identifier les voitures, les passants, les panneaux de signalisation, la route et les trottoirs, etc (figure 4). Mais aussi de la décision, y compris pour éviter une collision avec un passant ou un autre véhicule, avec parfois des conflits à trancher.

Dans le cas des agents conversationnels, on s'approche du test de Turing, où l'agent répond comme un humain, et parfois apprend en temps réel, comme l'agent *Tay* de Microsoft.

Celui-ci a fait scandale, parce qu'il apprenait au contact des internautes, dont certains lui ont tenu des propos orduriers, et *Tay*, par imitation, a tenu des propos similaires, de nature raciste. On a pu constater, sur cet exemple, certaines limites des agents apprenants.

**Figure 4** Véhicules autonomes : En bas, la vue « naturelle » et en haut, l'analyse de la même scène par le système de vision du véhicule autonome (Google Car)

Sur ces deux exemples d'application très différents, on observe que se posent des questions de l'ordre de l'éthique « computationnelle ». Il s'agit de voir comment on peut introduire un certain nombre de *valeurs* dans ces agents autonomes. C'est une question tout à fait passionnante pour un chercheur, dont on perçoit l'étendue et l'acuité.

## 3. Réseaux de Neurones formels

Revenons maintenant aux débuts des réseaux de neurones formels et à la naissance de l'Intelligence Artificielle.

Bien avant l'atelier de l'été 56 au Dartmouth College, dès les années 40, les acquis de la neurobiologie ont inspiré les pères des réseaux de neurones artificiels. Les travaux de Golgi, par exemple, avaient permis, à la fin du XIX$^{e}$ siècle, de visualiser la structure du tissu cérébral[5]. Les travaux de Santiago Ramon Y Cajal avaient ensuite identifié l'axone et les dendrites, grâce auxquels les neurones s'interconnectent via les synapses : chaque neurone envoie des informations par son axone et en reçoit par ses dendrites, la connexion entre axones et dendrites étant réalisée par les synapses (figure 5).

[5] Camillo Golgi, médecin et histologiste italien ; a partagé en 1906, avec Santiago Ramon y Cajal, le prix Nobel de Médecine, « en reconnaissance de leur travaux sur la structure du système nerveux »

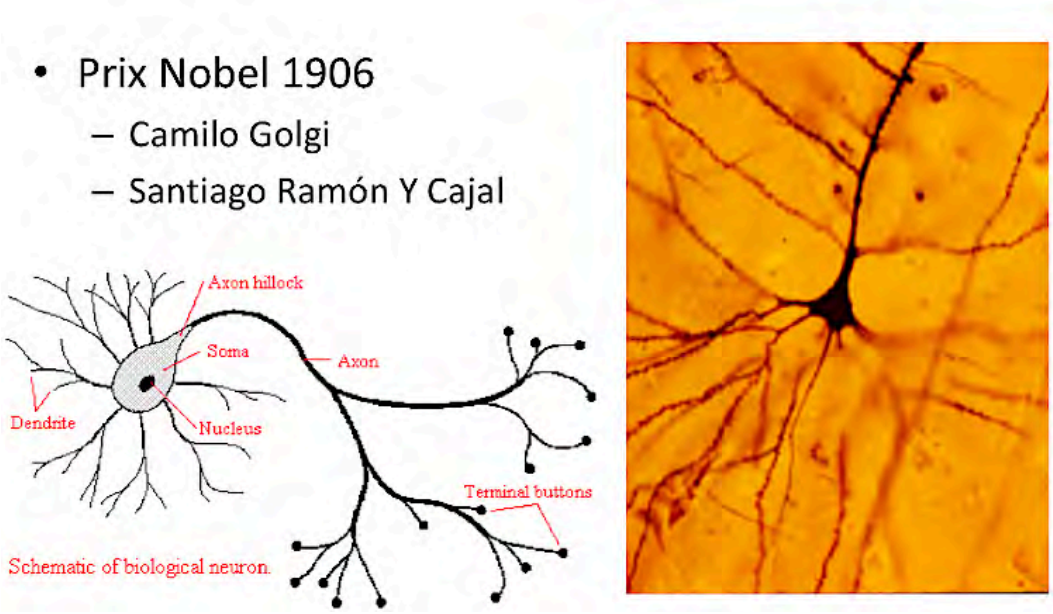



**Figure 5** 1906, découverte des neurones biologiques.

### 3.1 Une histoire mouvementée

Ces travaux en neurosciences font apparaître la structure des réseaux des neurones, et dans leur article de 1943[6], Warren McCulloch et Walter Pitts imaginent de simuler cette structure avec des automates. A cette époque où les ordinateurs électroniques n'existaient pas encore, les relais téléphoniques étaient utilisés pour faire du calcul « électrique », et McCulloch et Pitts recourent à ces relais pour réaliser leurs automates. Ils montrent qu'avec une structure de réseau à trois couches, à condition d'avoir suffisamment de neurones dans la couche centrale, d'une part, et d'autre part d'attribuer à chaque connexion la bonne pondération, on peut réaliser n'importe quelle fonction logique booléenne. On peut d'ailleurs noter que cette propriété d'universalité a été généralisée, bien plus tard, à toutes les fonctions continues bornées dans [0,1]

Dès les années 50 et les débuts de l'électronique, on se pose des questions ambitieuses touchant à l'utilisation du langage par l'ordinateur, à la théorie de la complexité, à l'auto-amélioration, à l'abstraction, à l'aléatoire et à la créativité. Et de fait actuellement, ces questions commencent à recevoir des réponses concrètes. L'apprentissage profond révolutionne l'IA et nos outils numériques, et le projet Blue Brain, en Suisse, essaie de reproduire complètement un cerveau (objectif en soi illusoire, parce qu'on ne peut pas

[6] Warren McCulloch and Walter Pitts, *A logical Calculus of Ideas Immanent in Nervous Activity ;* Bulletin of Mathematical Biophysics 5, 1943. Pitts avait alors 20 ans

simuler un assez grand nombre de neurones ni de connexions ; mais ce projet permet de mieux comprendre certains aspects des neurosciences).

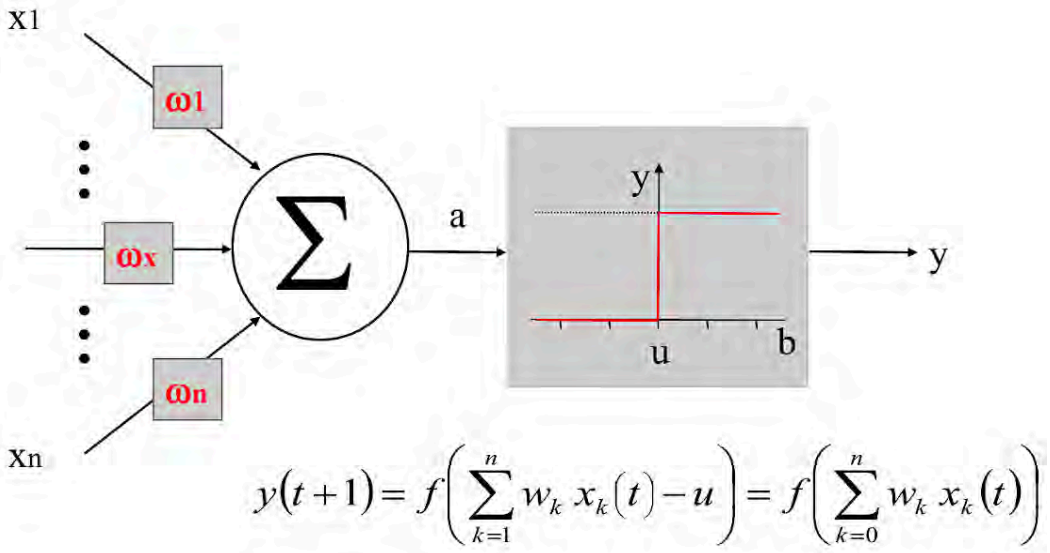



**Figure 6** Schéma d'un neurone formel : le vecteur d'entrée, les pondérations, la fonction d'activation et la sortie

*3.1.1 L'aventure des perceptrons*

Mais les réseaux de neurones formels, entre les années 40 et aujourd'hui, ont connu des hauts et des bas. Jusqu'aux années 90, les capacités de calcul limitent ce que l'on peut faire en pratique. Après les tous premiers essais de Warren McCulloch et Walter Pitts dans les années 40, à l'échelle de quelques neurones, les premières implémentations opérationnelles, utilisées dans les laboratoires, vont apparaitre dans les années 50. Marvin Minsky y travaille[7], en essayant de construire une machine qui apprend en modifiant les poids synaptiques et se rend compte qu'il s'agit d'un problème difficile. Il obtient son PhD à l'Université de Princeton en 1954 avec sa thèse intitulée *Theory of Neural-Analog Reinforcement Systems and its Applications to the Brain Model Problem*. Devant les difficultés rencontrées, inévitables compte tenu des machines de l'époque, Franck Rosenblatt[8] opte pour des implémentations limitées à deux couches ; partant de la métaphore de la rétine qui les inspire, ces premières implémentations, prennent le nom de « perceptron »[9]. Arrêtons-nous un instant sur ses travaux.

[7] Il faut voir les conditions effectives, dans ces années là, des approches « hardware » des réseaux neuronaux, conditions dans lesquelles le maniement du fer à souder s'avérait indispensable.

[8] Frank Rosenblatt et Marvin Minsky s'étaient connus pendant leurs années d'étude, dans les années 1945-1946

[9] Cf. Frank Rosenblatt : *The perceptron, A perceiving and Rognizing Automaton* ; Project Para Report N° 85-460-1, Cornell Aeronautical Laboratory (CAL), Jan. 1957 ; mais aussi, le lieu de la

L'apprentissage réalisé par Frank Rosenblatt, dans ses expériences de 1957-1958 (Mark I Perceptron) s'effectue en réalité sur les poids *d'une seule* couche de neurones, les poids $w_{i,k}$ affectés aux inputs *(k= 1,2, … p)* des *n* neurones *i= 1,2, … n ;* mais le vecteur [$x_1$, $x_2$, … $x_p$] des inputs est lui-même sortie d'une couche d'association, cette fois ci *câblée* - donc fixe dans chaque expérience - depuis la grille d'unités photo-sensitives, grille constituant un analogue simplifié d'une rétine, la vraie entrée du système[10]. Chacun des *n* neurones possède une seule sortie, correspondant à l'une des catégories d'objets que le système cherche à classifier : l'apprentissage est accompli lorsque seule s'active la sortie liée à la catégorie effective de l'objet présenté sur la rétine.

Pendant quelques années, les travaux autour du *perceptron* donnent de bons résultats : on développe en particulier des procédures d'apprentissage automatique efficaces. Mais les limites de cette architecture apparaissent vite, et Marvin Minsky lui-même, à la fin des années 60, dans « *Perceptrons: an introduction to computational geometry* »[11], montre que certaines fonctions ne peuvent pas être réalisées. Il montre notamment que les perceptrons, qui n'utilisent que des combinaisons linéaires, ne peuvent représenter qu'un ensemble limité de fonctions : celui des fonctions « linéairement séparables », c'est-à-dire, dont l'application revient à couper l'espace des possibles en deux régions séparées par un hyper-plan. Or il se trouve même des fonctions élémentaires, par exemple dans le domaine des fonctions logiques, qui ne peuvent pas être réduites à une séparation par un hyper-plan : c'est notamment le cas de la fonction « OU exclusif ».

*3.1.2 Un intermède utile, le développement des théories mathématiques de l'apprentissage*
Cette limite théorique condamne les réseaux de neurones formels, pendant une quinzaine d'années, à être délaissés. Comme on le verra plus loin, il fallut attendre 1986 pour que l'on généralise la procédure d'apprentissage du Perceptron à des réseaux de neurones à trois couches ou plus. Ceci étant, compte tenu des performances des ordinateurs, il était très difficile d'appliquer cette généralisation des réseaux de neurones formels à des problèmes de grande taille, avec beaucoup de données. C'est la raison pour laquelle, à partir du milieu des années quatre-vingt-dix, les approches statistiques de l'apprentissage prennent le dessus, avec notamment les machines à vecteur de support[12], et les machines à noyau, algorithmes qui sont encore d'usage courant aujourd'hui. Rappelons-en quelques bases : l'objectif est de déterminer automatiquement si un objet *x*, pris dans un ensemble

publication est intéressant à souligner : The perceptron, A Probalistic Model for Visual Perception ; *in Procs. Of the 15 International Congress of Psychology*, 1957 ; et plusieurs autres publications qui vont suivre en 1958, sous l'étiquette du CAL

[10] Il y a bien ainsi deux couches de neurones : une première couche, la couche d'association – 512 unités dans Mark-I Perceptron – ayant comme entrée la grille photo-sensitive, et la couche de réponse – 8 unités – dont les liens avec la première couche sont les poids $w_{i,k}$, les seuls modifiés par l'algorithme d'apprentissage

[11] Seymour Papert & Marin Minsky : *Perceptrons : An Introduction to Computational Geometry*. MIT Press, 1969, 268 p.

[12] Dites encore SVM, séparateur à vaste marge.

d'objets possibles *X*, peut être classé, ou non comme relevant d'une catégorie C donnée. On retrouve bien là le problème de *discrimination* auquel s'attaquait le perceptron. Pour ce faire, ces algorithmes opèrent sur un ensemble d'apprentissage (ou échantillon) constitué d'objets *x*, les uns relevant de la catégorie C (exemples) et d'autres non (contre exemples) ; chacun de ces objets est décrit par un *vecteur* d'un espace $X^p$ d'une certaine dimension *p* - un espace doté d'un produit scalaire et donc d'une distance - si bien que l'échantillon se concrétise par un ensemble de ces vecteurs $[x_1, x_2, \ldots x_n]$ ; il s'agit alors sur cet échantillon de trouver une fonction $f(x)$ prenant une valeur positive sur les vecteurs descriptifs des objets de la catégorie C et négative pour les autres objets. Bien entendu, cette fonction *f* doit satisfaire certaines propriétés assurant la qualité de la discrimination, l'assurance que les objets n'appartenant pas à l'ensemble d'apprentissage seront bien classés dans une grande majorité de cas ; et c'est là que les théories mathématiques de l'apprentissage entrent en jeu[13].

Dans les cas où, dans l'espace $X^p$ les vecteurs relevant de la catégorie C peuvent être séparés des autres par un simple hyperplan, le problème de discrimination est dit *linéairement séparable.* Il s'agit alors de trouver le « meilleur » hyperplan de séparation, c'est-à-dire la « meilleure » fonction discriminante de la forme $f(x) = v.x + a$, $v.x$ étant le produit scalaire d'un vecteur orthogonal à l'hyperplan et donc caractérisant son orientation dans l'espace, et *a* un scalaire caractérisant sa position. C'est dans la définition de ce qu'est ce « meilleur » qu'interviennent les théories, en définissant différentes *marges de confiance* fournissant – par maximisation - autant de critère de choix pour les couples de valeurs $(v,a)$.

Il faut bien sûr généraliser l'approche, car comme Minsky et Papert l'ont bien rappelé, tous les problèmes de discrimination ne sont pas linéairement séparables. C'est ici qu'apparait l'utilisation des *noyaux.* D'une manière très simplifiée, cette utilisation revient à trouver une « déformation » de l'espace $X^p$, c'est-à-dire une transformation $\Phi$ des descriptions $[x_1, x_2, \ldots x_n]$ en $[\Phi(x_1), \Phi(x_2), \ldots \Phi(x_n)]$, de façon à se retrouver dans le cas précédent d'une discrimination linéaire. D'une façon plus précise, un noyau est une fonction $k(x_i, x_j)$ évaluant une « séparation » entre deux vecteurs $x_i$ *et* $x_j$ . Le problème à résoudre est alors de déterminer, sur l'exemple d'apprentissage, la meilleure fonction discriminante[14] de la forme $f(x) = \Sigma_i \alpha_i y_i k(x_i, x) + a$, les $y_i$ étant les étiquettes (1 ou -1) affectées aux vecteurs selon qu'ils représentent ou non des objets de la catégorie C ; ici encore la recherche de

[13] Entrer dans ces théories dépasserait le cadre de cet article, mais il est utile de mentionner ici quelques références et notions : la théorie statistique de l'apprentissage de l'informaticien d'origine russe Vladimir Vapnik (autour des années 1990) mobilisant les notions de *trace* d'une famille *F* de sous-ensembles sur un sous-ensemble donné *C* , de *pulvérisation* de *C* par *F*, enfin la dimension de Vapnik-Chervonenkis de cette famille *F*. Ou encore la théorie de l'apprentissage PAC (Probably Approximately Correct) proposée par l'informaticien britannique Leslie Valiant en 1984

[14] Evaluant en quelque sorte la force à laquelle un vecteur *x* donné « se sépare » de l'ensemble des vecteurs de l'ensemble d'apprentissage, en tenant compte de leurs étiquettes respectives.

cette meilleure fonction procède de la maximisation d'une certaine marge de confiance, construite sur les valeurs des $\alpha_i$ et *a.*

Ces approches mathématiques et statistiques, au-delà des applications qu'elles ont permis de mettre au point, allaient se révéler très utiles pour comprendre l'efficacité des réseaux neuronaux.

*3.1.3 La résurgence des réseaux de neurones*

Dans la mi-temps des années 1980, quelques chercheurs, dont David Rumelhart et Yann Lecun[15], reprennent l'idée des réseaux de neurones formels, mais cette fois avec trois couches et plus. Ils introduisent une nouvelle méthode d'apprentissage, qu'ils appellent la rétro-propagation de gradient, et ajoutent en sortie de chaque neurone une fonction non linéaire, intitulée fonction d'activation. La fonction d'activation permet de dépasser le comportement linéaire trop limité du réseau. Et la rétro-propagation, inspirée de méthodes issues de la physique statistique, est utilisée lors de l'apprentissage, pour calculer, en remontant les couches de proche en proche depuis la sortie jusqu'à l'entrée du réseau, la correction qui doit être apportée dans chaque neurone à ses facteurs de pondération pour minimiser l'écart entre la sortie obtenue et le résultat attendu.

Donnons-en quelques détails. L'idée de base est d'évaluer, pour chaque exemple (vecteur des entrées) et pour une configuration donnée des poids du réseau, les sensibilités – à la valeur de ces différents poids - de l'écart *E* entre le vecteur des sorties attendues et le vecteur des sorties calculées par le réseau ; cet écart *E* peut être la moyenne quadratique des écarts sur chaque sortie élémentaire, encore appelés *erreurs*. Les sensibilités s'expriment par les dérivées partielles $\frac{\partial E}{\partial W(i,j,l)}$ (les coefficients de *gradient*) où chaque *W(i,j,l)* représente le poids « actuel » de la connexion entre le neurone *i* de la couche *l-1* et le neurone *j* de la couche *l.* On pourrait penser à priori que leur calcul est difficile, étant donné la complexité du réseau, l'éventuel grand nombre de couches, etc. La solution est, *sans changer les poids*, de *rétro-propager les erreurs* d'une couche à celle qui la précède, à l'aide d'une formule, déduite d'une démonstration mathématique, qui s'avère assez intuitive ; formule de plus très « efficace », car les sensibilités s'expriment alors de façon particulièrement simple en fonction de ces erreurs ainsi rétro-propagées sur chaque couche. Ces sensibilités une fois obtenues, il ne reste plus ensuite 1) qu'à modifier « légèrement » les valeurs des poids proportionnellement à l'opposé de la sensibilité de *E* à leur égard – si

[15] Il faut sans doute citer ici également Paul John Werbos et sa « dissertation » de 1974 sur la rétro-propagation des erreurs dans un réseau neuronal, et quelques autres pionniers tels que David B. Parker, avec son article de 1985 au titre évocateur : *Learning Logic : Casting the Cortex of the Human Brain in Silicon,* Geoffrey Hinton et Ronald Williams, co-auteurs avec D. Rumelhart de leur article de 1986 *Learning representations by back propagating errors.* Quant à Yan Le Cun, il publie plusieurs articles sur le sujet entre 1985 et 1988, notamment en lien avec l'équipe de Françoise Fogelman-Soulié à l'Université Pierre et Marie Curie ; il publie sa thèse en 1987, dans cette université, sous le titre *Modèles Connexionnistes de l'apprentissage.*

par exemple $E$ augmente avec un poids, on va évidemment diminuer ce poids, et inversement, 2) de recalculer avec ces poids corrigés un nouvel écart $E$ en principe plus faible, et 3) itérer le processus jusqu'à obtention d'un écart suffisamment faible pour être considéré comme satisfaisant. Bien entendu, de nombreuses variantes d'un tel algorithme ont été proposées et testées.

Après quelques débuts probants, les progrès sont lents, jusqu'à ce que deux facteurs essentiels, au début des années 2010, permettent aux réseaux de neurones de produire des résultats étonnants. Ces deux facteurs sont d'une part la croissance exponentielle de la capacité de calcul des ordinateurs - encore accrue par l'utilisation depuis quelques années des cartes graphiques, construites pour gérer l'affichage sur les écrans de nos ordinateurs, et dont la capacité à effectuer des opérations simples et répétitives à très grande vitesse est parfaitement adaptée aux réseaux de neurones - et d'autre part les énormes quantités de données qui deviennent disponibles pour entraîner les réseaux.

### 3.2 Techniques et performances actuelles

On utilise communément plus d'une dizaine de couches[16], réunissant des millions de neurones, que l'on entraîne à l'aide de millions d'exemples. Le rôle que jouent les différentes couches est conditionné par la manière dont elles sont connectées entre elles.

Les premières couches (cf figure 7) sont souvent connectées de manière à effectuer un ensemble de convolutions qui visent à détecter dans le signal d'entrée, sur une fenêtre glissante de taille variable, une série de motifs élémentaires, puis à composer ces motifs pour former des structures plus complexes extraites du signal d'entrée. Ces couches d'entrée sont généralement entraînées sur des jeux de données standard indépendants du sujet spécifique traité par le réseau ; ce sont les couches suivantes, qui sont alimentées avec les structures détectées par les couches d'entrée, que l'on entraîne avec des données spécifiques au problème que l'on veut traiter, jusqu'à la dernière couche, dont on compare la sortie avec le résultat attendu.

[16] Le terme de profond – deep - dans apprentissage profond - deep learning - fait depuis 2000 référence à la profondeur du réseau, le nombre de couches entre la couche d'entrée et la couche de sortie. Mais cette expression *Deep Learning* semble avoir été utilisée pour la première fois en 1986, introduite par l'informaticien américain Rina Dechter. Les années qui ont suivi ont vu une lente maturation de l'architecture des réseaux et des algorithmes d'apprentissage. A partir des années 2000, des applications industrielles du Deep Learning ont commencé à fonctionner, pour se développer sur une grande échelle après 2010.

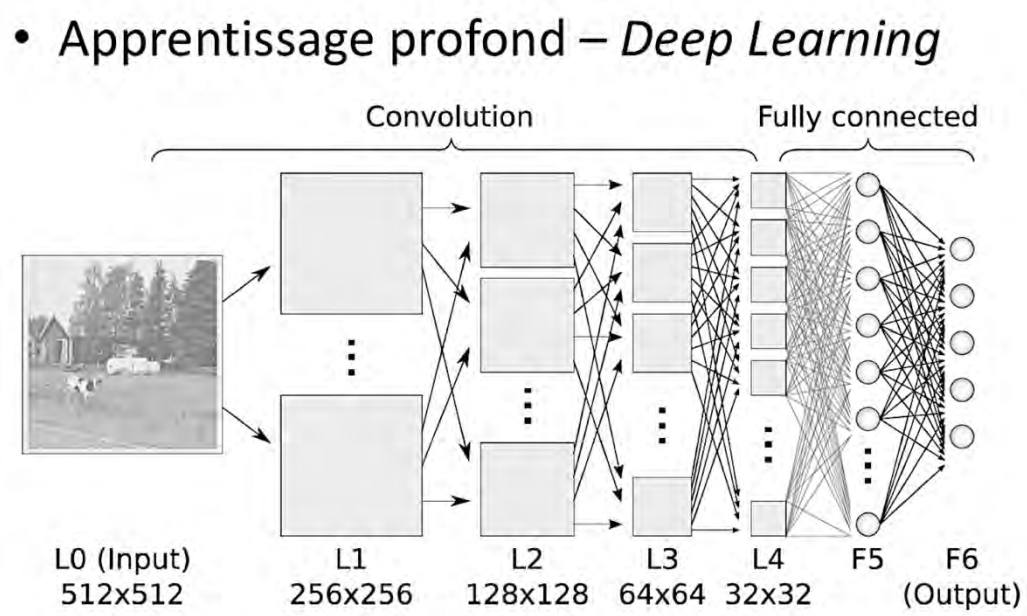



**Figure 7** Schéma d'un réseau de neurones pour le *Deep Learning*, l'Apprentissage profond

On fait ainsi de la reconnaissance d'images, de sons, de vidéos, de la recherche d'informations. La voiture autonome de Google est un exemple d'application de ces techniques. Les architectures se diversifient, ainsi que les méthodes d'apprentissage. On fait de l'apprentissage supervisé, non supervisé, ou par renforcement comme dans le projet Deep Mind d'Alpha Go. On construit des réseaux encodeurs-décodeurs qui apprennent à condenser l'information puis à la restituer, et on parvient même *à créer* à l'aide de réseaux génératifs « antagonistes », dérivés des encodeurs-décodeurs.

Mais c'est l'apprentissage supervisé qui est encore le mieux maîtrisé, avec cependant des limites pratiques. Il fonctionne en effet d'autant mieux que l'on dispose de grandes quantités de données pour entraîner le réseau. Le logiciel « Deep Face », de FaceBook, par exemple, est entraîné avec 4,4 Millions d'images pour 4030 personnes, soit plus de 1000 images par individu.

Et « Face Net », de Google, apprend sur 200 millions d'images pour 8 millions d'identités uniques, soit une vingtaine ou une trentaine d'images par personne, avec un taux de réussite de 99,63 %.

Dans ces contextes, où l'on dispose de beaucoup d'images pour l'entraînement, et où il s'agit d'identifier des visages isolés, la reconnaissance faciale donne de très bons résultats.

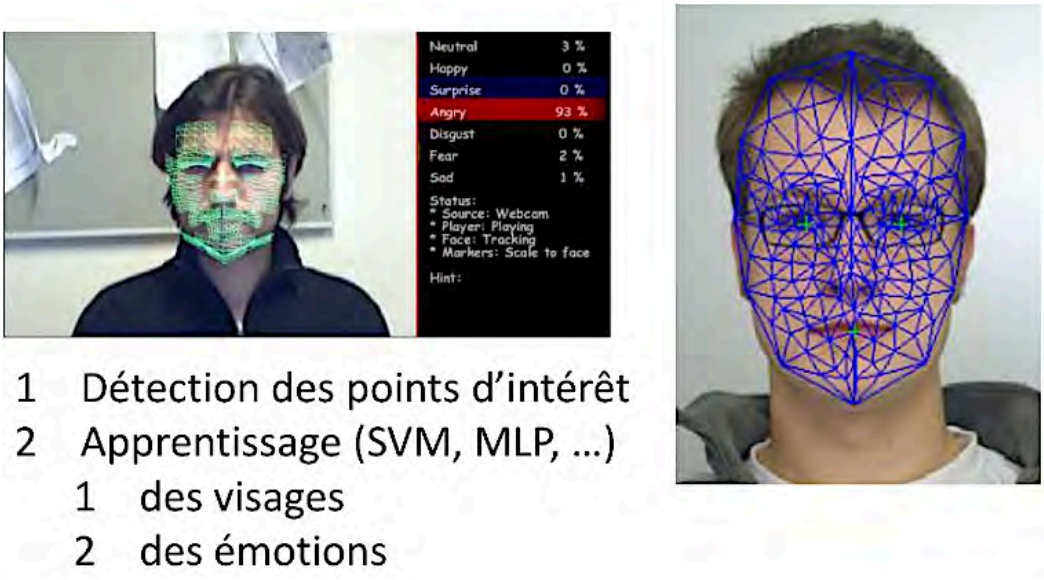



**Figure 8** Reconnaissance des visages, détection des émotions.

**Figure 9** Exemples d'images fournies pour la reconnaissance de trois identités

En revanche, dans un contexte de sécurité publique, où il s'agirait d'identifier des personnes suspectes au sein d'une foule, les conditions sont beaucoup moins favorables. La mairie de Nice avait par exemple proposé des portiques pour assurer la sécurité lors de l'Euro 2016 à l'aide de reconnaissance faciale ; mais cela semble impossible à mettre en pratique : d'une part les visages sont plus difficiles à reconnaître, quand ils sont noyés dans une foule, avec des attitudes et des expressions très variables, voire parfois grimés comme

certains supporters ; et d'autre part la mairie n'a tout simplement pas accès aux images des personnes fichées S pour entraîner le système. Nous avons là une situation paradoxale, où la puissance publique, pour des raisons éthiques, se trouve moins bien placée que des sociétés privées pour rendre des services d'identification de personnes, services qui relèvent pourtant de la sécurité publique.

Maintenant je voudrais aborder un autre aspect : celui de l'interface entre cerveau et ordinateur. On sait mesurer les ondes électromagnétiques engendrées par l'activité cérébrale, y compris l'activité des zones qui interviennent dans la préparation des gestes. On a ainsi accès à la signature électromagnétique de l'impulsion d'un geste, comme celui consistant à déplacer la souris de l'ordinateur vers la gauche. Et on peut entraîner un réseau de neurones à reconnaître ces signatures, et utiliser cette reconnaissance, par exemple pour déclencher la commande correspondante dans une prothèse.

## 4. Et pour demain ?

Les performances impressionnantes des réseaux de neurones, couplées aux possibilités qu'ouvre ce type d'interfaçage de notre cerveau avec l'ordinateur, nous conduisent à nous interroger sur les conséquences possibles de ces progrès techniques sur notre société humaine. Certains se demandent si l'ordinateur ne risque pas de dépasser l'homme : ainsi Stephen Hawking, avec plusieurs autres grands scientifiques, a signé une tribune dans ce sens, publiée dans « The Telegraph » du 1er Mai 2014. Dans cet esprit a notamment été reprise la notion mathématique de singularité, au sens mathématique de point critique d'une fonction. Appliquée ici aux capacités de la machine, et en imaginant que la loi de Moore se poursuive indéfiniment, on entrerait dans une spirale de croissance exponentielle des performances des machines, et donc de leur intelligence, ce qui reviendrait à une explosion de l'intelligence, où l'homme serait très vite totalement dépassé. De plus, on peut craindre que si l'homme a été capable de créer une machine consciente et plus intelligente que lui, celle-ci à son tour saura créer une nouvelle génération de machine plus intelligente qu'elle, etc. Certains parlent de « point de non-retour », dans l'évolution actuelle des machines, au-delà duquel l'homme perdra définitivement le contrôle.

Mais on peut ici considérer la distinction que fait John Searle entre intelligence artificielle faible, et intelligence artificielle forte. La première est notamment celle qu'implémentent actuellement les réseaux de neurones : elle est très performante pour des tâches spécialisées, y compris dans des jeux de stratégie comme le Go, mais il lui manque ce que l'on peut appeler le bon sens et la conscience. Tandis que l'intelligence forte, qui intègre bon sens et conscience, dépasse la force calculatoire brute, et repose sur des concepts qui nous échappent encore. John Searle tient même que cette version forte de l'intelligence, la seule qui rendrait la machine potentiellement dangereuse pour l'humanité, est réservée aux êtres vivants, seuls capables d'être doués de conscience.

# 8

# Former les machines à la compréhension du langage naturel[1]

**Antoine Bordes**
Directeur
Laboratoire
Facebook Artificial Intelligence Research
Paris

## Abstract

The performance achieved in the exploitation of relational databases through query languages and the ease with which information is accessed to, using keywords on search engines could suggest that we are close to a situation in which the algorithms "will understand" the "natural" language. Understanding in an operational sense, that is, being able to use this understanding to respond in a relevant way to complex questions expressed by a human in his language. But in reality, equipping machines with such a capacity is still a research task. Admittedly, the indexing capacities making it possible to gather in an almost instantaneous way the pieces of information containing virtually the answer to a question have been in place for a long time but the relevant exploitation of these pieces of information once gathered to give the correct answer is still a very difficult task for a machine. Antoine Bordes presents some of the work carried out on these themes, in the general line of neural networks and deep learning. The methods: encoding symbolic elements in continuous spaces, of which Antoine Bordes explains the principles, memory networks. The obtained results: in the field of interrogation of structured knowledge bases as in the field, more widely open, of the interrogation on text bases, like Wikipedia.

[1] Ce chapitre est la transcription, effectuée par Jacques Printz et Jean Pierre Treuil, membres de l'AEIS, de la conférence d'Antoine Bordes faite au colloque organisé par l'AEIS, à l'Institut Henri Poincaré, le 16 mars 2018 ; le texte a été relu et corrigé par le conférencier. Il est publié avec son accord.

# 1. Introduction

L'histoire de l'informatique et des ordinateurs est jalonnée par les progrès effectués dans la conception des langages – langages de programmation, de requêtes sur les bases de données – permettant de s'adresser à ces machines pour les utiliser ; mais le moyen idéal pour interagir avec elles serait bien de pouvoir leur parler comme nous le faisons entre nous, et donc de faire en sorte qu'elles puissent apprendre et comprendre le langage humain.

Un tel idéal parait atteignable relativement facilement, nombre de technologies déjà disponibles peuvent en effet le laisser penser. Les requêtes sur Google par exemple sont faites à l'aide d'un langage à base de mots clés que tout le monde s'auto-apprend à utiliser, ce langage peut être vu comme une espèce d'intermédiaire proche du langage humain ; autres exemples, les succès obtenus en reconnaissance de la parole, d'usage courant dans les smartphones, la transcription de la parole en texte, l'assistance à la conduite dans les véhicules récents, les agents conversationnels ou *chatbots* [2]. Il est clair que ces technologies vont être de plus en plus performantes, on parlera de plus en plus à son ordinateur ou à son téléphone, les personnes n'auront pas à apprendre à se servir de ces outils, il faudra juste leur parler.

Toutefois les interfaces en jeu dans ces technologies sont certes complexes, puisqu'elles requièrent notamment un certain décodage de la voix humaine ou celui d'un texte écrit, mais on ne peut dire actuellement qu'elles comprennent ce qui est dit de la même façon qu'un être humain le comprend. Affirmer cela, c'est bien sûr être à même de caractériser la spécificité de la compréhension humaine, et d'identifier les difficultés qu'ont actuellement les machines pour atteindre ce même niveau ; difficultés qui sont encore patentes, alors qu'on dispose de machines « surpuissantes », de très grandes capacités de stockage de données, et de concepts mathématiques et algorithmiques qui ont prouvé leur efficacité dans plusieurs domaines de l'intelligence[3]. L'écart éloignant les machines de la capacité à comprendre comme un humain est, pour résumer, la difficulté qu'elles ont à construire par apprentissage, à manipuler, à utiliser des *abstractions* de haut niveau propres aux humains : autrement dit se construire une représentation du monde sous forme de concepts, savoir les composer, les généraliser, et en conséquence pouvoir les utiliser dans plusieurs contextes. Traitements que les humains font beaucoup et relativement facilement, alors que les machines échouent largement à le faire : en simplifiant, on peut dire qu'actuellement les machines peuvent apprendre dans un contexte et appliquer ce qu'elles ont appris dans le même contexte, et cela pose bien sûr de fortes limites.

[2] Chat, pour discussion en ligne, et bot pour robot : un logiciel programmé pour simuler une conversation en langage naturel, dans le but d'apporter une certaine assistance à son utilisateur.

[3] Que l'on pense par exemple à la reconnaissance d'objets dans les images, la reconnaissance de visages, la capacité de jouer et de gagner au jeu de Go, etc.

Dans la confrontation à ces difficultés, l'Intelligence Artificielle oscille depuis le début entre deux « paradigmes », celui des systèmes symboliques et celui des réseaux de neurones[4] ; sans entrer dans le détail ni refaire l'histoire présentée dans un autre chapitre de cet ouvrage, rappelons brièvement ici les grands traits qui les distinguent, dans la perspective des approches décrites dans ce chapitre. A la base des performances des réseaux de neurones, il y a *l'apprentissage statistique*. Les réseaux de neurones peuvent intégrer de grands volumes de données diverses ; ils sont ainsi capables d'apprendre, à partir d'une multitude d'exemples de réalisations différentes d'un concept ou d'une tâche, et éventuellement de contre-exemples, à reconnaitre de façon fiable si telle réalisation particulière relève ou non de ce concept ou de pouvoir conduire cette tâche avec succès. Les performances atteintes sont assez robustes aux bruits se superposant à l'information brute portée par un texte, une image, un signal, le déroulement d'un processus, etc. Elles ne nécessitent pas l'intervention d'un expert : elles requièrent simplement les interventions élémentaires certifiant que telles réalisations sont bien des exemples, et telles autres des contre-exemples, au sein de *l'ensemble d'apprentissage.* Leurs points faibles, c'est d'abord cette nécessité, pour que l'apprentissage aboutisse, d'un très grand nombre d'exemples – les réseaux de neurones sont *very data hungry*[5] ; c'est ensuite leur *non-transparence*, leur grande difficulté voire impossibilité à rendre compte par eux-mêmes des raisons – au sens où un humain l'entend – qui les conduisent à tel ou tel choix ; par ailleurs ils ne peuvent pas facilement changer de contexte, par exemple utiliser ce qu'ils ont appris sur une tâche pour apprendre plus rapidement à réaliser une nouvelle tâche ; enfin leurs applications à l'apprentissage de la simulation du raisonnement humain restent encore élémentaires, les travaux que nous allons présenter dans ce chapitre visent précisément à élargir les performances dans ce domaine.

Concernant les systèmes symboliques, nous faisons ici référence essentiellement aux *bases de connaissances* avec leurs *ontologies* – les concepts mobilisés et leurs relations – et leurs *systèmes de règles*, basées sur la logique, permettant de conduire des raisonnements relativement complexes, compréhensibles et vérifiables par les humains. Ces *Knowledge Bases (KB)* ont fait depuis les années 1990 l'objet de travaux considérables, avec l'objectif de créer des bases de connaissances gigantesques, comportant des millions de nœuds représentant des concepts allant des plus généraux – des plus abstraits – aux plus spécifiques[6], avec l'objectif d'englober toute la connaissance sur un domaine donné. Leur construction nécessite – pour l'identification des concepts et des relations, pour la spécification des règles à prendre en compte – l'intervention de nombreux experts. Dans la réalité, de telles bases de connaissances se heurtent à la dynamique même des domaines qu'elles sont censées représenter, de telle sorte qu'il faut sans cesse rajouter des concepts,

[4] Une oscillation entre la logique et la géométrie, pour reprendre une expression de Stéphane Mallat.
[5] Pour parler trivialement, pour leur faire comprendre le sens d'un mot, le type d'objet qu'il désigne il faut leur donner 5000 exemples, alors que chez les êtres humains, trois exemples vont peut-être suffire.
[6] Pour donner une idée, une KB allant du concept de « chose » à celui de tel modèle de voiture ...

des relations et des règles pour ne pas qu'elles deviennent vite obsolètes. Et par ailleurs, la résolution de cas particuliers, la levée d'ambiguïtés, entraine une complexification des règles dont le nombre, pour atteindre les performances humaines, semble potentiellement infini. Aussi bien leur succès a-t-il été limité.

Par rapport à ce bref état des lieux, les travaux dont il va maintenant être question partent de la *conjecture* que les réseaux de neurones – et les algorithmes de la même mouvance – peuvent être adaptés pour dépasser les difficultés rencontrées par les deux paradigmes ; pour opérer (et apprendre) sur des éléments d'information symboliques, et pouvoir ainsi coder des bases de connaissances symboliques et découvrir des structures dans des ensembles de données non structurées.

## 2. Bases de connaissances : d'une approche symbolique à une approche géométrique


Le premier exemple de travail traite de la capacité des réseaux de neurones d'apprendre à coder des bases de connaissances déjà structurées – plus précisément ici des bases de données relationnelles – de manière à obtenir une structure qui puisse servir aux calculs de réponses correctes à des questions simples, alors même qu'au départ certaines relations peuvent avoir été oubliées, ou même s'avérer fausses. Le schéma ci-contre donne une idée sommaire du type d'inférence espérée : il représente une toute petite base « d'état civil » où les entités sont des personnes et des villes, les relations sont « né à » et « enfant de ». Bien que la relation soit manquante, il est *probable* (par inférence) que Jane soit née à Miami, puisque son frère John et sa sœur Patti, sont bien nés dans cette ville, et moins probable qu'elle soit née à Austin. C'est ce type d'inférence éventuellement quantifiée par une probabilité que le codage appris par le réseau de neurones devrait pouvoir faciliter. Le principe va consister à projeter la base dans un espace multidimensionnel continu – un espace vectoriel – puisque ces structures sont pour ainsi dire les structures natives sur lesquelles opèrent les réseaux de neurones ; l'inférence va passer d'une représentation logique à une représentation géométrique, en termes d'opérateurs de compositions d'opérateurs et de distance. Entrons un peu dans la technique des mécanismes en jeu.

## 2.1 Apprendre à coder géométriquement une base de données relationnelles

Le modèle relationnel structure, rappelons-le, une base de données comme une collection d'entités[7] et de relations entre ces entités ; dans « John *est né(e) à* Miami » John et Miami sont deux entités identifiées, occurrences respectives des classes d'entités « personnes » et « villes », *est né(e) à* est une relation, la relation « lieu de naissance ». Dans le modèle le plus simple, ces relations sont binaires, si bien que la base de données peut être vue comme un ensemble de triplets[8] $(h, l, t)$ où $h$ et $t$ sont les entités (*head* et *tail entities*), et $l$ (pour *label)* la relation qui les lie. On peut donc aussi, de manière plus géométrique, voir une telle base de données comme un *graphe orienté, dont les nœuds représentent les entités et les types d'arêtes, les relations.*

*2.1.1 Objectif et Principe*

Le développement des bases de données de très grandes dimensions (typiquement des centaines de milliers, voire des millions d'entités, des dizaines de milliers de relations) a amené de nombreuses questions impossibles à résoudre manuellement : comme par exemple le repérage de relations erronées ou, à l'inverse, de relations oubliées ; donc vérifier la cohérence de la base, pouvoir la rectifier et la compléter et répondre à des questions malgré erreurs et incomplétudes. Pour ce faire, il faut des algorithmes capables d'analyser sa structure – et d'en détecter des propriétés « cachées ». De nombreux travaux ont été publiés sur de telles recherches ; une des pistes[9] consiste à projeter la représentation géométrique « discrète », sous forme de graphe, dans un espace continu de relativement faible dimension $\boldsymbol{E} = \mathbb{R}^d$ ; les entités deviennent des points de cet espace et les types de relations deviennent des opérateurs géométriques, en l'occurrence, dans les travaux présentés, des translations.

Dans cette approche, chaque relation $l$ se trouve associée à un vecteur $\boldsymbol{l}$ définissant une translation dans l'espace continu $\boldsymbol{E}$. Les symboles $\boldsymbol{h}$ et $\boldsymbol{t}$ désignant respectivement les images dans $\boldsymbol{E}$ des entités $h$ et $t$, toute paire d'entités $(h, t)$ *liées par une relation* $l$ devrait vérifier l'équation $\boldsymbol{t} = \boldsymbol{h} + \boldsymbol{l}$ ; et, inversement, toute paire d'entités $(h, t)$ *non liées* par la relation $l$ devrait vérifier $\boldsymbol{t} \neq \boldsymbol{h} + \boldsymbol{l}$**.** La réalisation exacte d'une projection satisfaisant ces propriétés idéales n'est en général pas possible[10] mais, la dimension de l'espace continu $\boldsymbol{E}$

[7] Ces entités peuvent représenter des objets concrets du monde réel, comme dans l'exemple « jouet » donné ci-après, mais aussi des concepts, comme par exemple des classes d'objets, etc.

[8] Pour reprendre la notation utilisée dans Bordes et al 2015 qui nous sert ici d'article de référence.

[9] Sur laquelle nous nous focalisons donc ici, à partir notamment de l'article de Bordes et al 2015.

[10] Pour des raisons qui peuvent tenir à la structure de la base de données : par exemple, si toutes les personnes qui sont nées à Miami ne sont pas des salariés, il est clair que ces personnes ne peuvent être toutes projetées sur un même point de l'espace $E$, elles peuvent tout au plus former un « cluster » et ce cluster, regroupant les natifs de Miami, sera lui-même séparé en deux sous-groupes, les personnes salariées et celles qui ne le sont pas ; la *cardinalité* des relations (relations 1-1, 1-n, n-

étant judicieusement choisie, on peut s'en approcher. Pour ce faire, une grandeur $\mathcal{L}$ est définie[11], une tension mesurant l'écart séparant les propriétés effectives d'une projection donnée $(\ldots h, t \rightarrow \boldsymbol{h}, \boldsymbol{t}, \ldots)$ des propriétés idéales. Partant d'une projection initiale déterminée aléatoirement, un algorithme de « descente du gradient » permet de la modifier progressivement en diminuant peu à peu la grandeur $\mathcal{L}$ jusqu'à atteindre un niveau jugé acceptable.

*2.1.2. Résultats obtenus*

L'approche a été conduite sur une version de la base *Wordnet*, de l'ordre de 40.000 entités, puis sur deux sous-ensembles de Freebase[12] ; dans Wordnet, les entités représentent le sens d'un mot, ou de plusieurs mots synonymes, liés par des relations lexicales. Freebase est une base de connaissances collaborative de faits généraux ; une première expérimentation a été menée sur un sous-ensemble de 15.000 entités environ, liées par quelques 1300 relations ; une seconde expérimentation a cherché à tester l'approche sur des volumes nettement plus importants, de l'ordre du million d'entités.

Les résultats sont évalués sur la capacité de la configuration finale des entités dans l'espace $\boldsymbol{E}$ à prédire correctement les triplets correspondant à des relations vérifiées et ceux qui n'y correspondent pas[13]. Le principe est de classer, en fonction des écarts $\boldsymbol{d}(\boldsymbol{t}, \boldsymbol{h} + \boldsymbol{l})$, chaque triplet $(h, l, t)$ présent dans la base de données dans l'ensemble de tous les triplets possibles $(h, l, t')$ ou $(h', l, t)$ partageant la relation $l$ et l'une des entités $h$ ou $t$, mais valant contre-exemples, où la relation n'est pas vérifiée. Une bonne configuration finale sera celle où les triplets présents dans la base seront en moyenne classés correctement, par exemple dans le groupe des dix premiers triplets présentant les écarts $\boldsymbol{d}(\boldsymbol{t}, \boldsymbol{h} + \boldsymbol{l})$ les plus faibles. Deux critères précis sont ainsi utilisés pour comparer avec d'autres approches : le rang moyen de classement des triplets de la base, et la proportion de ceux qui sont classés dans le groupe des dix premiers.

---

1, n-n) est de fait un critère important dans l'évaluation détaillée des performances de ces approches (cf. Bordes et al déjà cité, table 4).

[11] $\mathcal{L}$ est un bilan des écarts entre $\boldsymbol{t}$ et $\boldsymbol{h} + \boldsymbol{l}$ . Ces écarts $\boldsymbol{d}(\boldsymbol{t}, \boldsymbol{h} + \boldsymbol{l})$– possiblement des distances euclidiennes – sont calculés sur les images dans $\boldsymbol{E}$ de triplets $(h, l, t)$ constituant *l'ensemble d'apprentissage.* Ces triplets sont des exemples (*h* et *t* sont liés par la relation *l*, présents dans la base) ou des contre-exemples (*h* et *t* ne sont pas liés par *l*). La valeur de $\mathcal{L}$ sera grande lorsque les écarts mesurés sur les triplets du premier type l'emportent sur ceux du second (cf. Bordes, réf. citée, équation 1).

[12] Fermée par Google en 2015, le contenu étant transféré dans Wikidata.

[13] Il faut préciser que si l'ajustement de la position des entités dans l'espace $\boldsymbol{E}$ est effectué sur l'ensemble d'apprentissage, l'évaluation est effectuée sur un *ensemble de test*, constitué de triplets jamais vus par l'algorithme durant la phase d'apprentissage. On évalue ainsi sa capacité de généralisation, de découverte de liens oubliés, ou inconnus mais qui s'avèrent vérifiés (capacité de *link prediction*).

Les expérimentations menées par Bordes et al 2015 *[1]* sur Wordnet et Freebase, ont abouti à des performances tout à fait encourageantes, notamment par rapport à d'autres méthodes : par exemple, sur la version de Wordnet comportant quelques 40.000 entités et 5000 triplets de test, l'expérimentation atteint 89% de triplets bien classés (c.-à-d. classés dans le groupe de tête des dix premiers), la dimension *k* de l'espace continu $\boldsymbol{E}$ étant de 20 ; cette proportion est encore de 47% sur le sous-ensemble de Freebase comportant de l'ordre de 15.000 entités mais quelques 60.000 triplets de test, la dimension *k* de l'espace étant cette fois-ci de 50.

## 2.2 Cas particulier de bases de données structurées par une hiérarchie

L'espace continu utilisé dans les projections $(h, l, t) \rightarrow (\boldsymbol{h}, \boldsymbol{l}, \boldsymbol{t})$ est sous-tendu par une métrique euclidienne, et les images $\boldsymbol{l}$ des relations sont des translations ; on peut légitimement se demander si d'autres options n'aboutiraient pas à de meilleurs résultats, l'option optimale dépendant de la nature des relations et de la structure sous-jacente de la base de données. Des recherches ont donc été menées dans ce sens, notamment pour des bases de données structurées, de façon latente ou explicite, par une hiérarchie ; autrement dit celles dont le graphe associé est un *arbre*, ou est proche de l'être.

### *2.2.1 Arbres et espaces hyperboliques*

On sait qu'un arbre régulier, dont chaque nœud possède le même nombre $b$ de descendants immédiats, peut être projeté sur un espace de dimension 2. Dans une telle projection, les $N(l)$ nœuds de même rang $l = 1,2,3, \ldots$ vont être placés concentriquement autour du sommet de l'arbre [14] sur un cercle $C_l$ ; pour que la densité de la distribution des nœuds demeure constante d'un rang à l'autre, il faut que la circonférence du cercle croisse proportionnellement au nombre de nœuds qu'elle doit contenir et donc, croitre, en tendance, de façon exponentielle avec $l$ [15]. Or il existe des espaces bidimensionnels qui remplissent cette condition de façon « naturelle » ; ce sont les espaces de courbure constante négative, les espaces hyperboliques. Dans un référentiel où les points d'un tel espace sont repérés par des coordonnées « polaires » $\rho, \theta$ autour d'un point origine choisi arbitrairement, l'ensemble des points de coordonnées $\rho = l, 0 \ll \theta < 2\pi$ forme un cercle $C_l$ dont la circonférence est $2\pi \sinh(l)$ . Cette circonférence croit donc de façon exponentielle avec $l$, ce qui est bien la propriété recherchée : *les espaces hyperboliques peuvent donc être pensés comme des versions continues d'arbres et vice versa, les arbres peuvent être vus comme des espaces hyperboliques discrets*[16].

Cette constatation sous-tend les travaux de Maximilian Nickel et Douwe Kiela, du centre de recherche de Facebook, autour, notamment, d'une base de données taxonomiques extraite de Wordnet ; les éléments de cette base de noms projetés sur une représentation

[14] Le nœud de rang $l = 0$.

[15] Ces propriétés n'exigent pas l'absolue régularité de l'arbre ; en général, le nombre de nœuds reste en relation exponentielle avec la profondeur.

[16] Comme il est dit dans la publication Nickel et Kiela citée en référence.

*finie*[17] d'un espace hyperbolique, appelée dans la littérature disque de Poincaré en dimension 2, et plus généralement boule ouverte de Poincaré. L'objectif est de reconstruire l'arbre taxonomique complet à partir d'informations partielles du type « le *mustang* ***est un*** *mammifère* » ; ce, par un processus d'apprentissage ; ce processus projette aléatoirement les taxons de la base sur une boule de Poincaré $\mathcal{B}$ de faible dimension, puis déplace progressivement leurs images de façon à ce que celles des taxons liés par la relation ***est un*** soient globalement proches selon la métrique de cet espace de projection.

La méthode est similaire à celle utilisée précédemment ; une fonction $\mathcal{L}(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_N})$ évalue la configuration des images dans $\mathcal{B}$ des différents taxons $(u_1, u_2, \dots, u_N)$ en exprimant une tension entre cette configuration et les relations, de type ***est un,*** inscrites dans l'ensemble d'apprentissage[18]. Un algorithme de descente du gradient est ensuite appliqué, avec cette fois-ci la particularité mathématique que ce gradient doit être calculé sur un espace non euclidien.

*2.2.2 Résultats obtenus*

La base extraite de *WordNet* sur laquelle les expérimentations ont été menées comportait quelques 82.000 noms et 740.000 relations (asymétriques) ***d'hyperonymie ;*** l'objectif a été de reconstruire la structure hiérarchique[19] du graphe sous-jacent, à partir de la donnée de tout (*Reconstruction*) ou partie (*Link Prediction*) de ces relations ; et par là même, de montrer l'efficacité de la démarche, à l'aide de procédés d'évaluation du même type que ceux mentionnés précédemment. La *figure 1* illustre, pour simplifier en dimension 2 (disque de Poincaré ), la progression du processus mis en œuvre sur le sous-arbre de la base constitué des seuls mammifères : dans une étape intermédiaire, les images des taxons sur

[17] C.-à-d. que les noms de la base taxonomique, les taxons, sont projetés sur des points dont la distance euclidienne à l'origine est strictement inférieure à 1 : donc effectivement restant à l'intérieur d'une « boule ouverte » dont la surface reste hors du domaine de projection. En dimension 2, la transformation passant des coordonnées $\rho, \theta$ de l'espace hyperbolique $(0 \ll \rho < \infty)$ aux coordonnées $r, \theta$ sur le disque de Poincaré ( $0 \ll r < 1$) est donnée par $r = \frac{\sinh(\rho)}{1+\cosh(\rho)}$ , et donc inversement $\rho = Arcosh\ \frac{1+r^2}{1-r^2}$ . L'expression de l'élément de distance, soit $ds^2 = d\rho^2 + (\sinh\rho)^2\ d\theta^2$ dans l'espace hyperbolique, devient ainsi $ds^2 = \left(\frac{2}{1-r^2}\right)^2 [dr^2 + r^2 d\theta^2]$ dans sa représentation par le disque de Poincaré. La distance, dans cette métrique, entre le centre du disque et un point se rapprochant du bord du disque augmente indéfiniment : le disque de Poincaré peut potentiellement recevoir, sa racine étant placée au centre, la projection de tous les nœuds d'un arbre de profondeur infinie.

[18] $\mathcal{D}$ étant cet ensemble, la fonction $\mathcal{L}$ met en rapport la distance $d(\boldsymbol{u}, \boldsymbol{v})$ de chaque couple de termes $(u, v) \in \mathcal{D}$ (qui satisfont donc la relation ***est un***, *exemples positifs*) avec les distances $d(\boldsymbol{u}, \boldsymbol{v'})$ de couples de termes $(u, v') \notin \mathcal{D}$, constituant donc des contre-exemples ou *exemples négatifs* (Voir la formule 6 de l'article cité). La valeur de $\mathcal{L}$ est élevée lorsque les distances associées aux exemples positifs sont grandes et celles associées aux exemples négatifs petites.

[19] En termes mathématiques, assurer *la clôture transitive* de cet ensemble de relations.

le disque de Poincaré forment encore une configuration assez confuse, avec plusieurs noms d'espèces – tel que « écureuil » – encore placés au centre du disque.

La configuration finale obtenue à l'issue de processus fait au contraire clairement apparaitre la structure hiérarchique, avec le taxon de rang le plus élevé, « mammifères » près du centre, et les noms des différentes espèces – les feuilles de l'arbre – réparties quasi uniformément à proximité de la circonférence du disque.

D'autres expérimentations ont été menées dans des contextes différents, expérimentations qui ont montré également l'avantage d'une projection sur une boule de Poincaré par rapport à une projection sur un espace euclidien ; ainsi en a-t-il été dans la prédiction de liens dans certains réseaux sociaux (par exemple de collaboration entre scientifiques) dont on pense qu'ils sont en partie structurés par une hiérarchie latente. De même dans le traitement de réseaux sémantiques, dont les liens représentent des *assertions graduées*, c'est-à-dire affectées d'une valeur de vraisemblance, d'un degré de certitude.

## 3. Réseaux de mémoire

John dropped the Milk
John took the Milk there
Sandra went to the bathroom
John moved to the hallway
Mary went to the bedroom
*Where is the Milk ?*

Comme suggéré dans l'introduction de ce chapitre, il est difficile de doter une machine de la capacité à *comprendre* un texte, dans l'acception suivante : être à même *d'inférer* la réponse correcte à une question à partir des faits énoncés dans ce même texte. Cette difficulté se manifeste même lorsque ce texte raconte une histoire très simple, comme par exemple l'histoire donnée dans l'encadré ci-contre[20], et que la question porte sur un fait pourtant très vite inféré par un être humain. Pour réaliser de telles tâches[21], on a pu penser qu'une approche d'apprentissage profond classique pourrait être une voie suffisante, en s'appuyant sur la fourniture d'un ensemble d'apprentissage composé d'un grand nombre de ces petites histoires très courtes, des questions et réponses associées. Or il n'en a rien été ou plus exactement le taux de réussite, environ d'un cinquième dans de premières tentatives, s'avérait insatisfaisant.

[20] Extrait de *Sainbayar Sukhbaatar et al. 2015 [3].*

[21] En fait, une vingtaine de catégories de tâches ont été identifiées pour comparer différents algorithmes proposés par la communauté (cf. *J. Weston, A. Bordes et al 2015, Towards AI-Complete Question Answering : A set of Prerequisite Toy Tasks [5])*. Ces catégories se différencient par exemple par le nombre de faits (*supporting facts*) ou par le nombre d'arguments dont le raisonnement doit tenir compte, ou encore par la prise en compte de notions temporelles, spatiales, etc.

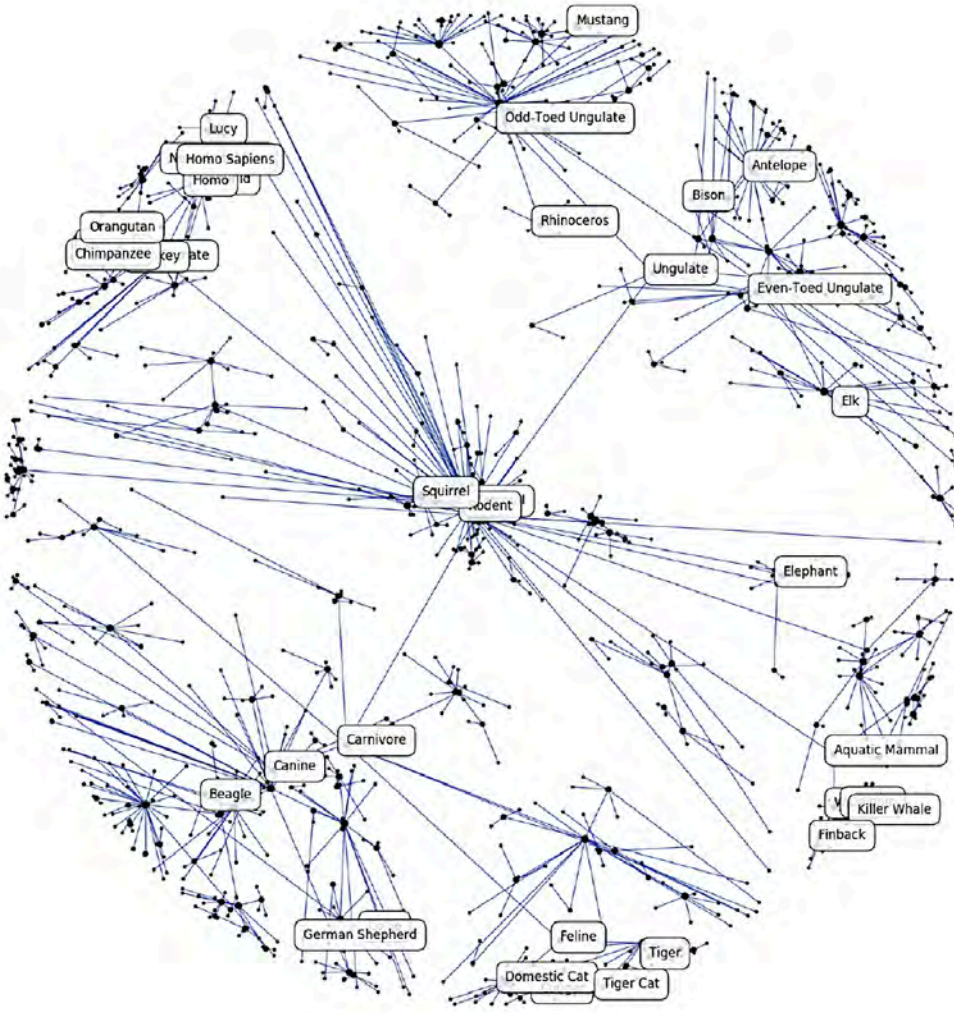



*Figure 1a : étape intermédiaire*

**Figure 1** Progression de l'apprentissage projetant la taxonomie des mammifères sur le disque de Poincaré, pour en refléter la structure hiérarchique. Image extraite de Maximilan Nickel et Douwe Kiela 2017 [8]

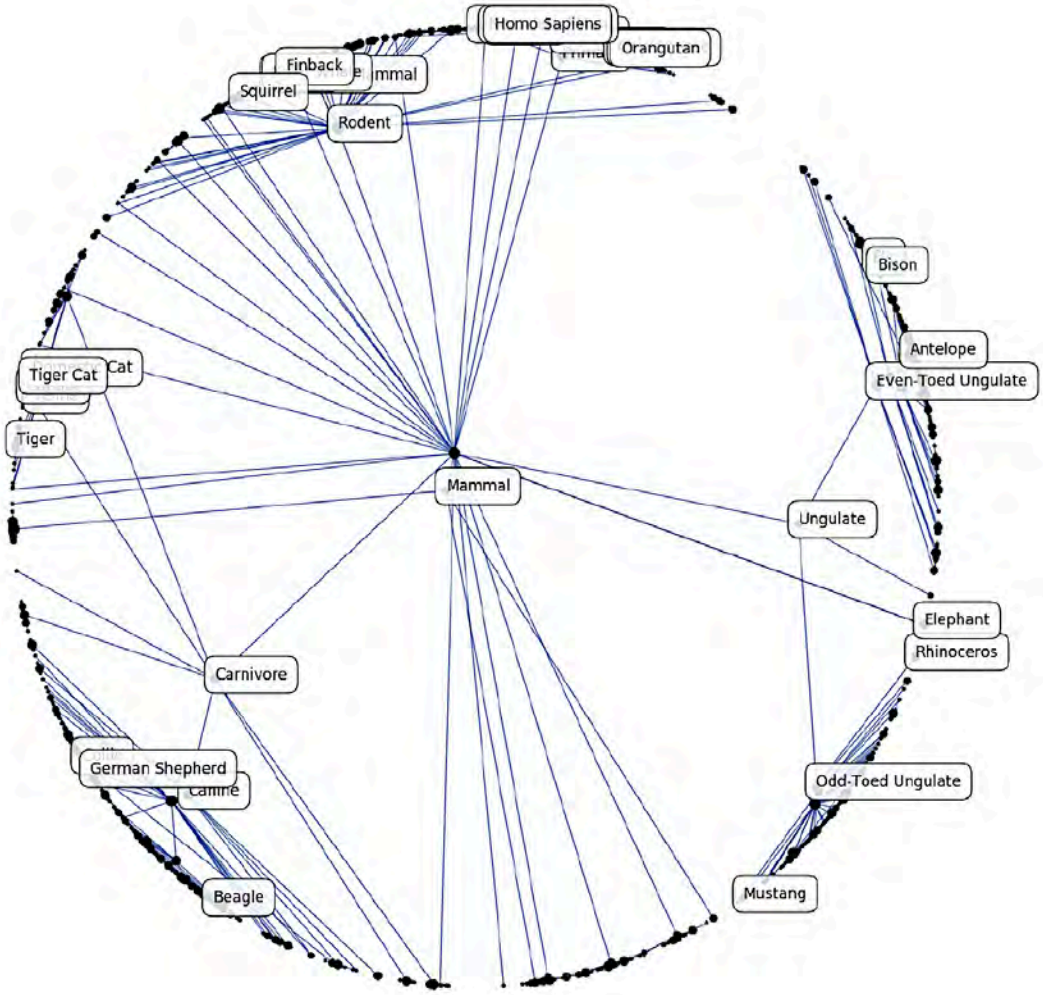



*Figure 1b : étape finale*

### 3.1 Principes

Ces premières tentatives ont conduit à la conclusion qu'il manquait une mémoire à court terme, capable de stocker les éléments pertinents pour la réalisation de la tâche. On s'est alors orienté vers des algorithmes capables d'apprendre *simultanément*, pour ainsi dire au sein d'un même processus, la manière d'encoder dans cette mémoire à court terme ce que racontent les textes et les questions posées, d'élaborer et d'encoder des éléments intermédiaires de raisonnement et enfin d'inférer les réponses. Telle est l'idée à la base des réseaux de mémoire[22] ; de nombreuses publications leur ont été consacrées, nous allons dans un premier temps, pour donner au lecteur *un exemple* de tels processus, reprendre l'une d'entre elles, celle de *S. Sukhbaatar & al. 2015, End-To-End Memory Networks [3]*, avec application à la compréhension d'histoires simples. Dans un second temps, nous décrirons brièvement l'application des mêmes principes algorithmiques à des applications visant à répondre à des questions posées à de grandes bases de données et de grandes bases de textes.

### 3.2 Répondre à des questions sur des histoires simples

La recherche d'une réponse à une question du type *Where is the Milk ?* dans l'exemple de l'encadré passe intuitivement par la recherche en mémoire des énoncés qui paraissent le plus en rapport, à travers les termes mis en œuvre, avec la question posée. Une recherche itérative : on commence par retenir un premier énoncé, qui pourra être *John took the milk there*, puis la combinaison de son contenu et de celui de la question initiale sert de base à une seconde recherche, qui sélectionnera par exemple *John moved to the hallway*. C'est bien ce type de démarche itérative qui est mise en œuvre dans les algorithmes ; l'originalité réside dans la manière dont sont codés les énoncés, la question posée, les contenus intermédiaires et la réponse à fournir ; elle réside aussi dans la manière dont s'effectue les rapprochements de contenu dont il vient d'être question. Tous les calculs s'effectuent en effet dans un espace de *représentations internes* particulières dont on pourrait dire, dans la mesure où elles permettent à l'algorithme de fournir des réponses correctes, que d'une certaine manière elles capturent le sens des énoncés.

Dans la publication citée en référence, l'idée est d'inscrire ces représentations internes dans *un espace vectoriel d'une certaine dimension d* ; puis d'effectuer les rapprochements par la mesure de relations entre différents vecteurs ; on retrouve ainsi le parti pris géométrique utilisé pour raisonner sur les bases de données. La transformation d'un énoncé en sa représentation interne s'effectue en deux étapes : une première étape code d'une manière simple et déterminée par le concepteur, l'énoncé dans un premier espace vectoriel de

[22] Le lecteur pourra trouver dans *J.Weston, S.Chopra & A. Bordes 2015, Memory Networks [4],* une description algorithmique générale du fonctionnement de ces réseaux de mémoire.

grande dimension $V$[23] ; une seconde étape applique une matrice $A$ de $d$ lignes et $V$ colonnes, ($avec\ d \ll V$) pour projeter cette première représentation dans la représentation interne finale désirée ; mais, cette fois-ci, les coefficients de $A$ et ceux d'autres matrices $B$, $C$ du même type, ainsi que ceux de la matrice de « sortie » $W$ de dimension $V$ x $d$[24] ne sont pas fixés par le concepteur : *ils vont être déterminés par apprentissage.*

Très schématiquement, l'algorithme d'élaboration de la réponse comporte une succession d'étapes (*hops* ou *layers*) indicées par $k = 1, 2, 3$ ... où interviennent les opérations suivantes :

- Mise en mémoire d'une première série (*inputs*) de représentations internes $m_i$, des différents énoncés $x_1, \dots, x_i, \dots, x_n$ de l'histoire :
$$m_i = A^k x_i$$
- Mise en mémoire d'une seconde série (*outputs*) de représentations internes $c_i$ , issues des différents énoncés $x_1, \dots, x_i, \dots, x_n$ de l'histoire, qui vont servir directement, à cette étape, à l'élaboration de la réponse :
$$c_i = C^k x_i$$
- Appariement de ce qu'on peut appeler *l'état de la réponse à cette étape de l'itération*, savoir un vecteur $u^k$, avec les différents représentations internes $m_i$, pour apprécier leur pertinence, à cette étape, dans l'élaboration de la réponse finale ; cette pertinence est codée, pour chaque $m_i$, par une probabilité $p_i$[25].

- Élaboration, à partir de l'état de la réponse $u^k$, des vecteurs $c_i$ et de leur poids de pertinence $p_i$, d'un nouvel état de la réponse $u^{k+1}$, utilisable à l'étape suivante.

Dans la première étape, le vecteur $u^1 = Bq$ est simplement la représentation interne de la question ; lorsque la dernière étape est atteinte[26], $u^{k+1}$ est la représentation interne de la réponse $r = Wu^{k+1}$.

L'apprentissage s'effectue à partir de la donnée d'un grand nombre[27] de triplets histoires, questions, réponses fournies ; il s'agit donc d'un apprentissage faiblement supervisé : les énoncés marquants, déterminants pour la réponse à la question – les *supporting facts* – ne

[23] Le codage le plus simple – qui ne tient pas compte de l'ordre des mots – transcrit l'énoncé dans un espace dont la dimension est la taille du vocabulaire, chaque mot de ce vocabulaire étant associé à une composante de ce vecteur ; une valeur 1 d'une composante de ce vecteur signalant que le terme associé est présent dans l'énoncé (méthode identifiée sous l'expression *bag-of-words*).

[24] Transformant la représentation interne de la réponse à son énoncé en mots du vocabulaire.

[25] Une manière de faire cet appariement est d'effectuer le produit scalaire $m_i . u^k$ (dans l'espace vectoriel des représentations internes, de dimension $d$).

[26] Avec, par exemple, $k = 3$, modèle à trois étapes ou couches.

[27] Typiquement 1000 ou 10.000 exemples.

sont nullement indiqués comme tels durant la phase d'apprentissage[28]. Cet apprentissage opère par descente du gradient dans l'espace des paramètres, en ajustant progressivement les coefficients des matrices $A^k, C^k, \ldots, k = 1, 2, 3, \ldots,$ applicables aux différentes étapes ainsi qu'aux matrices $B$ et $W$. Pour réduire la dimension de cet espace, certaines contraintes sont posées sur ces matrices ; par exemple $A^{k+1} = C^k$ : les outputs $c_i$ d'une étape (éléments retenus pour le calcul du nouvel état de la réponse à cette étape) seront les inputs $m_i$ de l'étape suivante ; ou encore, plus drastiquement, les matrices $A$ et $C$ ne varient pas d'une étape à l'autre.

Les publications faites sur ces travaux analysent systématiquement leurs performances en fonction de différentes options possibles sur l'algorithme présenté, en fonction également des performances atteintes par d'autres approches. La publication de *S. Sukhbaatar 2015* commentée ici donne ainsi une idée des résultats atteints[29] : retenons ici qu'avec la meilleure option, dans un apprentissage sur 10.000 exemples effectué indépendamment pour chacune des 20 catégories de tâches, *le taux d'erreur* dans les réponses apportées est *en moyenne sur l'ensemble des catégories* de 6,6%, ce taux excédant 5% seulement dans 4 catégories. Ces résultats s'avéraient à l'époque nettement supérieurs à ceux d'autres travaux, excepté l'un d'eux utilisant un apprentissage fortement supervisé. Récemment (2017) des approches similaires (*EntNets, Query-reduction Networks*) ont réussi à résoudre correctement l'ensemble des 20 catégories.

### 3.3 Répondre à des questions sur de grandes bases de données ou de textes

La recherche d'une réponse à une question portant sur des histoires simples composées d'une dizaine ou d'une centaine de phrases est le prélude à la réalisation d'ambitions plus vastes : pouvoir interroger en « langage naturel » de grandes bases de connaissances, comportant des millions d'éléments, et obtenir des réponses pertinentes.

Que ces bases de connaissances soient structurées en bases de données (ensembles de triplets entités-relation $\boldsymbol{h, l, t}$) ou simplement constituées par un ensemble de textes, le schéma d'élaboration de la réponse est le même et se déploie en deux étapes : une première étape de *filtrage* rassemble, à partir des mots de la question – un relativement petit nombre d'éléments d'information – de triplets entités-relation ou de phrases de textes et les insère dans la mémoire du réseau. Une seconde étape utilise des méthodes inspirées de celles que nous avons vues précédemment pour construire la réponse.

S'agissant des bases de données relationnelles, on pourrait penser que le problème est simple : un utilisateur connaissant la structure de la base et le langage d'interrogation SQL[30]

[28] Contrairement à l'apprentissage fortement supervisé, appliqué par certains algorithmes, où l'on indique à chaque étape l'énoncé – le *supporting fact* – qui doit être utilisé.
[29] Table 1 de la publication.
[30] Structured Query Language.

n'aura aucun mal à poser dans ce langage une question précise et en obtiendra la réponse ; la clé d'une obtention rapide de cette réponse réside dans le bon fonctionnement de mécanismes d'indexation. Mais ce problème change de nature lorsque l'utilisateur ne sait rien de cette structure et qu'il pose sa question comme l'on parle ; même lorsque la réponse requiert un raisonnement très simple s'appuyant sur un seul « fait » énoncé par un triplet présent dans la base[31]. Car en pratique l'étape de filtrage va encore laisser ouverte plusieurs possibilités dont une seule contient le fait permettant de répondre à la question ; dans l'exemple illustré[32] par la figure, trois triplets, entre autres[33], sont sélectionnés à partir de la question *What year was the movie Blade Runner released ?* , le « fait » permettant la réponse étant évidemment « Blade Runner, release_year, 1982 ».

L'apprentissage conduisant le système à donner les bonnes réponses à ce type de question s'effectue selon le même principe que pour les questions sur les histoires simples : le réseau de mémoire code *dans le même espace vectoriel* aussi bien la question, les triplets issus de l'étape de filtrage et ses propositions de réponse ; il affine, sur la base de l'ensemble d'apprentissage, les paramètres de ce codage de manière à ce que l'évaluation des valeurs de pertinence[34] de chaque « fait » fournisse la réponse correcte.

Mais l'ambition peut être portée plus loin. En effet, l'information initiale, concernant ces connaissances générales traitées ici, est portée par des articles, des images, des livres, etc. ; leur encapsulation dans des bases de données structurées servant ensuite de socle aux questions des utilisateurs, apparait comme une étape intermédiaire qu'il est peut-être possible d'éviter ; à condition bien sûr de savoir opérer directement à partir des textes, par exemple à partir d'articles de *Wikipédia*. C'est bien à cette tâche que se consacrent plusieurs travaux actuels.

Bien sûr, le problème est plus difficile car, comme indiqué dans Miller et al. 2015, dont nous reprenons ici le travail à titre d'exemple, l'expression de l'information est dans les textes souvent indirecte, ambigüe et dispersée dans de multiples documents ; aussi bien ces auteurs ont spécifié un réseau de mémoire utilisable dans les deux contextes, base de données et textes, aux fins de comparaison des performances accessibles. La structure de ce réseau est celle de S. Sukhbaatar et al. Les deux représentations internes $m_i$ et $c_i$ (les inputs et outputs de l'algorithme décrit précédemment) reçoivent respectivement les rôles *Key* et *Value ;* un élément d'information donné est partagé entre ces deux mémoires d'une manière différente selon qu'il s'agit d'un triplet (cas des bases de données, *figure 2*) ou d'une phrase (cas des textes, *figure 3*).

---

[31] Comme il est dit dans *Bordes et al. 2015, Large-scale Simple Question Answering with Memory Networks [2].*

[32] Reprenant celui discuté dans Miller et al. 2016 Key-Values Memory Networks for Directly Reading Document.

[33] Comme "Black Runner *starred_actors* Harrison Ford, Sean Young", etc.

[34] Par des opérations de type produit scalaire.

Des tests de performances ont été réalisés, en expérimentant différents types de partage des éléments d'information entre *Key* et *Value*, et ce sur trois contextes : sur une base de données créée à partir d'une documentation de base par des opérateurs humains, sur une base de données générée automatiquement à partir de la même documentation[35] et enfin directement sur cette documentation de base, constituée d'un ensemble d'articles de Wikipédia dans le domaine cinématographique. Treize catégories de questions ont été mobilisées pour ces tests. L'ensemble d'apprentissage a comporté de l'ordre de 100.000 couples questions-réponses. Les meilleures performances ont été atteintes sur le premier contexte – base de données construite par des opérateurs humains, avec un taux de bonne réponse de 93,9%. Viennent ensuite celles obtenues sur le troisième contexte opérant directement sur les articles Wikipédia, avec 76,2% de réussite, et descendent à 68,3% dans le cas d'une base de données construite automatiquement. Ces performances que Miller et al analysent en détail par type de question sont tout à fait intéressantes lorsqu'on les compare à d'autres méthodes appliquées sur les mêmes contextes.

## 4. Conclusion

La compréhension du langage par les machines fait des progrès rapides dus à la combinaison de méthodes capables d'apprendre sur de grandes quantités de données avec des modèles et des concepts venant de l'Intelligence Artificielle symbolique. Néanmoins, il reste encore de multiples étapes à franchir, de multiples verrous scientifiques à lever, avant de pouvoir converser avec une machine comme on le fait avec un humain. En effet, nos conversations s'appuient toutes sur le socle de notre connaissance commune, que l'on appelle le sens commun, qui informe nos paroles et les ancrent dans le monde physique, philosophique et sociologique. Pour l'instant, ce sens commun est inatteignable par les machines. Comment le sens commun peut-il être encodé dans un programme ? Doit-il être précodé ou appris à partir d'expériences ? Les recherches sur de telles questions sont encore aujourd'hui très préliminaires et recèlent encore de nombreuses inconnues.

[35] Le principe de cette génération automatique étant d'exploiter les regroupements sujet-verbe-objet.

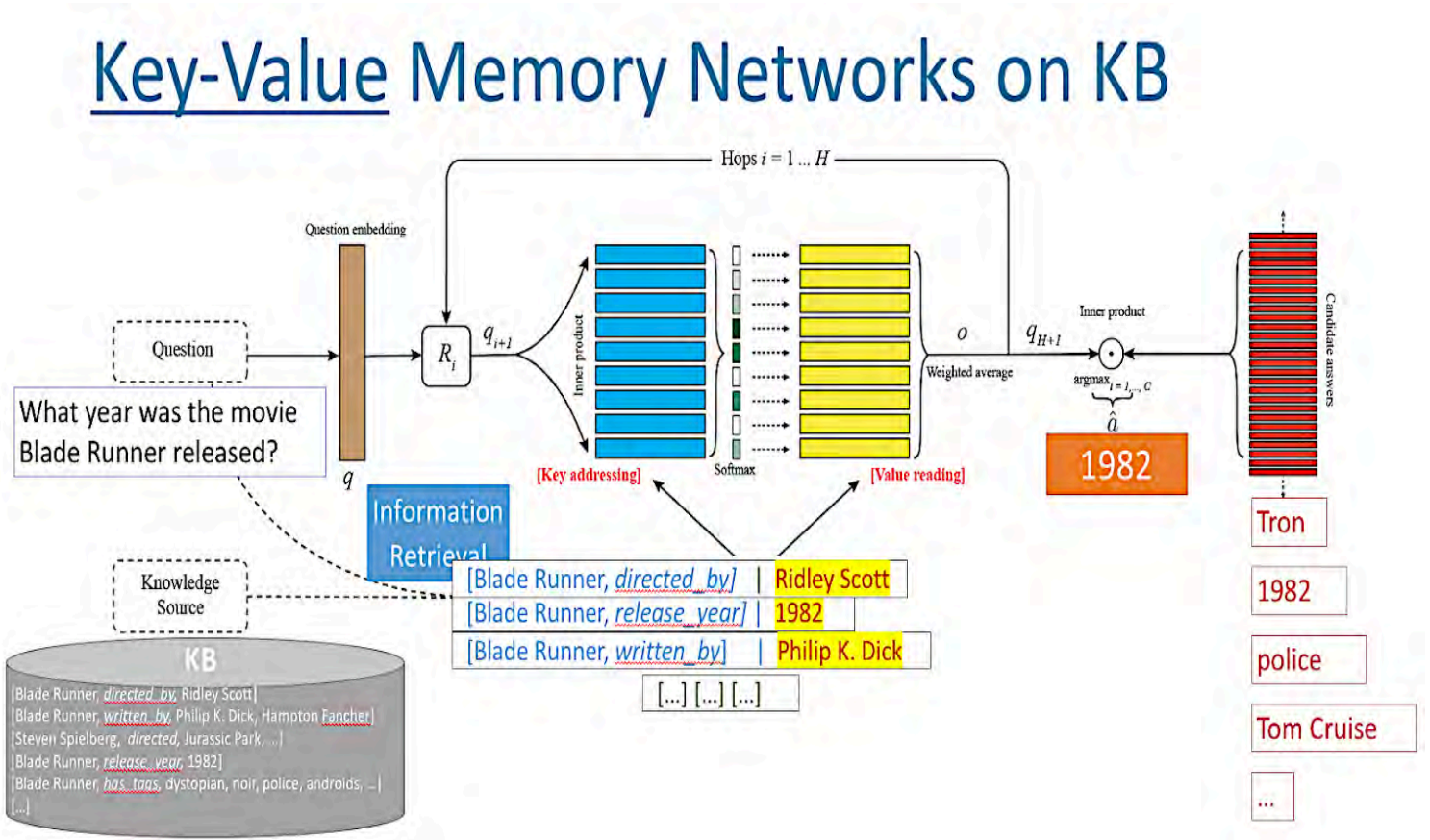



**Figure 2** illustrant les deux étapes de la recherche de la réponse à une simple question sur une base de données cinématographique ; image construite par A. Bordes à partir de Bordes et al 2015 [2] et Miller et al. 2016 [6].

# Key-Value Memory Networks on Text

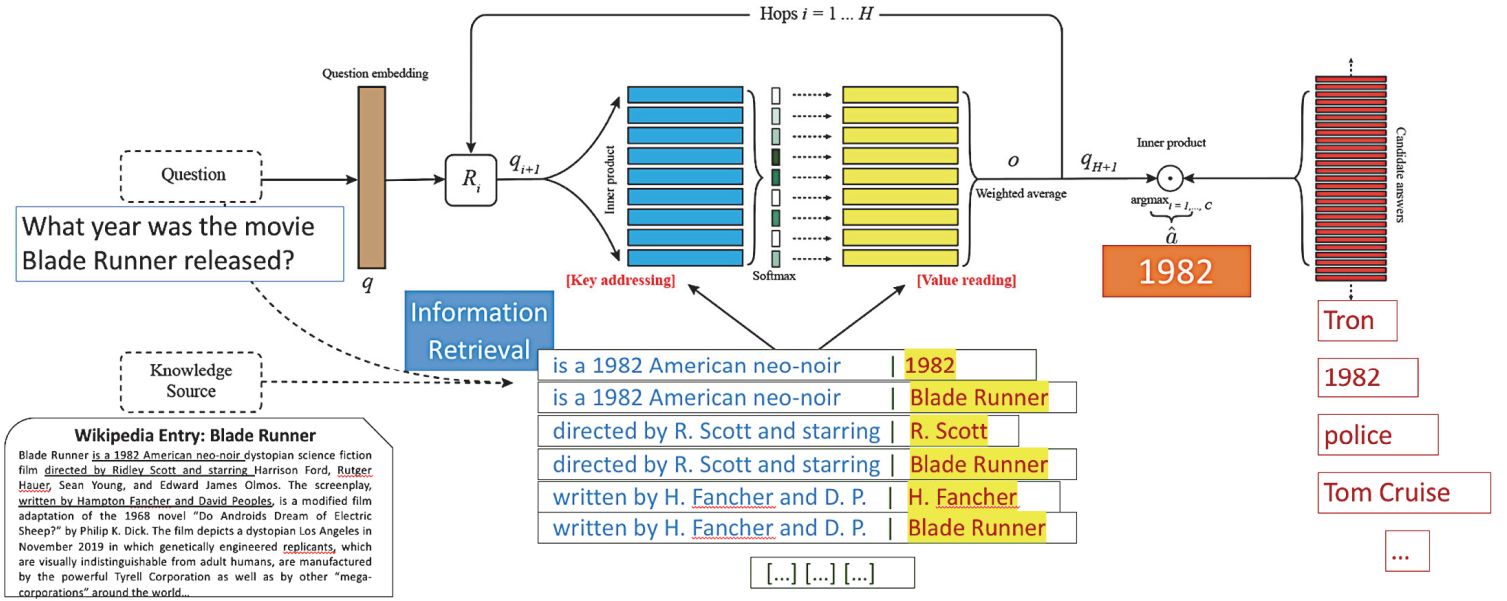



**Figure** 3 illustrant les deux étapes de la recherche de la réponse à une simple question sur une base de texte (Wikipédia) ; image construite par A. Bordes à partir de Bordes et al 2015 [2] et Miller et al. 2016 [6].

**Références**

[1] Antoine Bordes, Nicolas Usumier, Alberto Garcia-Duran, Jason Weston. Translating Embeddings for Modeling Multi-relational Data, *Advances in Neural Information Processing Systems 26 (NIPS 2013).*
[2] Antoines Bordes, Nicolas Usumier, Sumit Chopra, Jason Weston. Large-scale Simple Question Answering with Memory Networks, *arXiv 5 Jun 2015.*
[3] Sainbayar Sukhbaatar, Arthur Szlam, James Weston, Rob Fergas. End-to-End Memory Networks, *arXiv 24 nov 2015, Advances in Neural Information Processing Systems 28 (NIPS 2015).*
[4] Jason Weston, Sumit Chopra, Antoine Bordes. Memory Networks, *arXiv 29 Nov 2015, conference paper at ICLR 2015.*
[5] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander Rush, Bart van Merriënboer, Armand Joulin, Tomas Mikolof. Toward AI-Compete Question Answering ; a Set of Prerequisite Toy Tasks. *arXiv 31 Dec 2015.*
[6] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, Jason Weston. Key-Value Memory Networks for Directly Reading Documents, *arXiv 10 Oct 2016.*
[7] Danqi Chen, Adam Fisch, Jason Weston, Antoine Bordes. Reading Wikipedia to Answer Open-Domain Questions. *arXiv 28 Apr 2017.*
[8] Maximilian Nickel, Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. *arXiv 20 May 2017.*

# 9

# L'Origine et l'Evolution du Langage[1]

**Luc STEELS**
ICREA
Institut de Biologia Evolutiva
(UPF/CSIC), Barcelona
Université de Bruxelles (VUB)

## Abstract

How can we explain the appearance, the evolution and the diversity of languages in human societies? To address these questions Luc Steels supports the hypothesis of an analogy between the mechanisms of language evolution and those of biological evolution: namely processes of replication / transmission, mutation and selection, to which are added prioritization processes at different levels of organization. After showing examples of the intervention of such processes in languages, Luc Steels explains the methods emerging from Artificial Intelligence and robotics which allow him to test his hypothesis through various experiments. In these experiments he first shows how Artificial Intelligences, interacting with one other, brings out a common vocabulary, and a common meaning given to each word of this vocabulary. And beyond that, still through such interactions, he shows how can emerge and why, there can also be emergence of elementary grammatical structures such as number or gender agreements or even more complex structures such as sentence nests. To conclude, he reasserts the idea that languages are permanently changing cultural systems, under the effect of dynamics of a nature similar to those at work in the evolution of species, while acknowledging that this idea is far from winning unanimous support.

[1] Ce chapitre est la transcription, effectuée par Alberto Oliverio (Université de Rome), Ernesto Di Mauro (Dipartimento di Scienze Ecologiche e Biologiche, Università della Tuscia, Italie), et Jean Pierre Treuil (AEIS) de la conférence de Luc Steels faite devant l'AEIS le 12 juin 2017. Le texte a été relu par le conférencier ; il est publié avec son accord

# 1. Introduction

Le thème de ce chapitre concerne l'origine du langage et les mécanismes de son évolution. Nous présentons ici les principes de notre approche de cette question très débattue, puis montrons comment on peut l'explorer et la valider en utilisant les techniques de l'intelligence artificielle.

L'origine du langage : que s'est il passé au début ? Comment est-il apparu ? Cette question, à laquelle il faut essayer de répondre scientifiquement, est très liée à une autre, celle de la diversité des langues ; un changement a pu se produire, suite à une mutation survenue chez un individu, se manifestant donc dans *un* cerveau. Mais comment se fait-il qu'il y ait des millions de gens qui parlent la même langue et que parallèlement il y ait des milliers de langues différentes. Et que, à l'intérieur d'une même communauté linguistique, existe encore énormément de diversité.

Face à de telles questions, nous pensons pouvoir affirmer que, pour le moment, il n'existe pas vraiment de théorie qui fasse l'unanimité. Force est de constater un certain retard, par rapport à la biologie et la théorie de l'évolution des espèces. La biologie s'est révélée être pour nous une piste et nous avons trouvé, dans le Laboratoire de Biologie Evolutionniste de Barcelone, où l'on travaille en particulier sur les Primates, un environnement très favorable à notre propre réflexion.

Il faut rappeler en effet qu'en linguistique, le thème de l'évolution est très peu abordé, voir un peu tabou, témoin le fait que, à la création de la Société Linguistique de Paris, on a pu affirmer « nous, on ne parlera pas d'évolution ». C'était un peu le trait fondateur de cette science. Et même, plus récemment, des linguistes comme Chomsky, par exemple, étaient complètement opposés à l'idée qu'on pouvait formuler une théorie de l'évolution du langage.

Pourquoi le problème de l'évolution du langage est-il considéré comme aussi difficile ? La raison réside dans le fait que cette évolution se manifeste sous plusieurs aspects : un premier aspect biologique, celui de changements très probablement survenus dans le cerveau, affectant sa structure et augmentant ses capacités ; un second aspect éthologique ou sociologique, celui des changements affectant les interactions sociales, les systèmes sociaux, et corrélativement, les besoins de communication. Enfin, un troisième aspect relevant plus spécifiquement de la linguistique, celui des changements affectant les systèmes de communication eux mêmes. Ces trois aspects sont en interaction, comme le montre la figure 1 ; une capacité plus grande du cerveau va permettre un langage plus complexe, lequel à son tour permettra une plus grande complexité des relations sociales, etc. On se retrouve ainsi au carrefour entre trois disciplines : la biologie, l'éthologie sociale (ou l'anthropologie sociale pour les humains) et la linguistique. On va donc examiner les relations entre ces différents aspects.

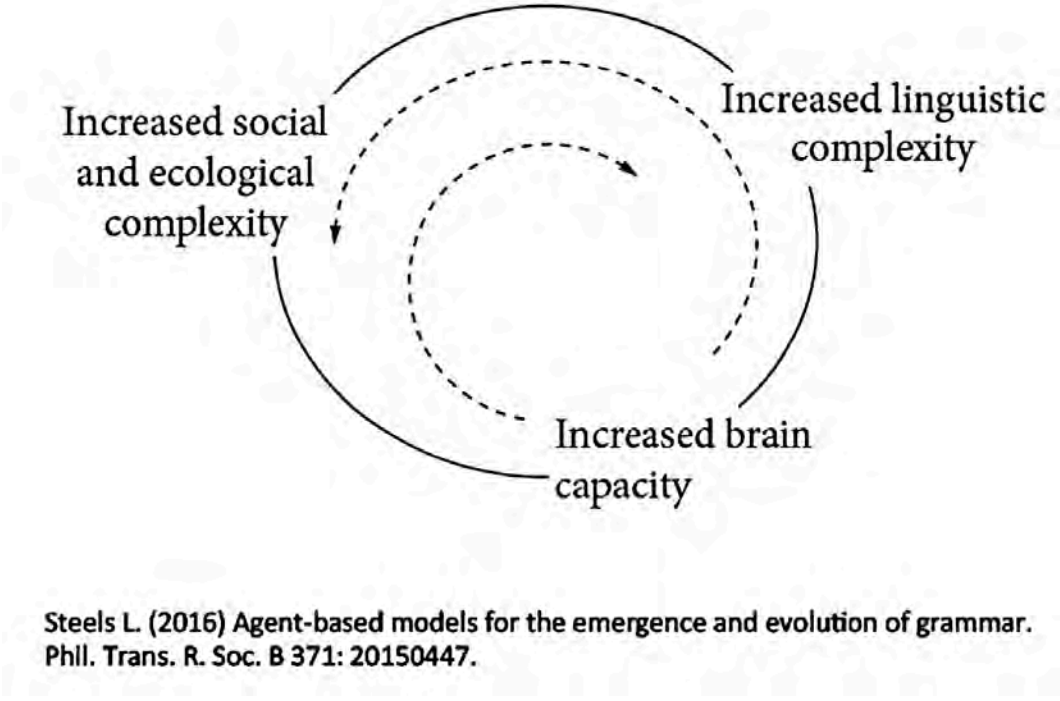



**Figure 1.** Les évolutions biologiques, sociales et culturelles interagissent. Cette interaction amène à une augmentation réciproque de complexité, jusqu'au moment où un bilan fonctionnel est atteint [1].

### 1.1 S'inspirer de la biologie

Lorsque l'on parle avec des biologistes, leur question est toujours « comment peut-on passer d'un état A à un état B ? ». Par exemple, chez les salamandres, on trouve une espèce qui n'a pas de poumons, alors que les autres salamandres ont toutes des poumons. Alors, que s'est-il passé ? L'état A et l'état B sont tous deux présents, comment cela a-t-il pu se produire ? Autres exemples : il y a des poissons, les cyclides, qui ont des taches sur leurs queues, tandis que d'autres espèces de cyclides en sont dépourvues. A un certain moment, dans une espèce qui en était dépourvue, les taches sont apparues. Il y a des espèces qui sont au début solitaires mais chez lesquelles, à un certain moment, apparait un comportement collectif, comme chez les abeilles ; en remontant encore plus loin dans le temps, des organismes unicellulaires deviennent multicellulaires ; encore plus en arrière dans le temps, certaines réactions biochimiques se manifestent et, à un certain moment, la vie émerge : des organismes peuvent se copier eux-mêmes.

Telles sont les questions qu'affronte la biologie. Nous pensons qu'on peut poser les mêmes types de questions pour le langage (cf. figure 2). Il existe des langues pour lesquelles il n'y a pas, ou très peu, d'expressions pour la couleur, qui ne peuvent exprimer, par exemple

que le *noir* et le *blanc* mais pas le *jaune*, le *bleu*, etc. De nombreuses autres langues ont au contraire plusieurs façons d'exprimer des couleurs. Un autre exemple concerne la présence d'articles : dans beaucoup de langues, il n'y a pas d'articles, par exemple le japonais, et il existe évidemment des langues qui ont des articles ; et il y a ce phénomène appelé progression : le latin par exemple n'avait pas d'articles, tandis qu'ils existent en français, en italien, espagnol, etc

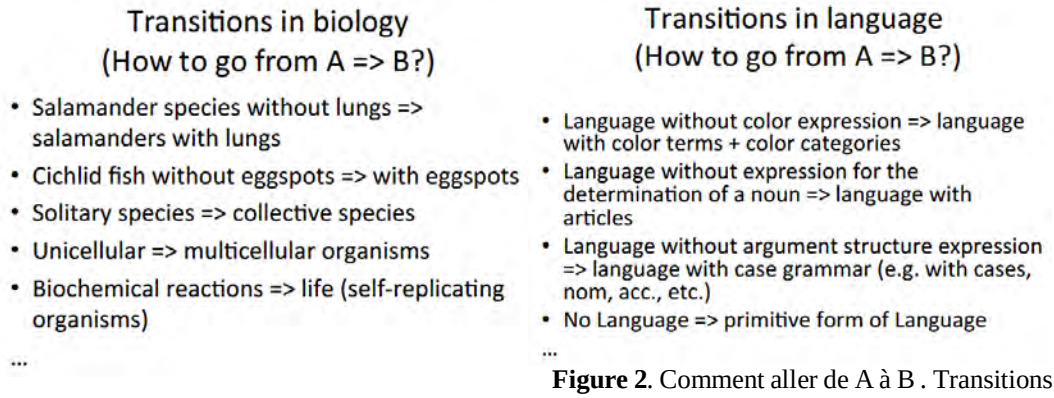



**Figure 2**. Comment aller de A à B . Transitions en biologie et transitions dans le langage [15, et 16, Planche 10].

Un autre exemple concerne la possibilité d'exprimer le rôle des objets dans une action : *Jean donne un livre à Marie* ou *Marie donne un livre à Jean* ; dans ces phrases le rôle de Jean est différent de celui de Marie. En français, cet effet s'obtient en changeant l'ordre des mots, en variant la façon de les situer dans la phrase. Mais en latin par exemple le même effet s'obtient par l'usage des cas, savoir le nominatif, l'accusatif, le datif, ce qui constitue une autre façon de construire le discours. Et il existe des langues qui, à un certain moment, perdent la capacité de le faire, par un phénomène qu'on peut qualifier d'érosion.

Nous pensons que cette prise de parti scientifique, postulant de fortes similitudes entre les mécanismes à la base de l'apparition de la vie et de l'évolution biologique et ceux à la base de l'apparition du langage et de l'évolution des langues, est la bonne. Notre hypothèse est qu'elle permet d'aller aux origines, au moment de l'histoire humaine où l'on peut dire : là il y a des humains qui sont sans doute très intelligents, qui possèdent peut-être déjà des systèmes de communication, mais qui ne possèdent pas de langage tels que nous le connaissons. Puis cette capacité est apparue et il s'agit de comprendre le processus de cette apparition.

### 1.2 Priorité à la question du sens

Classiquement les linguistes s'intéressent d'abord à la morphologie (la forme des mots) et à la syntaxe. C'est en effet ce qui est concret, visible. Nous pensons qu'il est au contraire préférable de commencer par le sens et poser la question : « comment le sens est-il exprimé ? ». On constate que le langage utilise à cette fin plusieurs procédés : un premier

procédé mobilise le contexte : quel est le sens que la phrase exprime dans son contexte ? Cela implique que pour deviner le sens, il faut effectuer un raisonnement, une inférence. Considérons par exemple la phrase *Je vais ce soir à Bruxelles* (comme en anglais *I go to Bruxelles tonight*) ; « *Je vais* » stricto sensu, entendu isolément, ne comporte pas, ne renvoie pas une expression du futur ; cependant tout le monde comprend que je parle du futur. Pourquoi ? Parce que je suis ici (et non à Bruxelles), nous sommes maintenant (et non ce soir) et ce soir c'est aujourd'hui. De nombreux éléments de sens ne sont donc pas exprimés, mais nous sommes des êtres intelligents qui peuvent les inférer, les deviner. Un second procédé – lexical – est l'expression directe par un mot, d'un fragment, d'un segment de sens, *ordinateur*, *table*, *fenêtre,…* Un troisième procédé – syntaxique – compose un fragment de sens à travers un assemblage ordonné de mots qui y conservent cependant leur individualité : *the Roman poet* ; cet assemblage agit comme une unité qui peut être combinée à d'autres unités formant un assemblage plus complexe porteur d'un sens plus riche : *The Roman poet wrote pretty boring sonnets*. Un quatrième procédé enfin – morphologique – exprime un fragment de sens par des mots complexes formés d'un cœur lexical associé à des « marqueurs » qui modifient la forme du mot et expriment des notions telles que le nombre, le genre, ou encore un temps … *open* (le cœur), *ed* (le marqueur), *opened*.

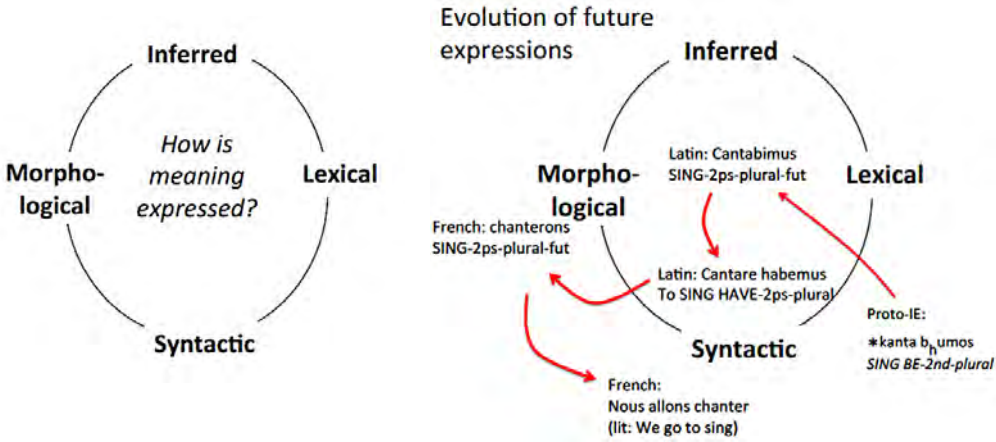



**Figure 3.** Comment le sens s'exprime-t-il ? [16, planches 41 et 42]

### 1.3 Des changements permanents dans la manière d'exprimer un sens

Il existe donc 4 façons d'exprimer le sens dans le langage. Or les linguistes ont observé que ce système présente des cycles. Ainsi (cf. figure 3), dans l'expression du futur d'un mot défini, *chanter,* construit dans le proto-indo-européen sur la racine *kanta*, on le trouve associé au verbe $b_h$*umos*, à savoir le verbe *être*, à la première personne du pluriel. Mais lorsqu'on examine l'évolution du système, si, par exemple, on regarde le Latin, on voit que les deux mots séparés sont devenus un seul mot, *cantabimus.* C'est un exemple typique : des structures initialement syntactiques (assemblages de mots isolés) deviennent morphologiques. Plus tard encore, dans le Latin apparait *cantare habemus*, associant l'infinitif *cantare* et *have* [*habere*]. Ce n'est plus *être* mais *avoir*, à la première personne

du pluriel. Et de fait, quand on se tourne vers le français, on observe que le futur s'exprime à nouveau par deux mots : *nous allons chanter*. C'est *aller*, en fait, qui est devenu le verbe pour exprimer le futur.

Il y a donc bien des changements dans le temps, une évolution qui mobilise la création des mots, le recrutement de mots et leur assemblage syntaxique pour exprimer certains choses, des opérations de morphologie accompagnées d'une érosion. Ce processus est typique de la dynamique des langues et, en général, est réactivé lorsqu'un nouveau mot apparait.

On observe également que la dynamique des langues comporte une composante collective, à travers des processus de propagation, d'introduction de nouveautés qui se diffusent dans la communauté. C'est le cas pour la négation en français (cf. figure 4) ; partant du *non* du Latin, l'expression de la négation s'est d'abord transformée en *ne.* C'est typiquement un phénomène d'érosion ; mais *ne* n'exprimant pas de façon très claire la négation, un deuxième mot est apparu pour la renforcer, comme *mie*, *point* ou *pas* ; ainsi dit-on *je ne veux pas*, ce qui signifie *je ne veux (faire) un pas*. A un certain moment, il existe donc plusieurs manières de renforcer la négation. Puis l'une d'entre elles gagne la bataille, pour devenir dominante.

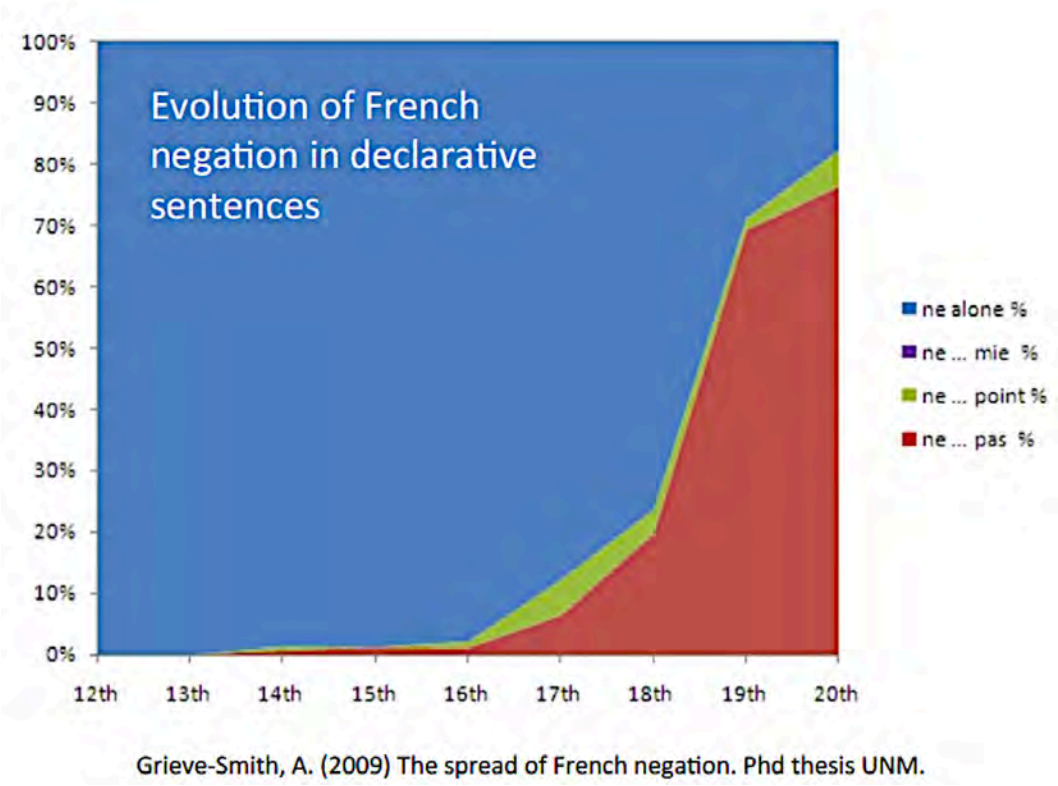



**Figure 4** Evolution de la négation dans les phrases déclaratives dans la langue française [2]

Ces exemples illustrent ce qui se passe dans ces processus d'évolution ; de nombreuses données sont disponibles sur ce thème, les livres de linguistique historique en sont remplis. Mais disposer de nombreux exemples ne fournit pas une théorie. C'est un peu comme à l'époque de Darwin : des biologistes avaient donné des descriptions, dressé des typologies, mais ce n'était pas encore la théorie de l'évolution.

## 2. Hypothèses pour une théorie de l'évolution du langage

Présentons d'abord quelques-unes de ces hypothèses, pour ensuite examiner comment elles peuvent être testées. Pour ce faire, rapprochons-nous à nouveau de la biologie. En biologie, en fait, l'explication des phénomènes concernant l'évolution mobilise plusieurs théories. La première est la théorie de Darwin sur les origines, couplant une dynamique de réplication, *replicator dynamics* – dynamique de populations d'entités capables de se copier elles-mêmes de façon plus ou moins exacte – et l'intervention d'une sélection – *natural selection*. La seconde théorie est la celle de la formation d'une hiérarchie de *niveaux*, la théorie de l'auto-organisation, de l'émergence de niveaux d'organisation supérieurs, dotés de propriétés nouvelles.

### 2.1 Dynamiques de réplication

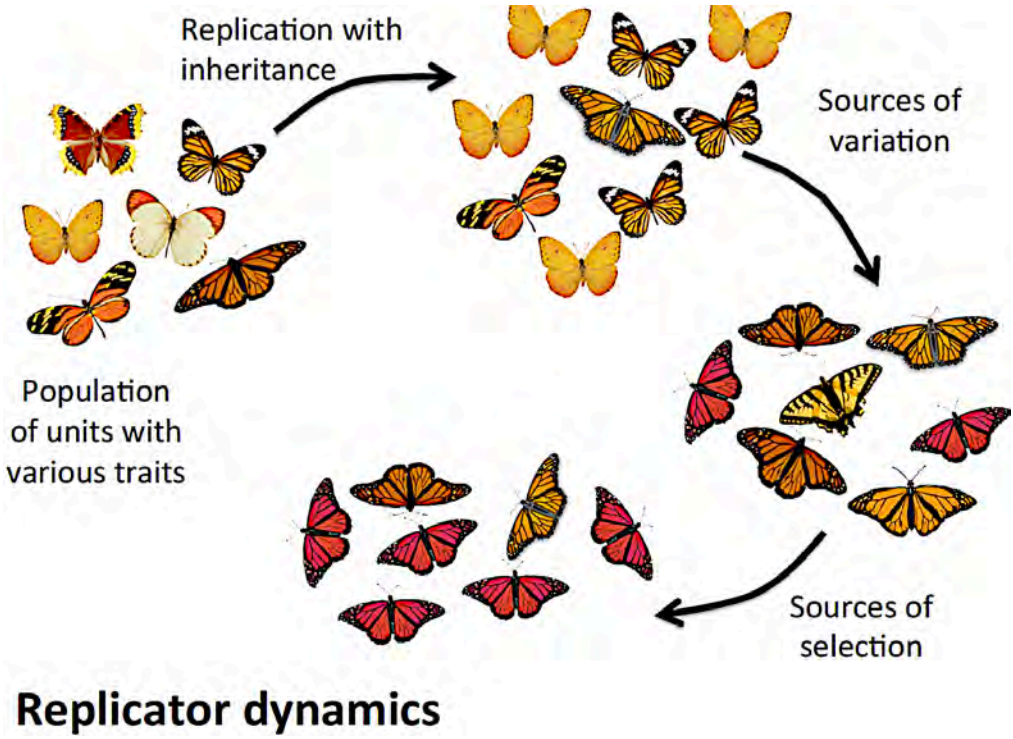



**Figure 5.** Principes de la théorie de l'évolution : dynamiques de réplication [16, planche 30].

Quelques mots sur la dynamique de réplication, *replicator dynamics*. Elle requiert d'abord une population, un ensemble d'organismes qui partagent certaines caractéristiques ; elle requiert également la disposition chez ces organismes d'une possibilité de se copier avec précision et de transmettre ainsi leurs caractéristiques à leurs copies *(inheritance)*. Elle requiert, enfin, une source de variation. Dans le processus de copie, des erreurs peuvent survenir, lesquelles peuvent s'avérer favorables, notamment lorsque des changements sont intervenus dans l'environnement. Certaines copies vont s'avérer mieux adaptées, vont avoir la possibilité de créer encore plus de copies.

### 2.2 Emergence de multiples niveaux d'organisation

Venons-en maintenant à la théorie de la formation des niveaux d'organisation, dont nous pensons qu'elle a une importance équivalente à la théorie des dynamiques de réplication. Nous allons en donner un exemple biologique, mais il en existe aussi en physique et en chimie.

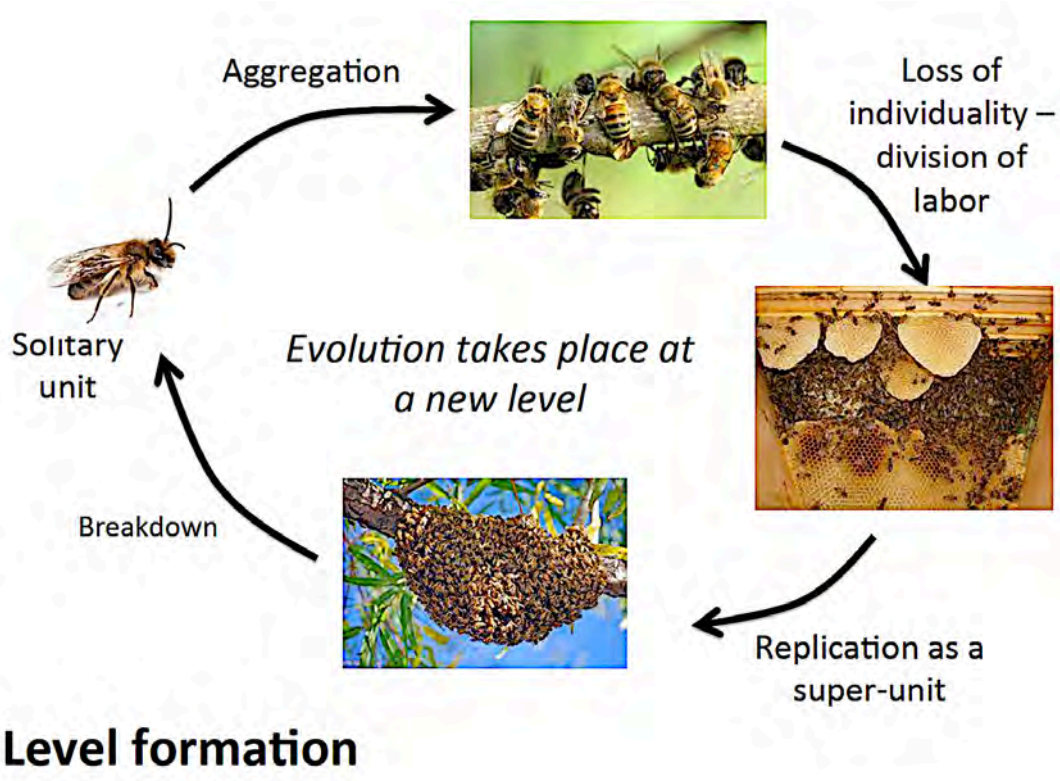



**Figure 6.** Principes de la théorie de l'évolution : formation d'une hiérarchie de niveaux. Auto-organisation [16, Planche 48].

Considérons donc des unités solitaires, des abeilles solitaires par exemple, sujettes à un processus de réplication. Mais supposons que soient sélectionnées plus particulièrement

celles qui ne vivent pas seules, le fait d'être proches des autres favorisant la survie. Après un certain temps, un processus peut se mettre en route, un processus *d'auto-organisation*, dans lequel ces organismes perdent peu à peu leur individualité ; apparaissent alors des *colonies*, avec un commencement de division du travail. La reproduction n'est plus l'affaire de chaque individu, mais devient celle d'un organisme spécialisé, la reine. Au bout du compte, émerge un niveau d'organisation supérieur, savoir celui des *ruches*, à tous égards des super-organismes, capable de se copier en tant que tels et transmettre leurs caractéristiques d'organisation. A contrario, peut survenir un processus de *breakdown* dans lequel les individus composants des super-organismes qui ont émergé redeviennent solitaires.

Ces deux théories sont les deux racines, les deux piliers de notre approche. Approche qui est en effet de les rapporter, de les transposer en linguistique. Comment faire cette transposition ?

### 2.3 De la biologie évolutive à la linguistique évolutive

En premier lieu, il nous faut un système avec une population (un ensemble dynamique d'éléments), des processus de (re)copie de ces éléments, des sources de variation et des processus de sélection.

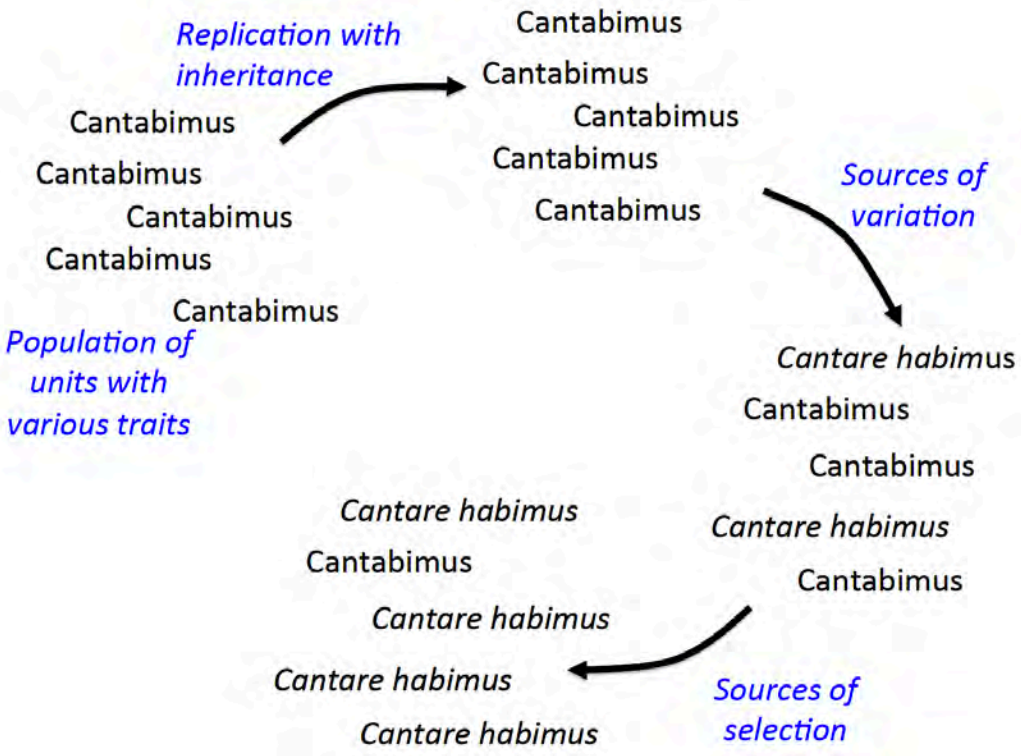



**Figure 7.** Dynamiques de réplication à l'œuvre dans l'évolution du langage [16, Planche 31].

Dans notre cas, la population est constituée d'éléments linguistiques, savoir des phonèmes, des mots, des constructions grammaticales, des éléments de sens. Par exemple, en français, le mot *un* (mot qui, notons-le, n'existe pas en latin) est caractérisé par le son nasal. Pour exister, il faut que ces éléments soient connus par les individus (d'un groupe humain) : chacun d'eux (phonèmes, mots, etc.) doit être créé puis reproduit, pour devenir partie de cette population.

Le mécanisme de copie mobilise l'apprentissage ; par exemple, l'apprentissage d'un nouveau mot par un individu et, quand un individu apprend un nouveau mot (ou une nouvelle expression), il est probable qu'il va le (la) réutiliser. Peut-être pas exactement dans le même sens mais c'est tout de même une façon de coopérer au processus de « mise au monde » et de survie des éléments linguistiques. La copie se fait lorsqu'il le réutilise effectivement et les possibles variations interviennent lors de cette réutilisation.

Venons-en à la sélection. Existe-t-elle et comment opère-t-elle ? Plusieurs critères entrent en jeu ; le premier est celui de l'utilité. Il faut que les éléments linguistiques soient utiles, savoir qu'ils soient porteurs de sens importants à exprimer dans la communauté. S'il y a des coutumes, des objets, des choses qui ne sont plus importants, le mot qui les désigne va disparaître. Le second est le succès de l'élément dans la communication ; si je demande quelque chose en utilisant un mot mais que je n'arrive pas à en communiquer le sens, ce mot va être abandonné, disparaître. Intervient également la complexité cognitive qui doit être réduite, pour faciliter l'apprentissage : si je réutilise un mot ou une structure grammaticale et qu'il ne se trouve personne capable de l'apprendre, c'est la fin, ce mot ou cette structure grammaticale ne vont pas continuer à exister. En fin de compte, pour rester vivant, le langage se doit d'être efficace et optimisé.

Il y a donc bien variation et sélection : il y a une population (on voit : *cantabimus, cantabimus, cantabimus*), les gens répètent, c'est appris par les enfants, etc. mais on commence à voir des variations et à un certain moment, en fait, on voit que certaines variations vont devenir dominantes pour la population et d'autres vont disparaître.

Considérons maintenant le deuxième pilier de la théorie : celui de la formation de niveaux – *level formation*. Certains éléments linguistiques vont perdre leur individualité, s'agréger pour former des structures plus complexes, en tenant compte de diverses limites ou contraintes. Ainsi les voyelles et consonnes se combinent-elles pour former des syllabes. Puis les syllabes se combinent pour former les mots. On voit là donc qu'il y a des niveaux de complexité dans le langage comme dans d'autres systèmes. Il y a par exemple des mots qui existent en tant que tels, de façon indépendante, avec un sens qui leur est propre ; mais après un certain temps, ils sont également répliqués – utilisés – en tant que membres d'une structure grammaticale, avec un emploi et une signification liés à cette structure. Donnons-en un exemple typique : le verbe *vouloir* ; c'est un verbe comme les autres, son emploi implique un sujet et un prédicat, emploi dans lequel il s'accorde avec le sujet : par exemple, le *s* dans *wills* marque l'accord avec un sujet à la troisième personne. Mais à une certaine époque *vouloir* devient un auxiliaire et, employé comme auxiliaire, il perd sa capacité à

s'accorder avec le sujet. Ce changement, cette mutation de son usage a commencé à se répliquer et cette diffusion a introduit dans la population des éléments linguistiques une structure de plus haut niveau, une structure grammaticale qui s'est mise à exister par elle-même.

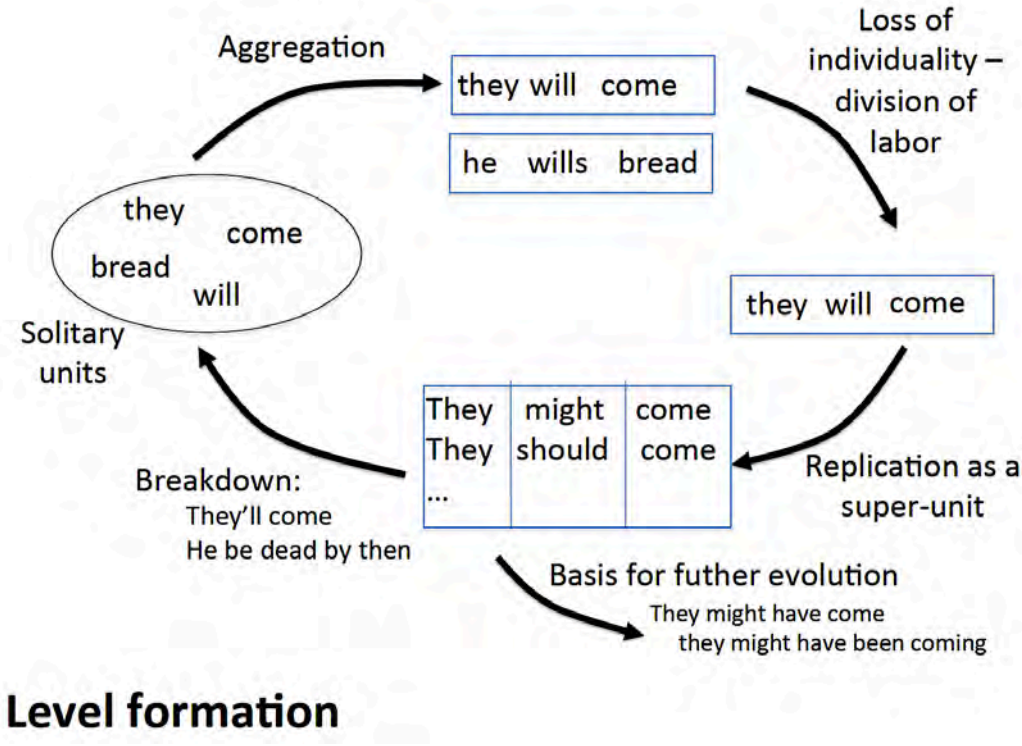



**Figure 8.** Formation de niveaux à l'œuvre dans l'évolution du langage [16, Planche 49, 15].

## 3. Des hypothèses à l'expérimentation

Pour expliquer la genèse, puis l'évolution des langages, notre hypothèse est donc qu'il faut faire appel aux théories des dynamiques de réplication et de l'émergence de multiples niveaux d'organisations. Comment – par quelles méthodes – aborder ce problème d'un point de vue scientifique ? Notre démarche, comme dans d'autres sciences, est de construire des modèles et examiner si ces modèles peuvent reproduire les faits constatés.

### 3.1 Modèles multi-agents

A la base de ces modèles, est l'idée, implicite dans les développements précédents, que l'émergence d'un langage partagé et relativement stable au sein d'un groupe naît des *interactions* entre individus et non, par exemple, de décisions prises par quelque institution. Ce parti pris affirmant le rôle central des interactions entre individus dans la structuration d'une collectivité - ségrégation, répartition de rôles, division du travail, etc. - est l'apanage

d'un courant de la sociologie, parfois nommé individualisme méthodologique. Il conduit à des modèles informatiques de processus sociaux appelé *modèles multi-agents – Agents-based models.* Ces modèles mettent en oeuvre un ensemble d'*agents* – c.-à-d. des entités artificielles – des modules de programmes informatiques, ou même dans notre cas, on va le voir, des robots – dotés de fonctions minimales et de capacités d'interagir les uns avec les autres. L'observation de ce à quoi conduit la succession d'un grand nombre d'interactions, à la fois dans l'organisation globale de l'ensemble des agents et dans l'état interne de chacun d'eux permet de tester la pertinence du modèle pour expliquer les dynamiques réelles.

Un des premiers exemples significatifs de cette approche a été celle de Thomas Schelling, prix Nobel d'économie en 1978. Thomas Schelling s'est intéressé à l'émergence d'une structuration d'une ségrégation spatiale d'une population en différents groupes séparés, par exemple dans une ville comme New-York (cf. figure 9). Son point de vue était que cette structuration n'est pas le résultat d'une planification, d'une organisation urbaine décidée par une administration. C'est pour une large part le résultat d'un processus spontané. Par le biais d'un modèle à base d'agents, il a pu proposer une explication scientifique de cette structuration. Pour ce faire, il a imaginé un monde virtuel (numérique) constitué d'une grille spatiale dont chaque cellule représente une maison, et une population d'agents, chacun d'eux étant caractérisé par un ensemble d'attributs très simples : l'origine – le pays où ils sont nés – la langue, etc.

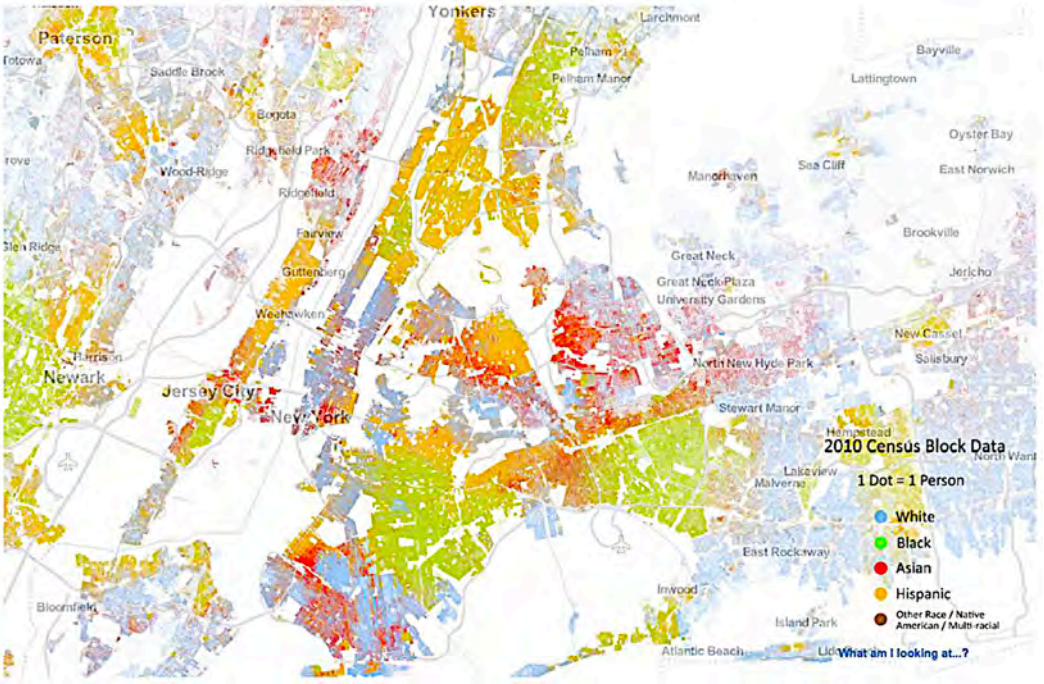



**Figure 9**. 2010 US Census New York City map, couleurs en fonction de la race [16, Pl. 15, 15].

Tous ces agents ont les mêmes règles de comportement : par exemple, habitant d'une maison, l'agent regarde quels sont ses voisins et si ces voisins partagent avec lui certaines caractéristiques – certains attributs – il est content et reste dans cette maison. Sinon il a tendance à déménager. La réflexion scientifique, la construction du modèle consiste à trouver les règles de comportements *individuels* les plus simples possibles permettant de reproduire, par simulation dans ce monde virtuel, la structuration observée dans le monde réel. On se demandera par exemple si une reproduction convaincante est assurée par la règle du choix aléatoire fixant la maison dans laquelle l'agent va déménager, par un choix au hasard dans l'ensemble des maisons libres. La figure 10 montre ainsi la configuration spatiale initiale de ce monde artificiel numérique, homogène, et la configuration finale, où apparait clairement une ségrégation par quartiers différents et séparés ; reste à examiner ensuite si, *statistiquement* (fragmentation, taille des agrégats, forme des contours, etc.), la ségrégation obtenue est similaire à la ségrégation observée.

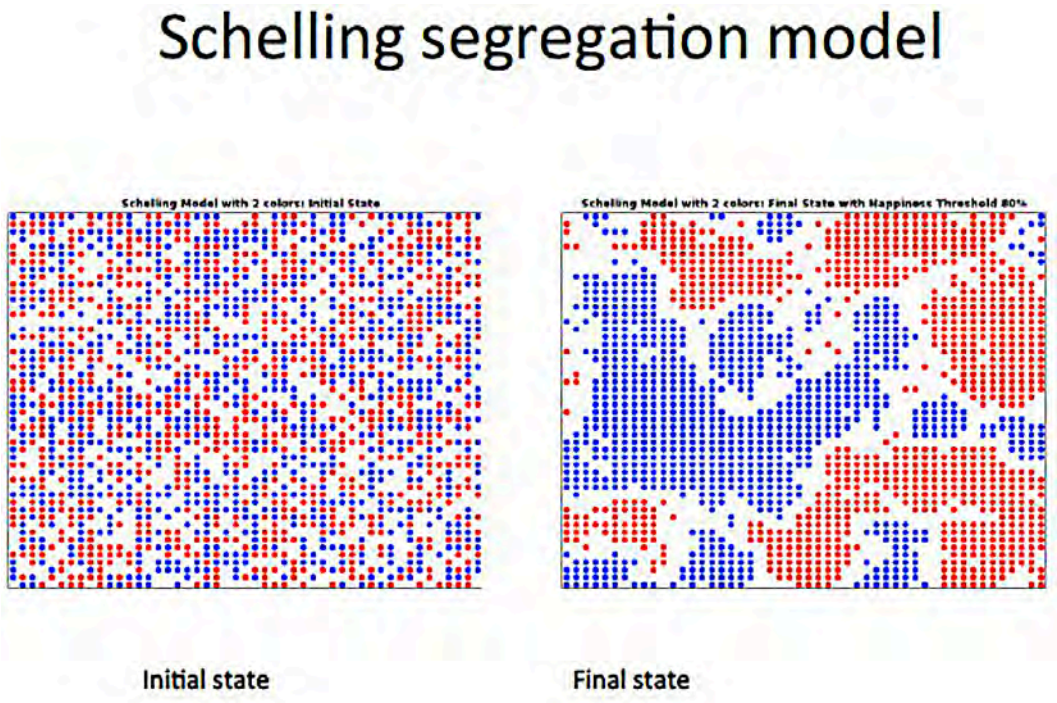



**Figure 10.** Le modèle de ségrégation de Schelling [16, Planche 17, 15].

Dans notre investigation sur la genèse d'un langage partagé par une communauté, nous avons utilisé la même approche. Nous nous sommes donnés une population « d'agents linguistiques » dotés de certaines capacités minimales, par exemple d'émettre et de recevoir des « mots » mais sans que ces mots soient partagés au départ d'un agent à un autre, ni aient pour eux une signification commune ; puis on les fait interagir entre eux. On examine alors ce qui se passe et on constate, après un certain temps, l'émergence, au sein de cette

population, d'un langage commun. Cette approche, initiée en 1996, a rencontré à ses débuts – et rencontre toujours – un certain scepticisme. Mais nous pensons qu'elle a fait ses preuves et a de grandes potentialités. Les agents mis en œuvre sont plus complexes que ceux de Schelling. Initialement ces agents étaient de simples programmes interagissant entre eux au sein d'un ordinateur – c'était évidemment plus facile et plus économique –

mais très vite nous avons investi une approche physique, chaque agent étant incarné par un robot, un vrai robot. Si en effet il peut être utile de faire des simulations numériques, il s'avère nécessaire, à un certain moment, d'expérimenter dans la réalité physique : les mots prennent en effet d'abord leur sens dans cette réalité physique, ils parlent de cette réalité, en mobilisant les processus de perception : on regarde tel objet, on parle et on dit : une table. C'est ce qu'il faut expliquer : comment, à partir de la perception de cet objet physique, une table, en est-t-on arrivé à le désigner par un mot comme « la table ». En montant des expériences avec des robots, on est sûr - du moins est-ce notre hypothèse - d'attaquer les problèmes d'une façon similaire à la manière dont le font les cerveaux.

### 3.2 Modèles d'agents et évolution du langage

Quelques exemples de notre travail sont illustrés par des vidéos. Une première vidéo montre ainsi un jeu très simple de mouvement : deux robots apprennent à associer certains mouvements – ouvrir la main… – à des mots précis – formés de suites de syllabes arbitraires – sur lesquels ils vont progressivement s'accorder : un agent demande à un autre agent, en utilisant un mot de son invention – une suite de syllabe arbitraire –, de faire un certain mouvement, par exemple d'ouvrir la main. L'autre agent, face à ce mot qu'il entend pour la première fois, dit « *je ne sais pas* » ; le premier agent lui montre alors le mouvement. Et le second apprend, voire d'autres agents qui ont également entendu le mot prononcé et pu voir le mouvement. A travers une série d'interactions du même type, se stabilise un vocabulaire partagé désignant des mouvements possibles. Une autre vidéo montre des robots entrant dans un environnement formé de différents objets (cf. figure 11). Le but est d'amener progressivement ces robots à communiquer par un mot une demande d'attention à tel ou tel de ces objets, une bouteille d'eau par exemple, et bien sûr que cette communication soit efficace, savoir que les robots récepteurs du message prêtent attention à l'objet concerné. Donc arriver à une situation courante chez les humains : si on me dit : veux-tu me donner du papier pour que j'écrive sur …, je vais aller chercher du papier et le donner.

Sans entrer dans toutes les techniques mises en œuvre pour réaliser ces expériences, techniques qui ressortent de l'Intelligence Artificielle, analysons-en brièvement quelques unes.

### 3.3 Distinguer et nommer des couleurs

Une de ces expériences concerne l'émergence d'un vocabulaire pour les couleurs (cf. [14]). Cette expérience mobilise des agents, de fait des robots, environnés d'un ensemble d'objets colorés de différentes couleurs et nuances. Au départ, ces agents n'ont aucun vocabulaire

pour désigner des couleurs ; mais à la fin du processus, ils en partagent un. Un tel vocabulaire en a émergé, similaire à celui dont nous disposons.

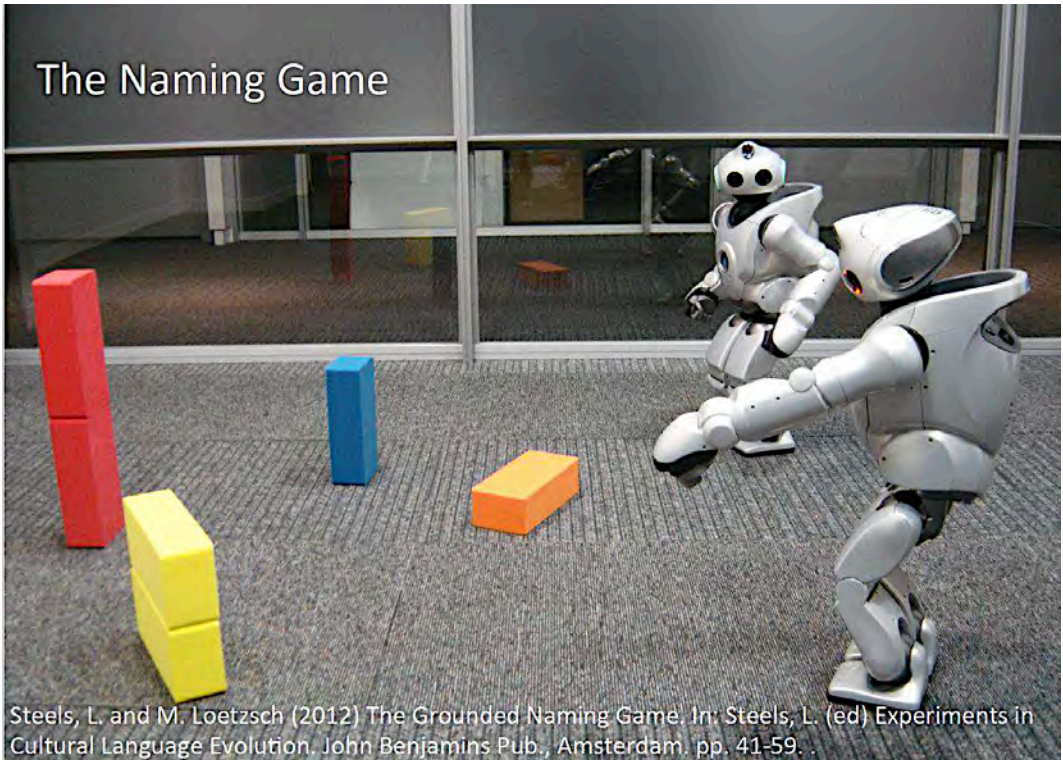



**Figure 11.** The naming game [3].

De quoi faut-il doter ces agents pour que cette expérience aboutisse ? Il faut d'abord qu'ils aient une capacité de *percevoir*, donc qu'ils disposent chacun d'une caméra. Il faut qu'ils soient ensuite armés pour projeter chaque objet (ou partie d'objet) perçu dans un certain espace abstrait – un espace de configuration, un *feature space* –, dont chaque dimension représente une caractéristique mesurée sur cet objet – plus spécifiquement dans ces travaux, pour être proche de l'œil humain, une projection sur le diagramme tridimensionnel CIE/Lab (un axe de luminence, allant du noir au blanc et deux axes de couleurs allant respectivement du vert au rouge et du bleu au jaune, voir figure 12). Il faut également qu'ils soient équipés pour être à même de délimiter, dans cet espace de configuration, des régions distinctes, qui constitueront autant de catégories – ici des catégories de couleurs. Cette capacité de catégoriser conduit les agents, à chaque étape de ce « jeu de langage », à un ensemble de catégories provisoires et propres à chacun d'eux. Ces agents sont en outre capables d'associer librement à ces catégories provisoires des *mots* (des suites arbitraires de syllabes).

L'expérience consiste en une succession d'interactions entre deux robots. Lors de chacun de ces *jeux*, un robot *speaker* s'adresse à un robot *hearer* en lui proposant un mot et lui demande de *deviner*, c.à.d. de désigner par un geste un objet qui ressort de la catégorie que

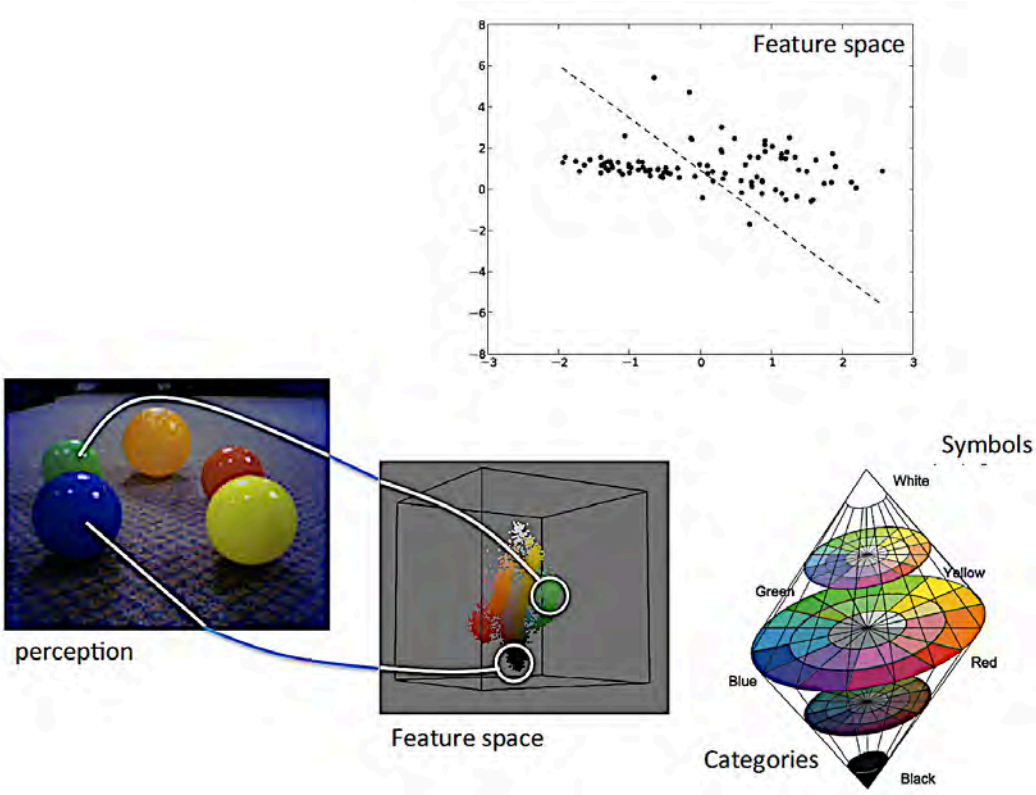



**Figure 12.** Perception, Feature space, Catégories, Symboles [16, Planche 34, 15].

le *speaker* avait associée à ce mot. Si le *hearer* devine bien, la communication entre les deux est un succès et l'association mot-catégorie est renforcée dans le cerveau des deux robots, ainsi que validité de leur catégorisation respective pour l'objet désigné[2]. Parallèlement d'autres associations voient leur force diminuée (processus d'inhibition). Au début du processus, les agents possèdent toutes les capacités mais n'ont aucun vocabulaire ni aucun répertoire de catégories. L'objectif est qu'au terme du processus – en fait un processus d'apprentissage coopératif – aient émergé parallèlement des catégories – en l'occurrence de couleurs – et des mots pour les désigner, communs à tous les agents.

[2] D'autres type d'interactions peuvent bien sûr être mises en oeuvre, comme par exemple celles où un robot attire l'attention d'un autre robot en lui désignant par un geste et en prononçant le mot correspondant à la couleur qu'il lui attribut.

Des expériences ont donc été conduites avec une population de cinq ou dix robots : les espaces de configuration, les catégorisations formées et les associations avec les mots proposés sont accessibles à l'expérimentateur ; ce dernier peut voir en quelque sorte ce qui se passe dans le « cerveau » de chaque robot et se faire une image exacte de l'évolution du langage à chaque étape. Au début, les mots proposés sont complètement différents d'un agent à l'autre, et les succès dans la communication sont rares, dus au hasard. Puis des mots liés à des succès de communication sont retenus, repris et deviennent partagés, émerge ainsi progressivement un lexique commun qui se stabilise à un niveau relativement riche. On constate parallèlement l'homogénéisation progressive des catégories de couleurs désignées par ces mots chez ces agents. La figure 13 illustre bien ce processus. La partie gauche représente la situation au bout de 1000 jeux. Trois mots communs ont émergé mais les *prototypes de couleurs* qu'ils désignent chez chacun des cinq robots sont encore assez différents aux yeux d'un humain. Après 2500 jeux (partie droite de la figure), il y a certes encore quelques différences, par exemple sur la nuance de couleur désignée par le mot *vamasi*, mais il est clair que la variance de ces différentes interprétations est assez faible.

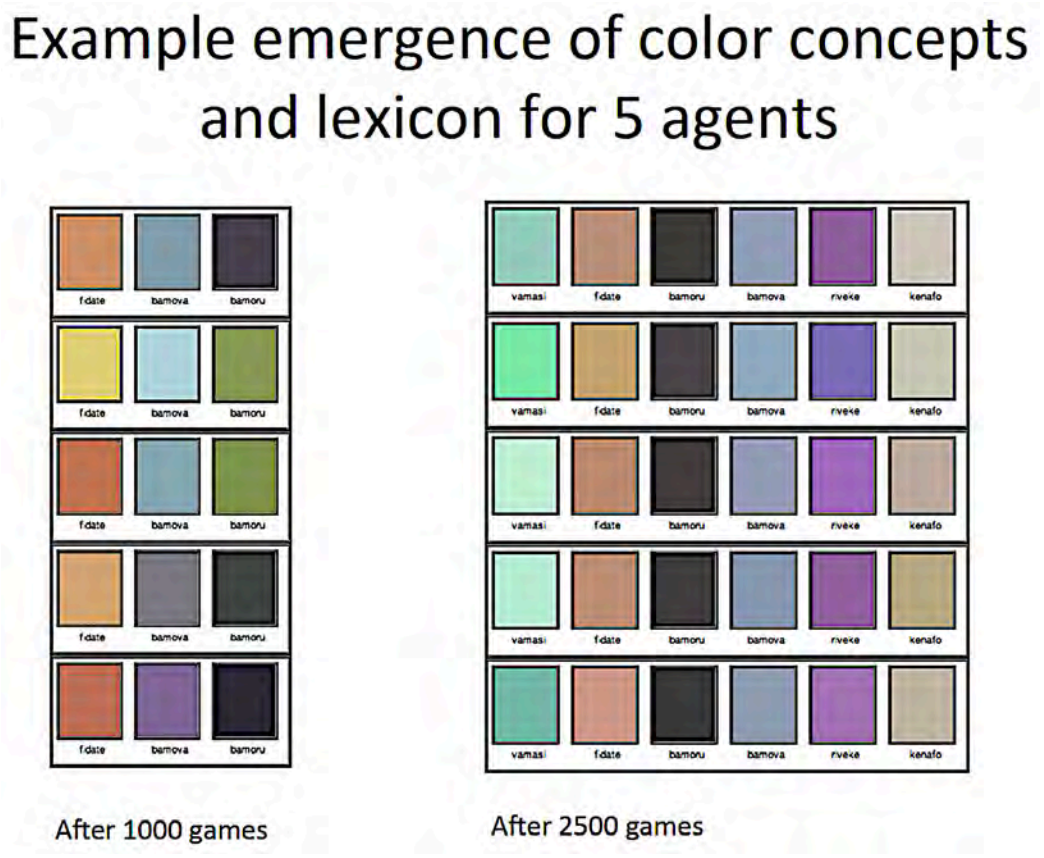



**Figure 13.** Emergence de prototypes de couleurs et d'un lexique associé pour 5 agents [16, Planche 36, 15].

Lorsque le système est stabilisé, il est possible d'en modifier le cadre : en introduisant de nouveaux objets, et/ou de nouveaux robots. On constate que le système peut alors se réorganiser (par exemple par l'apparition d'un mot désignant une couleur supplémentaire), ou encore par un remaniement plus profond.

Bien d'autres thèmes ont été ainsi abordés, avec des méthodes similaires ; jeux de langage où les robots doivent apprendre à donner des noms aux différents objets placés dans leur environnement, ou acquérir un vocabulaire partagé désignant les mouvements qu'ils peuvent faire avec leur corps. Ce dernier apprentissage mobilise simultanément trois types de représentations : la perception interne du mouvement - *proprioception* chez les humains - la *vision* de ce même mouvement par le biais des caméras, et enfin les mots qui sont associés à ces mouvements au cours du jeu. Ont également fait l'objet d'expériences les notions relatives à l'espace, à l'orientation spatiale (cf. figure 15) et au vocabulaire associé (notions d'*arrière*, d'*avant*, de *gauche*, de *droite*). Mais en rester simplement à l'apprentissage d'un lexique est insuffisant. Les linguistes nous le disent : et la grammaire, qui parait spécifique, ou du moins particulièrement structurée dans le langage humain ?

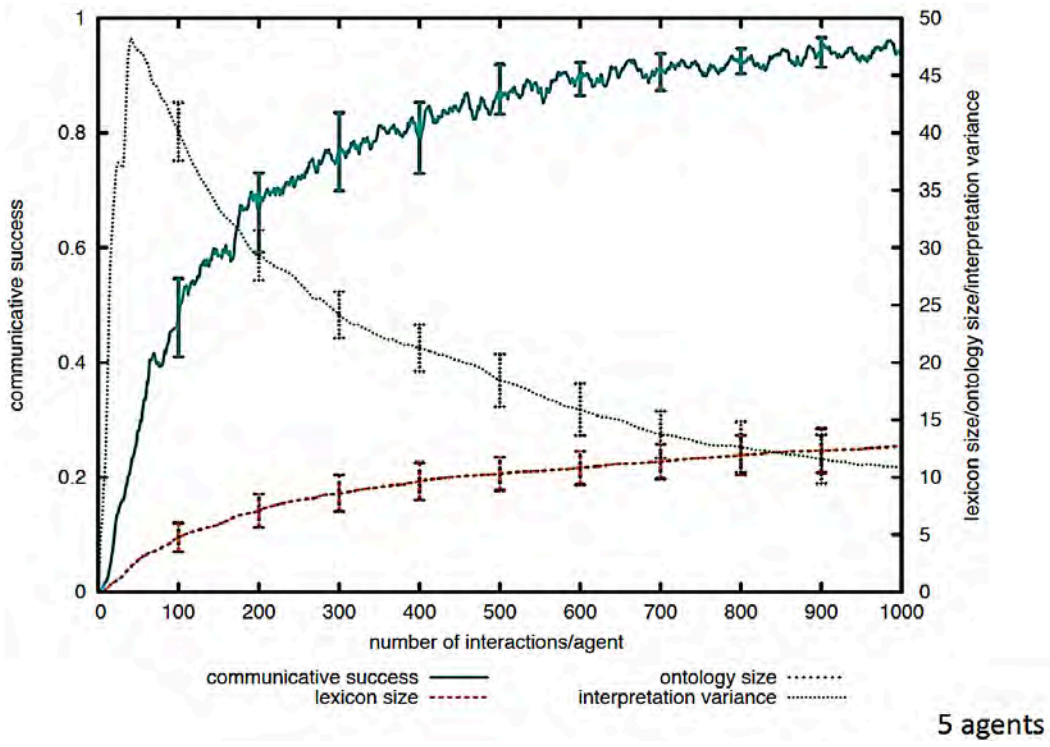



**Figure 14.** Emergence d'un vocabulaire pour les couleurs [16, Planche 38, 15].

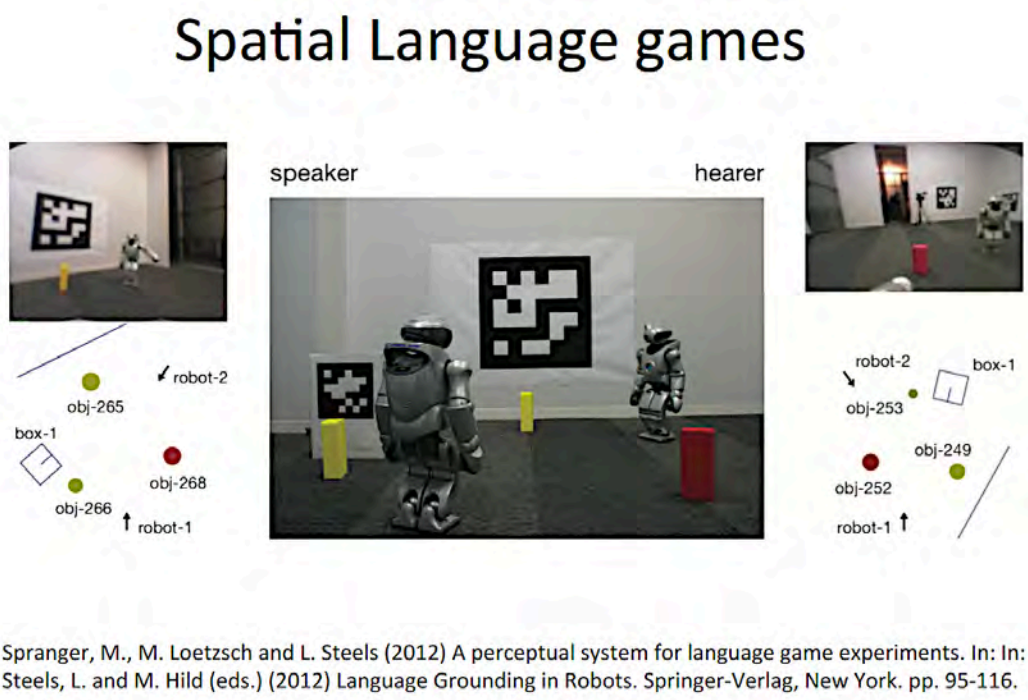



**Figure 15.** Spatial Language games [4].

### 3.4 Aller plus loin : comprendre la formation des structures grammaticales

L'apparition de grammaires complexes, puis leurs évolutions, peuvent-elles être abordées par des démarches similaires ? Cette question nous occupe depuis une dizaine d'années. Plusieurs expériences, dont les résultats ont été publiés en 2013, ont ainsi été menées sur l'émergence des *agreements*, des *accords* – accords entre sujet et verbe, entre article, adjectif et nom (exemples : *une belle fille* ou encore, en latin, *illarum bonarum feminarum*, en français *de ces deux bonnes femmes*, cumulant l'accord sur le genre et sur le nombre). Nous avons également travaillé sur l'apparition des *quantifiers* (few, several, many). Nous nous sommes enfin attaqués au problème controversé de la genèse de structures hiérarchiques (récursives) dans les phrases.

Selon nous, la racine de toutes ces structures réside dans une recherche d'efficacité du langage, comme nous l'avons déjà souligné plus avant. Cette recherche vise à réduire ou même supprimer les ambiguïtés d'interprétations (volet sémantique), à réduire également les difficultés de son emploi (volet articulatoire et phonétique, lexical, syntaxique). Concernant le volet sémantique, imaginons que pour décrire une certaine scène du monde réel, un *speaker* utilise un certain ensemble de mots. De multiples interprétations peuvent correspondre à cet ensemble dans l'esprit du *hearer*, s'il doit se contenter de cette liste de mots fournie « en vrac », sans une information complémentaire. Bien sûr, sa bonne

connaissance du contexte pourra lui faire deviner la bonne interprétation et lui permettre de se représenter correctement la scène en question. Mais cela ne sera pas toujours le cas. Pour reprendre un exemple donné dans une de nos publications, l'assemblage des quatre mots *brown, good, book, idea* reste ambigüe : certes une idée peut difficilement être qualifiée de « brune » mais le qualificatif de *good* peut être attribué aussi bien à *book* qu'à *idea*. La solution, pour éviter cette possible « explosion combinatoire » des interprétations, est de fournir des informations complémentaires encapsulées dans la forme des mots et/ou leur place et organisation dans la phrase. Et, de ce point de vue, les linguistes ont observé une grande diversité des stratégies utilisées par les différentes langues. Certaines privilégieront l'ordre des mots, d'autres les systèmes d'accord ou de cas, etc.

### *3.4.1 Un exemple, l'émergence des accords*

Les travaux concernant l'émergence des accords, comme par exemple le travail de Katrien Beuls et Luc Steels publié en 2013 [5], illustrent bien cette approche (cf. figure 16). Ils mettent en œuvre une population d'agents et un environnement constitué d'un ensemble d'objets dont chacun est décrit par les valeurs de certaines propriétés telles que la couleur, la taille, la texture. Le vocabulaire pour désigner ces valeurs de propriétés – avec des mots comme *green, small, smooth* est déjà en place et partagé par tous les agents, de même que leur capacité, à l'aide de ce vocabulaire, à décrire et reconnaître chaque objet. La tâche assignée à un *hearer*, dans un jeu, est d'identifier précisément un *ensemble* d'objets décrits par le *speaker* à l'aide de mots qualifiant leurs propriétés. Les agents sont dotés de la capacité de déterminer les différentes solutions possibles de ce problème d'identification d'un ensemble d'objets, partie d'un ensemble plus vaste. Mais il est clair qu'une simple liste de mots, sans convention d'aucune sorte, par exemple la liste *green, smooth, small, medium, red*, est insuffisante. Le *hearer* peut en effet deviner qu'il y a peut-être deux objets, mais si tel est le cas, l'objet *red* est-il *medium* ou *small* ? Dans les travaux mentionnés plusieurs solutions sont expérimentées pour lever de telles ambiguïtés : la première, celle des *marqueurs formels*, est de permettre aux *speakers* d'associer aux mots, caractérisant les propriétés des objets de l'ensemble qu'ils veulent désigner, des sortes de stickers ou *marqueurs* sans signification autre que celle d'indiquer que tel et tel mot se réfère au *même* objet, sans toutefois dire lequel, cette association s'exécutant par l'intermédiaire d'un suffixe. Par exemple au lieu de dire *green, smooth, small, medium, red,* quelque chose comme *greenaa, smoothxy, smallaa, mediumxy, redxy*. Le *hearer*, disposant du vocabulaire partagé des propriétés est en mesure de repérer les suffixes *aa, xy* comme des marqueurs – lesquels peuvent déjà ou non faire partie de sa propre liste de marqueurs – et ainsi pointer les objets qu'il pense avoir été désignés par le *speaker*. Il n'y a pas d'autorité centrale régissant une liste de marqueurs à utiliser : chaque agent speaker est a priori libre d'en inventer de nouveaux à chaque interaction ou de réutiliser ceux déjà entendus. Aussi l'un des objectifs de telles expériences est de tester différentes méthodes à mettre en œuvre lors des interactions pour stabiliser une liste partagée de marqueurs de taille raisonnable. On observe cependant que dans nombre de langues humaines, les marqueurs sont le plus souvent choisis parmi des mots ayant déjà un sens et non de simples chaines de caractères arbitraires, sans signification. Par exemple, un marqueur sera lié au caractère d'objet

inanimé ; ce ne sont plus des *formal markers*, mais des *meaning markers.* Les travaux déjà cités étendent donc leur approche au cas des marqueurs porteurs d'un sens, en exploitant un formalisme combinant le sens du marqueur et le sens du mot associé pour lever les ambiguïtés.

Emergence of grammatical agreement systems

*ki-kapu* *ki-kubwa* *ki-maja*
ki.basket ki.large ki.one
'One large basket' [Swahili]

*ill-arum* *du-arum* *bon-arum* *femin-arum*
those.arum two.arum good.arum women.arum
'of those two good women' [Latin]

Beuls, K., & Steels, L. (2013). Agent-Based Models of Strategies for the Emergence and Evolution of Grammatical Agreement. PLOS ONE, 8(3), e58960. http://dx.plos.org/10.1371/journal.pone.0058960

**Figure 16.** Emergence de systèmes d'accords grammaticaux [5].

D'autres expériences sont menées, avec les mêmes approches – jeux de langage entre agents – pour rendre compte des processus de généralisation et d'érosion de formes observés dans les langues humaines : au cours de ces processus, les marqueurs abandonnent peu à peu une partie de leur forme, leur indépendance et leur sens, pour n'être à la fin plus rien qu'un son (cf. figure 17).

*3.4.2 Autres travaux : des adjectifs aux quantifieurs, structures de phrases récursives*

D'autres travaux montrent comment, toujours à travers des jeux de langages, des robots commencent à former des quantifieurs [8]. Donnons-en quelques détails. Pour mener ces expériences sur l'émergence d'adjectifs – tels *some, many* – exprimant une certaine quantité floue, il a d'abord fallu modéliser leur sémantique, en lien avec le décompte réel d'objets pointés par un agent dans une scène. Cette sémantique s'exprime par une sorte de *nombre flou*, en fait une forme de distribution de probabilité sur un certain intervalle de nombres cardinaux, par exemple de 0 à 9 dans une première approche.

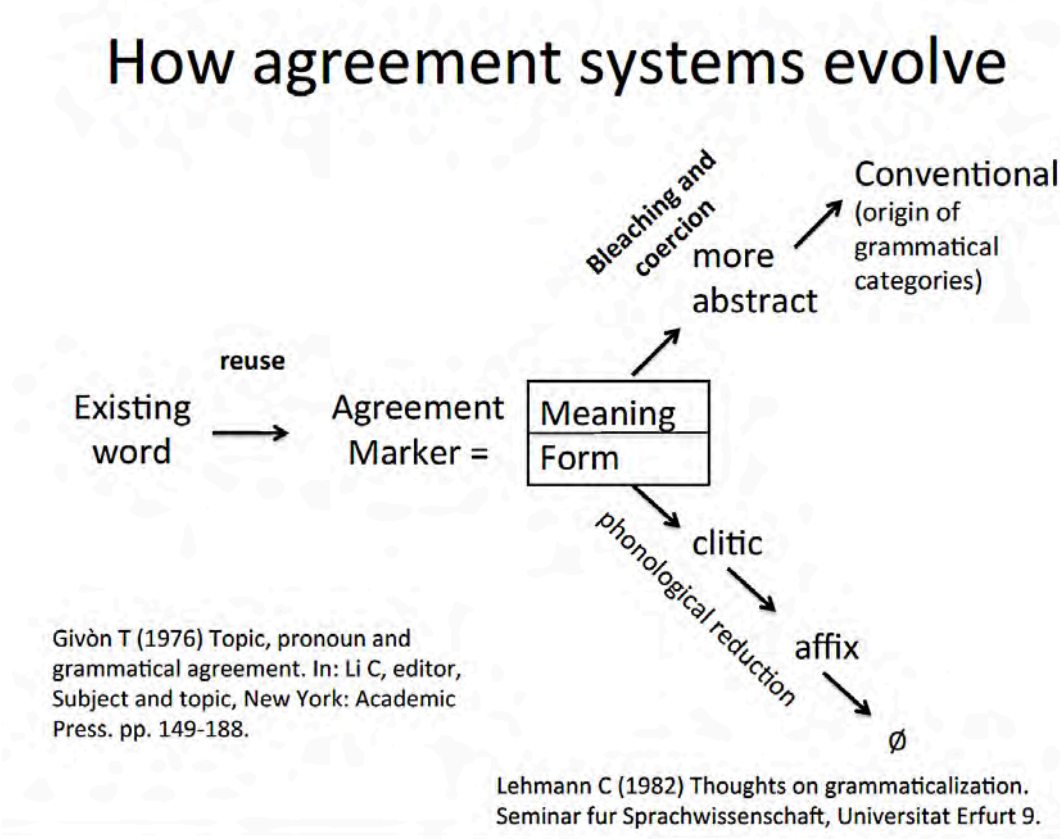



**Figure 17.** Comment évoluent les systèmes de l'accord [6, 7].

Les auteurs ont d'abord étudié le fonctionnement de la communication, lorsque les deux quantifieurs (*some* et *many*) et leur sémantique étaient déjà en place chez les robots. *Un jeu* consiste à mettre deux robots en présence de deux scènes, différenciées entre elles par le nombre d'objets qui y figurent. Le *speaker* est ainsi à même de choisir l'une des deux scènes puis de l'évaluer pour fournir, afin de la désigner au *hearer*[3], une expression utilisant le quantifieur qui lui semble le mieux approprié. Et chaque *hearer* est à même d'interpréter le quantifieur entendu et en conséquence de pointer vers la scène qui lui semble correspondre. Les processus d'élaboration puis d'interprétation de l'expression étant par nature probabilistes, le succès de la communication n'est pas assuré et il dépend bien sûr du contraste entre les scènes présentées ; mais ce taux de succès (évalué sur un ensemble de jeux successifs) sera acceptable si la sémantique des deux quantifieurs est bien adaptée à la diversité des scènes mobilisées dans les expériences.

[3] Dans une scène comprenant, par exemple, deux paquets d'objets de cardinalité différente.

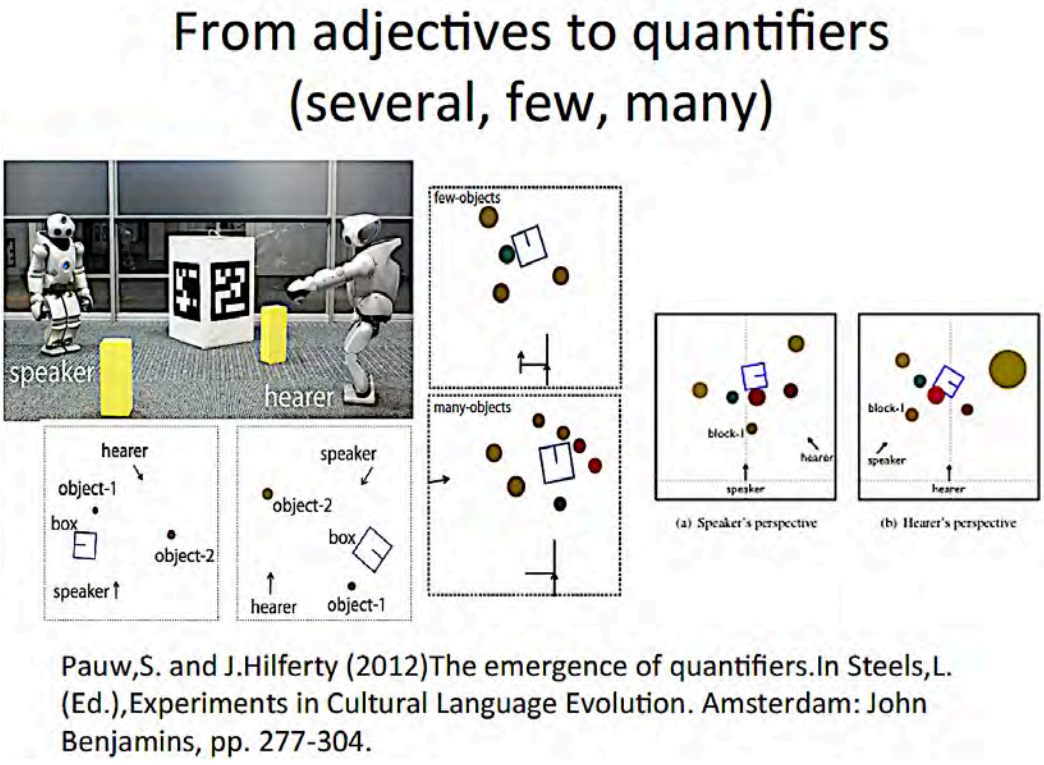



**Figure 18.** L'émergence des quantifiers [8].

Deux autres séries d'expériences ont ensuite été menées pour étudier comment cette adaptation peut se mettre en place. Dans la première, l'un des agents joue le rôle de *tuteur* ; il est le seul à être équipé d'un lexique de quantifieurs et de leur sémantique. L'autre agent *l'élève* – n'en dispose pas, mais dispose de deux opérateurs d'apprentissage. Le premier est mis en œuvre lors de la première audition d'un quantifieur (l'élève entend par exemple pour la première fois le mot *many*) ; l'élève associe alors à ce mot la cardinalité *exacte* de la scène désignée (par exemple 4). Le second opérateur permet ensuite à l'élève d'élargir la distribution de probabilité du quantifieur à d'autres cardinalités proches et de se rapprocher ainsi peu à peu de celle attribuée à ce même quantifieur par le tuteur. Des heuristiques de renforcement et d'inhibition permettent de contrôler ce processus d'apprentissage sur l'ensemble des quantifieurs utilisés par le tuteur et de le faire converger. La seconde série d'expériences est encore plus ambitieuse, puisqu'elle a pour but, comme dans le cas des couleurs, de faire émerger la sémantique – savoir des notions comme *peu, quelques, beaucoup*, voire plus différenciées encore – en même temps que le vocabulaire correspondant, l'équivalent des *few, some, many*, … Pour ce faire, les robots *speaker*, face à une scène pour laquelle aucun des quantifieurs déjà en place ne leur semble adapté, ont la capacité, soit de modifier la sémantique de l'un de ces quantifieurs, soit d'en inventer un

nouveau, lequel viendra alors se rajouter à la population existante. Ici encore, des contraintes de mémoire et diverses heuristiques assurent la convergence du processus. Enfin les travaux cités abordent la question de la *relativité* de ces quantifieurs, leur sens dépendant évidement du contexte ; une extension simple du formalisme adopté pour la version *absolue* du sens des quantifieurs *some, many,* etc. dans les expériences décrites précédemment permet de conserver la même démarche expérimentale.

On a également beaucoup travaillé sur l'émergence de *structures de phrases,* par exemple les phrases nominales, et sur le phénomène qu'on appelle *récursion,* une phrase nominale pouvant être partie d'une autre phrase nominale. Les linguistes en ont beaucoup discuté. A nouveau [9,10], on a fait des expériences pour trouver des petites règles qu'on donne à chaque agent. Et voir s'ils sont capables d'utiliser ces règles pour faire émerger de tels systèmes grammaticaux. On peut répéter ces expériences avec différents paramètres, voir les variations de dynamique qui en résultent, et on a développé des théories mathématiques pour déterminer quel est le temps nécessaire pour arriver à convergence et stabilisation, pour une population de taille donnée.

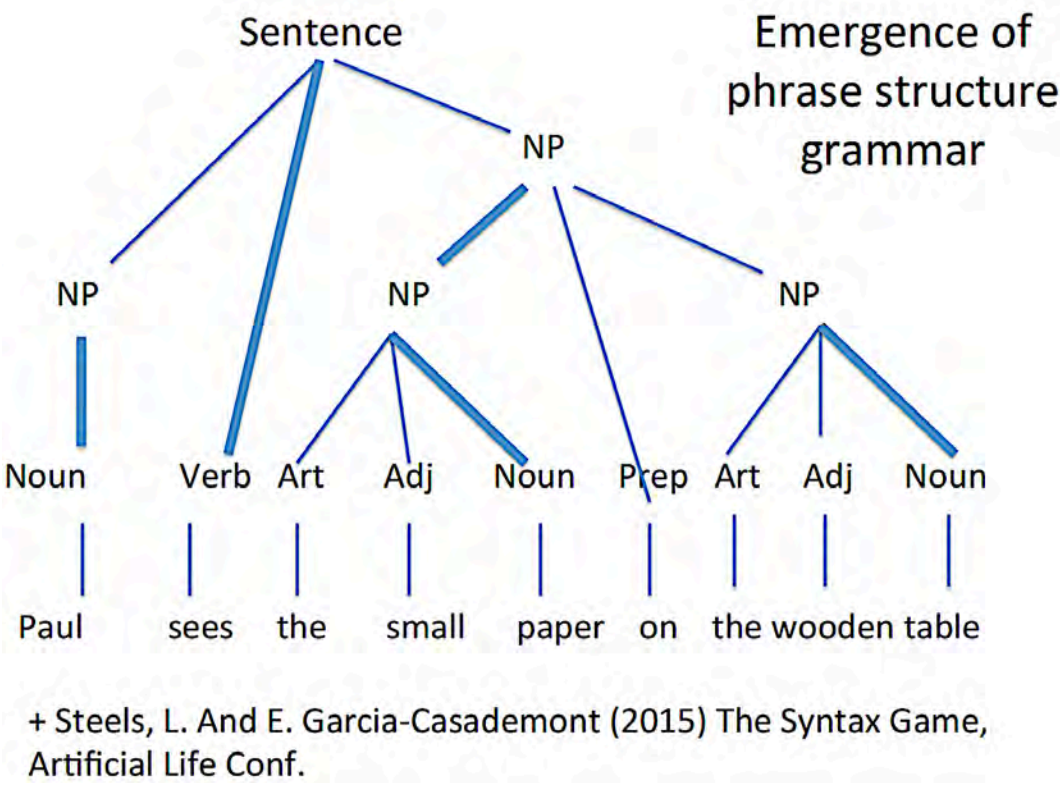



**Figure 19.** Emergence de la grammaire à structure de phrases [10]

## 4. En conclusion

Les thèses présentées dans ce chapitre sont loin de recueillir l'unanimité. Elles font à vrai dire l'objet de grands débats dans la communauté linguistique. Beaucoup de linguistes, et non des moindres, s'y opposent. Ainsi Berwick et Chomsky, dans un ouvrage paru en 2016, écrivent-ils « Languages change, but they do not evolve. It is unhelpful to suggest that languages have evolved by biological and non-biological evolution (…) The latter is not evolution at all. » Alors que de notre coté nous écrivions la même année que « Language is a culturally evolving complex adaptative system recruiting available cognitive capacities. » Que recouvre exactement cette opposition ?

Le point de vue de Chomsky et d'autres chercheurs comme Pinker est que la structure du langage – la grammaire universelle – est d'origine génétique, elle est liée au génome humain, déterminée par ce dernier. La génétique fixe complètement les modalités de son acquisition. En conséquence, les changements culturels qui affectent les langues ne sauraient être que superficiels, sans importance ; leur étude ne relève pas de la discipline linguistique, dont l'objet est la mise en évidence de cette structure par delà la diversité dans laquelle elle s'exprime.

Nous affirmons, quant à nous, que les langues sont des systèmes culturels qui changent en permanence, de façon continue, sous l'effet de dynamiques de réplication, de formation de hiérarchies de niveaux, de nature similaire à celles à l'œuvre dans l'évolution des espèces. De tels changements affectent ces langues en profondeur, sous l'effet d'innovations, de sélections et de disparitions progressives, qui interviennent constamment. Pour nous par exemple, le fait que la catégorie des articles n'est pas présente, puis apparait à un moment donné de l'histoire d'une langue (ainsi dans le passage du latin au français) signifie que les catégories syntaxiques ne sont pas innées, elles ne sont pas déjà en place dans le cerveau…

Il est clair qu'il faut un cerveau doté de fonctions capables de supporter tous ces processus. Capables, par exemple, de détecter des structures syntaxiques. Mais nous pensons que ces fonctions ne sont pas nécessairement spécifiques au langage ; le langage certes les utilise mais ces mêmes fonctions peuvent servir à d'autres tâches, comme la reconnaissance de formes, etc. Et il n'est pas exclu qu'à l'inverse, le langage ayant atteint un stade donné contraigne le cerveau à s'adapter sur certains niveaux de son organisation ou de ses fonctions, dans une interdépendance mutuelle.

Tout ce débat n'est pas sans rapport avec celui concernant la date à laquelle le langage est apparu dans l'histoire humaine ; longtemps on a pensé que le langage était apparu tardivement, il y a probablement environ quarante mille ans, au moment de la « révolution symbolique », avec l'apparition des peintures rupestres. Puis cette date a commencé à reculer car on a trouvé des manifestations symboliques datant de plus de 100.000 ans. Actuellement un linguiste reconnu comme K. David Harrison pense que le langage est apparu plus tôt encore, avant la séparation entre Néanderthaliens et Humains actuels, car en effet nombre de chercheurs pensent que les premiers disposaient également d'un

langage… Cette ancienneté ne plaide-t-elle pas pour la réalité d'une évolution du langage sur l'évolution biologique, et plus particulièrement peut-être, sur l'évolution du système vocal ?

**Références**


[1] Agent-based models for the emergence and evolution of grammar. Phil. Trans. R. Soc. B 371: 20150447.
[2] Grieve-Smith, A. (2009) The spread of French negation. PhD thesis UNM.
[3] Steels, L. and M. Loetzsch (2012) The Grounded Naming Game. In: Steels, L. (ed.) Experiments in Cultural Language Evolution. John Benjamins Pub., Amsterdam, pp. 41--59.
[4] Spranger, M., M. Loetzsch and L. Steels (2012) A perceptual system for language Game experiments. In: Steels, L. and M. Hild (eds.) (2012) Language Grounding In Robots. Springer-Verlag, New York, pp. 95--116.
[5] Beuls, K., & Steels, L. (2013) Agent-Based Models of Strategies for the Emergence and Evolution of Grammatical Agreement. PLOS ONE, 8(3), e58960. http://dx.plos.org/10.1371/journal.pone.0058960.
[6] Givòn T. (1976) Topic, pronoun and grammatical agreement. In: Li C. (ed.) Subject and topic, New York, Academic Press, pp. 149—188.
[7] Lehmann C. (1982) Thoughts on grammaticalization. Seminar für Sprachwissenschan, Universität Erfurt.
[8] Pauw, S. and J. Hilferty (2012) In Steels, L. (ed.) Experiments in Cultural Language Evolution. Amsterdam: John Benjamins, pp. 277--304.
[9] Steels, L. and E. Garcia-Casademont (2015) How to play the Syntax Game. In: Andrews, P. et al. (2015) Proceedings of the European Conference on Artificial Life 2015. The MIT Press, Cambridge Ma pp. 479--486.
[10] Steels, L. and E. Garcia-Casademont (2015) The Syntax Game, Artificial Life Conference.
[11] Pauw, S. and J. Hilferty (2012) The emergence of quantifiers. In Steels, L. (ed.) Experiments in Cultural Language Evolution. Amsterdam: John Benjamins, pp. 277--304.
[12] Berwick, R. C. and N. Chomsky (2016) *Why Only Us: Language and evolution.* Cambridge, MA: MIT Press. p. 52.
[13] Steels, L. (2016) Language evolution or language change? Journal of Neurolinguistics. Vol 52.
[14] Steels, L. and Belpaeme, T. Coordinating perceptually grounded categories through language : A case study for color. Behavioral and Brains Sciences, Sept. 2005.
[15] Steels,L. How can we explain the (culturel) evolution of language, And how technology can help finding and validating answers. Exposé présenté à Paris, le 12 juin 2017, devan l'AEIS.
[16] Steels,L. Intelligence artificielle et modèles théoriques de l'origine du langage ; Conférence présentée le 8 janvier 2018, Collège de France.

# 10

# Les machines « pensantes » Un panorama de l'intelligence artificielle

**Jean-Paul Haton**

Université de Lorraine
Reconnaissance des Formes en IA
LORIA/INRIA/NANCY

## Abstract

Jean-Paul Haton first traces the history of Artificial Intelligence; he sets out the main types of problem that AI is tackling and details the different lines of work that have been implemented in this discipline: symbolic, neuro-mimetic, probabilistic and statistical approaches. The learning issues and methods are then tackled: symbolic learning versus digital learning, supervised learning versus non-supervised learning, reinforcement learning. Then there is an overview of deep neural networks: hardware and software conditions that have allowed their emergence and current performance, different types of networks used, their limits and areas of application. After evoking of the major private or institutional players concerned, Jean-Paul Haton ends by evoking current trends, which cannot be reduced to deep neural networks, in the direction of Strong Artificial Intelligence. As well as the ethical, societal, legal problems, which the progress already in place or expected does not fail to cause.

# 1. Introduction

Depuis le début des années 2010, l'intelligence artificielle (IA) occupe le devant de la scène médiatique, après avoir été longtemps cantonnée aux laboratoires de recherche et à la science-fiction.

Le but de l'intelligence artificielle (IA) est double [1]. D'une part, l'IA s'attache à résoudre des problèmes qui relèvent d'activités humaines ou animales de nature variée : perception, planification, interprétation de données, diagnostic, prise de décision, compréhension du langage, conception. D'autre part, l'IA cherche à mieux comprendre et modéliser l'intelligence. Elle se rapproche ainsi des sciences cognitives dont elle s'inspire par ailleurs pour la conception de modèles (mémoire, raisonnement, apprentissage).

Les systèmes formels ont montré leurs limites intrinsèques pour la modélisation du raisonnement. Les travaux de Gödel et de Church ont démontré qu'un système de raisonnement de logique formel, à l'image du raisonnement mathématique, ne pourrait jamais prouver exhaustivement l'ensemble des propositions vraies. De même, la nécessité de restreindre un raisonnement à un champ d'application délimité et d'appuyer ce raisonnement sur des connaissances de nature diverse est apparue rapidement en IA. Cette approche symbolique de l'IA a donné lieu aux systèmes à bases de connaissances.

Une autre approche, dite connexionniste, tente de s'inspirer du fonctionnement du cortex cérébral. Un réseau neuronal est formé par l'interconnexion d'un grand nombre de neurones artificiels. Il présente des propriétés intéressantes, notamment la capacité d'apprendre à partir d'exemples.

Des succès récents, comme la victoire du programme de jeu d'échecs Deep Blue contre le champion du monde G. Kasparov et du programme de jeu de go DeepMind contre les meilleurs joueurs mondiaux ou du programme Libratus au poker, la mission des robots martiens, le jeu américain de questions-réponses Jeopardy, la reconnaissance d'images dans la compétition mondiale ILSVRC (*ImageNet Large Scale Visual Recognition Challenge*), la reconnaissance de la parole [41], les jeux vidéo tels que Dota 2 ou StarCraft II, le jeu de construction stratégique Jenga et d'autres ont médiatisé l'IA. Cette dernière a aussi inspiré les scénaristes de films comme A.I. Intelligence artificielle ou I, robot.

L'IA aborde un vaste champ d'activités que l'on peut classer en grands domaines : la reconnaissance et l'interprétation de formes et de données, le diagnostic, la planification d'actions et la robotique, l'aide à la décision, le traitement de la langue naturelle écrite et parlée, la formation assistée. Elle est ainsi entrée dans notre vie quotidienne.

Elle fait partie des sciences du numérique dont elle est une composante importante. Son évolution doit être envisagée dans ce cadre pour en saisir la globalité. En deux générations humaines, le numérique a vu la puissance des processeurs gagner quinze ordres de

grandeur, comme énoncé par la loi de Moore : l'humanité n'a jamais connu au cours de son histoire une telle situation dont l'influence sur l'évolution de l'IA est majeure.

# 2. Intelligence… et intelligence artificielle

## 2.1 Définition de l'intelligence

L'intelligence est une notion multiforme et difficile à préciser. Philosophes et scientifiques se sont attachés à la définir depuis des millénaires. Pour les besoins de l'IA, nous nous contenterons de qualifier l'intelligence par un ensemble de capacités, notamment de mémorisation, de structuration de la connaissance et de conceptualisation, de perception, de raisonnement, de prise de décision, d'apprentissage, de communication et de dialogue.

Ces capacités se retrouvent, à des degrés divers, dans les systèmes d'IA actuels. Ces systèmes se nourrissent des avancées dans ces différents domaines. Inversement, comme il a été dit, l'IA contribue à ce que nous comprenions mieux l'intelligence.

## 2.2 Les paradigmes de l'IA

La mise en œuvre d'un système d'IA se fonde sur un ensemble de paradigmes qui ont fait l'objet de très nombreux travaux depuis les premiers temps de l'IA. Citons :

- l'algorithmique de la résolution de problèmes,
- l'apprentissage,
- la reconnaissance de formes,
- la représentation des connaissances,
- la formalisation des raisonnements et des prises de décision,
- l'intelligence collective multi-agents.

Ces différents paradigmes s'appuient sur un ensemble de modèles qui seront décrits dans ce chapitre.

## 2.3 Les grands modèles de l'IA

Dès le début de l'IA dans les années 1950, deux grands types de modèles ont été proposés par les chercheurs pour concevoir des machines intelligentes :

- les *modèles symboliques,* correspondant à une approche (*Making a mind*) que l'on peut qualifier d'IA symbolique, permettent de doter les systèmes d'IA de mécanismes de raisonnement capables de manipuler les données symboliques qui constituent les connaissances d'un domaine. Cette approche fait appel aux modèles et méthodes de la logique. Elle a donné lieu aux systèmes à bases de connaissances [2]. A. Newell considérait le niveau de connaissance (*knowledge level*) comme le trait d'union entre l'homme et la machine intelligente, permettant de rationaliser le comportement d'un agent qui, à l'instar de l'être humain, peut prendre une décision en menant un raisonnement fondé sur ses connaissances [20]. Un aspect important

est la capitalisation et la diffusion du savoir et de l'expérience, éléments constituant la mémoire d'une organisation. Une base de connaissances se construit à partir d'ontologies qui la structurent et la contraignent. De telles bases interviennent dans de nombreux champs d'application comme l'intelligence économique, le droit, la production industrielle, la médecine, etc.

Pour être utiles, ces bases doivent inclure des modèles opérationnels du monde et du bon sens. Le projet Cyc lancé en 1984 par D. Lenat [26] a cet objectif ambitieux. Il s'agit du plus grand projet existant visant à formaliser les connaissances relatives à notre vie quotidienne (la version actuelle comprend des millions de faits et de règles formalisés dans un langage fondé sur le calcul logique des prédicats) et à les utiliser pour mener des raisonnements. Cyc comporte également des relations de causalité, reconnues très importantes dans le comportement humain. En 2018, Paul Allen (co-fondateur de Microsoft) a lancé dans son ***Institute for Artificial Intelligence*** le projet Alexandria, avec la même ambition d'apprendre le bon sens à une machine. Ce thème très important pour l'avenir de l'IA a également été repris par l'agence américaine DARPA sous la forme d'une proposition de projet *Machine Common Sense* ;

- les *modèles neuromimétiques* correspondent à une approche (*Modeling the brain*) que l'on peut qualifier d'IA connexionniste par analogie métaphorique. Ils reviennent à s'inspirer du fonctionnement du cortex cérébral. L'entité de base est un modèle formel très simplifié du neurone proposé par McCulloch et Pitts en 1943 [32]. Chaque neurone possède un certain nombre d'entrées synaptiques, chacune assortie d'un poids ; le neurone effectue la somme pondérée de ses entrées et active sa sortie, reliée à d'autres neurones, si la somme atteint un seuil prédéterminé. Un système est formé par l'interconnexion d'un grand nombre de tels neurones en couches successives. Cette approche a donné lieu aux réseaux neuromimétiques actuels, avec une grande variété des modèles, comme on le verra par la suite. La plupart de ces modèles, comme le perceptron multicouche, sont de type *feedforward*, ce qui signifie que les informations transitent dans un sens unique, de la couche d'entrée vers la couche de sortie. Comme d'autres modèles, les réseaux neuronaux sont capables de calculer toute fonction, d'où leur utilité en IA ;

- enfin, les ***modèles probabilistes et statistiques*** présentent un cadre formel intéressant pour capturer la variabilité inhérente au monde réel et en rendre compte. Ces modèles, tout comme les modèles neuromimétiques, sont capables d'apprendre à partir d'exemples. L'apprentissage revient ici à mémoriser des distributions de probabilités à l'aide d'algorithmes souvent complexes mais dont les propriétés sont parfaitement connues.

  Un modèle probabiliste largement utilisé est celui des réseaux bayésiens. Ces réseaux sont des graphes constitués de nœuds représentant les concepts d'un domaine et d'arcs représentant des relations de causalité probabilisées entre deux concepts (par exemple, tel état physiopathologique d'un patient peut être la cause

de tel symptôme, avec telle plausibilité). Un réseau bayésien permet de mener un raisonnement probabiliste sur des faits multiples grâce à des mécanismes de propagation de coefficients de probabilité à travers le réseau. Il est ainsi très intéressant dans des problèmes à choix multiples comme le diagnostic, notamment médical et industriel. Pour des applications en vraie grandeur, ces réseaux de dépendance conditionnelle peuvent atteindre des dimensions considérables. La mise au point et l'exploitation de tels réseaux sont des questions parfaitement maîtrisées.

Le temps est une dimension essentielle dans de nombreuses activités en IA. De ce fait, les modèles statistiques intégrant la variable temporelle, ou modèles stochastiques, comptent parmi les plus utilisés en intelligence artificielle. Le modèle stochastique le plus répandu est le modèle de Markov caché, ou MMC (*Hidden Markov Model, HMM*). C'est le cas en reconnaissance de la parole où chaque entité à reconnaître (mot, unité phonétique) est représentée par une source de Markov capable d'émettre le signal vocal correspondant à cette entité [3]. La reconnaissance revient alors à calculer la vraisemblance de la suite d'observations acoustiques constituant l'entité à reconnaître par rapport à chacun des modèles appris. Le modèle présentant la plus grande vraisemblance d'avoir émis cette suite d'observations fournit la réponse.

Les MMC ont également été utilisés avec succès dans d'autres domaines que la parole, en particulier l'interprétation d'images, la reconnaissance de l'écriture, l'interprétation de signaux (radar, sonar, biologiques, etc.) ou la robotique.

Les trois modèles présentés ci-dessus se rencontrent, parfois simultanément, dans les systèmes d'IA actuels. Les systèmes de traitement d'images et de reconnaissance de la parole, par exemple, sont le plus souvent fondés sur la complémentarité entre des modèles stochastiques MMC et des modèles neuronaux. Les modèles statistiques jouent également un rôle fondamental dans le traitement des grandes masses de données et leur exploitation, en particulier pour la découverte de régularités ou de connaissances.

Une caractéristique commune à tous ces types de modèles est leur capacité d'apprentissage à partir d'exemples. L'apprentissage, capacité fondamentale de l'intelligence, joue ainsi un rôle majeur dans le bon fonctionnement des systèmes d'IA (*cf.* par.4).

### 2.4 Les différents types d'IA

Il est d'usage de distinguer deux types d'IA en fonction des capacités des tâches envisagées : l'IA faible (*Narrow AI*) et l'IA forte (*General AI*). L'IA faible est celle des systèmes actuels, atteignant des résultats de très haut niveau, souvent comparables ou supérieurs à ceux d'êtres humains, mais dans des domaines restreints bien délimités comme évoqué ci-dessus (jeux, diagnostic, reconnaissance de la parole, identification d'images, etc.) pour lesquels un apprentissage spécifique a été mené. Les performances atteintes sont souvent spectaculaires, notamment depuis le début des années 2010 avec l'apport des

réseaux neuronaux profonds (*cf.* par. 5). Des exemples typiques, parmi beaucoup d'autres sont :

- la reconnaissance automatique de la parole. Les taux de reconnaissance atteints pour de très grands vocabulaires (de l'ordre de 100 000 mots) sont remarquables. En revanche, les systèmes n'ont pas la capacité de comprendre les mots ou les phrases reconnus,

- l'identification de personnes à partir de l'image de leur visage sur caméra vidéo.

L'IA forte tend vers celle de l'être humain, capable d'apprendre à mener des tâches complexes dans des domaines très différents, ou de comprendre et de raisonner sur des sujets variés en se fondant sur l'expérience acquise. Conscience de soi et du monde environnant, émotions, en sont des composantes de base. L'IA forte est encore dans les laboratoires de recherche et il est bien difficile de prédire quand elle en sortira…

## 3. Bref historique

L'intelligence artificielle est née dans les tout premiers temps de l'informatique. En 1950, Alan Turing publie un article désormais célèbre [4] dans lequel il pose la question « Les machines peuvent-elles penser ? » L'auteur propose un test, connu désormais sous le nom de *Test de Turing*, destiné à répondre à la question posée. Dans ce test, inspiré du jeu de l'imitation, une personne interroge une machine et un être humain qu'il ne voit pas. Lorsque, sur la base des réponses fournies, l'interrogateur ne peut plus distinguer l'être humain de la machine, la machine est déclarée intelligente. Les questions peuvent être de tout type. Pour l'instant, aucun programme n'a pu tromper un interrogateur pendant un temps suffisamment long. Les tentatives ont été nombreuses, à commencer par celles de J. Weizenbaum avec son système Eliza [31] qui simule le dialogue entre un psychothérapeute et son patient en recherchant des mots-clés dans le discours du patient, mais sans aucune capacité de compréhension. Les agents conversationnels (*chatbots*, et *callbots* via le téléphone) qui se sont multipliés ces dernières années augmentent progressivement leur compétence à mener un dialogue avec un être humain, comme le montre le prix Alexa lancé par Amazon [37].

Le terme *Artificial Intelligence* apparaît en 1956 lors d'une école d'été (*Dartmouth Conference*) réunissant un ensemble de jeunes chercheurs dont certains allaient devenir de grands noms du domaine : J. McCarthy, M. Minsky, etc. L'ambitieuse conjecture (toujours ouverte) énoncée par ces chercheurs est que: "*every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*". Au cours de cette rencontre furent étudiées les bases du raisonnement abstrait, de la résolution de problèmes, de l'apprentissage, qui constituent toujours le cœur de nombreux travaux en intelligence artificielle. Les travaux pionniers ont ainsi abordé les domaines des jeux (échecs, dames), de la reconnaissance de formes (parole, caractères écrits), de la résolution de problèmes (systèmes *General Problem Solver, GPS* [27] de A.

Newell et A. Simon, et Alice [28] de J.L. Laurière). Dans GPS, on trouve déjà la notion fondamentale d'heuristique (« qui aide à trouver », par opposition à algorithme [40]) :

*Many kinds of information can aid in solving problems: information may suggest the order in which possible solutions should be examined; it may rule out a whole class of solutions previously thought possible; it may provide a cheap test to distinguish likely from unlikely possibilities; and so on. All these kinds of information are heuristics: things that aid discovery. Heuristics seldom provide infallible guidance. . . Often they "work," but the results are variable and success is seldom guaranteed.*

La cybernétique, mouvement scientifique de l'après-Seconde Guerre mondiale autour de N. Wiener |29], W.R. Ashby, L. Couffignal et beaucoup d'autres, a joué également un rôle dans la genèse de l'IA, notamment avec la notion de régulation et le concept clé de rétroaction (*feedback*), aussi bien chez l'animal que dans la machine, toujours aussi important pour les réseaux neuronaux ou la robotique, spécialement pour des tâches de pilotage.

À la même époque, F. Rosenblatt inventait le perceptron monocouche [5], réseau neuronal classifieur linéaire, ancêtre des réseaux neuronaux profonds actuels qui sont toujours fondés sur la modélisation du neurone proposée en 1943. L'intérêt pour ces modèles a décliné vers 1969 avec la publication d'un livre sur leurs limites par M. Minsky et S. Papert [6]. Il faudra attendre le perceptron multicouche et la redécouverte de l'algorithme d'apprentissage associé de rétropropagation du gradient d'erreur [7] pour constater un regain d'activité dans ce domaine jusqu'aux réseaux neuromimétiques profonds actuels (*cf.* par. 5) qui occupent presque exclusivement la scène médiatique.

La conception de systèmes à bases de connaissances, et notamment, de systèmes experts, représente néanmoins toujours un domaine important de l'IA. Les systèmes experts sont conçus pour atteindre les performances d'experts humains dans des domaines limités en raisonnant sur un ensemble de connaissances acquises pour l'essentiel auprès de ces experts. Apparus vers 1975 (*cf.* le système de diagnostic des maladies du sang MYCIN [8]), ils ont eu un impact certain sur l'IA, et aussi un retentissement médiatique parfois exagéré car ils n'ont pas tenu toutes les promesses qui leur furent associées. Le terme de système expert disparaît peu à peu, au profit du concept plus général de système à bases de connaissances (SBC). Ce concept est fondé sur une séparation entre d'une part les connaissances nécessaires pour résoudre un problème et d'autre part les mécanismes de raisonnement qui exploitent ces connaissances.

Parmi ces travaux, les systèmes multi-agents occupent une part importante. L'idée est de parvenir à une décision commune à un ensemble d'entités par fusion d'informations et de points de vue, dans un cadre de coopération ou de compétition. Les modèles développés (tableau noir, acteurs, société d'experts) ont donné lieu à de nombreuses applications [38].

Dès les premières années de l'IA, la langue naturelle écrite a fait l'objet d'un ensemble de projets en traduction automatique (avec des résultats fort limités pendant longtemps, la difficulté de la tâche ayant été initialement sous-estimée) et en dialogue humain-machine. Le système SHRDLU [30], développé par T. Winograd, permettait un dialogue avec un

robot manipulateur dans un monde de blocs de formes (cubes, cônes, sphères) et de couleurs différentes. La capacité du système à interagir avec une personne et à répondre à des questions sur son univers de blocs a connu un succès considérable… malgré ses limites. En effet, le niveau de « compréhension » de la langue atteint par SHRDLU était essentiellement lié à la simplicité de son monde de blocs : nombre et types de blocs, nombre et niveau de complexité des actions possibles. Il a été impossible d'étendre le système à un univers plus complexe, sans parler du monde réel ! Toutefois, son influence a été notable dans les domaines de la langue naturelle et de la planification d'actions.

En ce qui concerne la langue parlée, les travaux en reconnaissance automatique de la parole ont débuté dès la fin des années 1950. Comme on l'a vu, l'utilisation de modèles stochastiques (modèles de Markov cachés), puis leur association à des modèles neuronaux ont permis d'atteindre des niveaux de performance remarquables en reconnaissance de mots ou de phrases, même pour des vocabulaires de dizaines de milliers de mots. En revanche, la compréhension de la parole demeure un sujet de recherche actif.

# 4. Apprentissage

## 4.1 Introduction

Comme on l'a vu, l'apprentissage, caractéristique fondamentale de l'intelligence, est partie prenante de la plupart des systèmes d'IA, permettant d'optimiser leurs performances. En 1959, Arthur Samuel, un des pionniers de l'IA, définit l'apprentissage automatique (*machine learning*) comme la capacité pour un ordinateur d'apprendre sans être explicitement programmé.

L'apprentissage automatique est une discipline très active et multiforme, selon les modèles sous-jacents. Toutes les méthodes impliquent l'existence de grandes bases de données d'exemples et parfois de contre-exemples, le plus souvent dûment étiquetés, sur lesquels se fonde l'apprentissage. Un système apprend à partir de chaque exemple, avec l'idée d'être capable de généraliser son comportement à de nouveaux cas non encore rencontrés, grâce aux bonnes propriétés des modèles appris.

On distingue deux grands types d'apprentissage en IA, l'apprentissage symbolique et l'apprentissage numérique. Nous résumons ci-dessous les grandes caractéristiques de ces méthodes.

## 4.2 Apprentissage symbolique

Ce type d'apprentissage, lié aux systèmes à bases de connaissances, concerne l'acquisition de concepts et de connaissances structurées [9]. Un exemple fondateur bien connu est Dendral, système expert en chimie organique, capable d'apprendre des règles d'explication à partir de données de spectrographie de masse [10]. Ajouter de nouvelles règles dans une base de connaissances pose des problèmes de gestion des bases, de maintien de cohérence

et, éventuellement, de généralisation des connaissances apprises par recherche d'explications.

Les nombreuses méthodes proposées relèvent de deux grands types d'apprentissage :

- l'apprentissage par détection de similarités (*Similarity-based learning*) dans lequel on apprend en détectant des similarités et des dissemblances dans une base d'apprentissage d'exemples et de contre-exemples. Cette recherche d'associations et de liens de causalité significatifs permet d'extraire des pépites de connaissances à l'aide de mécanismes de fouille de données associés à une évaluation par un expert du domaine. Cette coopération IA-expertise humaine se retrouve souvent dans les applications pratiques (*cf.* par. 7) ;
- l'apprentissage par recherche d'explications (*Explanation-based learning*) dans lequel on apprend à partir des explications extraites d'exemples et de contre-exemples.

Il faut également adjoindre à ces deux types d'autres méthodes telles que l'apprentissage par analogie, la construction automatique de taxonomies et d'ontologies, l'inférence inductive de connaissances ou encore le regroupement automatique de concepts (*cf.* ci-dessous), ainsi que les arbres de décision [11]. Un arbre de décision est une structure arborescente qui permet de classer une certaine entité (objet, cas, etc.) par un ensemble de questions. À chaque embranchement d'un arbre est associée une question portant sur les caractéristiques de l'entité étudiée. Les feuilles terminales correspondent aux différentes classes définies. L'apprentissage revient à apprendre l'ensemble des questions relatives aux caractéristiques des objets présentés. Plusieurs logiciels de construction d'arbres ont été développés (par exemple CART, C4.5, ID3, etc.).

Les travaux en apprentissage symbolique ont donné lieu à de nombreuses réalisations. En revanche, les méthodes proposées demeurent spécifiques et liées à un domaine particulier. Depuis le début des années 2000, ces méthodes sont moins étudiées que les méthodes relevant de l'apprentissage numérique.

### 4.3 Apprentissage numérique

**L**a majorité des progrès réalisés en IA depuis une dizaine d'années sont dus à l'utilisation de méthodes numériques d'apprentissage.

*Les modèles statistiques* (modèles de Markov cachés et réseaux bayésiens) ont déjà été brièvement présentés au par. 2. Un des intérêts de ces modèles réside dans l'automatisation de l'apprentissage des différents paramètres et distributions de probabilité du modèle à partir de données représentatives de l'application considérée (appelées observations). Le principe est de trouver un modèle qui maximise la probabilité d'un ensemble (ou dans le cas des MMC d'une séquence) d'observations, c'est-à-dire de déterminer le modèle qui explique le mieux cette séquence. Il n'est pas possible de trouver un tel modèle de façon analytique. L'apprentissage est alors assuré par des algorithmes itératifs d'estimation des

paramètres, notamment l'algorithme de Baum-Welch, cas particulier de l'algorithme Espérance-Maximisation, dit EM.

Dans les années 1990, un autre modèle numérique statistique, l*es machines à vecteurs supports*, appelées aussi Séparateurs à Vaste Marge (*Support Vector Machines, SVM*) ont contribué de façon notable au succès des méthodes numériques d'apprentissage [12]. Une SVM est essentiellement un classifieur discriminant à deux classes (qui peut être étendu à une SVM multiclasses) dont le critère d'optimisation est la largeur de la marge entre les deux classes, c'est-à-dire la zone vide autour de la surface de décision définie par les formes les plus proches. Ce modèle est intéressant, mais les réseaux neuronaux profonds se sont révélés plus performants que les SVM pour de nombreuses applications.

*Les modèles neuronaux* utilisent également un apprentissage numérique. Le principe est d'optimiser les poids des connexions entre les neurones des différentes couches du modèle. Comme il a été dit au par. 3, l'algorithme de rétropropagation du gradient d'erreur a permis le développement dans les années 1990 de modèles neuronaux comportant un nombre limité (quelques unités) de couches cachées. L'idée de ce type d'apprentissage est la suivante : si la réponse du système à une entrée donnée est incorrecte, un gradient d'erreur est calculé en fonction des réponses de la couche de neurones de sortie. Ce gradient est rétropropagé de couche en couche depuis la couche de sortie jusqu'à la couche d'entrée en modifiant les poids des connexions inter-neurones de façon adéquate. On démontre que, moyennant un nombre suffisant d'exemples d'apprentissage, l'algorithme converge vers une configuration stable de poids pour l'ensemble des neurones. Récemment, le nombre de couches cachées a été considérablement augmenté, tout en conservant la capacité d'apprentissage à partir d'exemples, ce qui a donné naissance aux réseaux neuronaux profonds (*Deep Neural Nets, DNN*) présentés au par. 5.

Après avoir appris sur un grand nombre d'exemples étiquetés, un réseau neuronal est non seulement capable de classifier correctement ces exemples, mais peut aussi traiter de nouveaux objets de même catégorie qu'il n'a pas vus durant la phase d'apprentissage ; cette capacité de généralisation est une des propriétés très intéressantes de ces modèles.

### 4.4 Supervisé *vs* non-supervisé

Les méthodes, symboliques et numériques, décrites ci-dessus sont toutes de type *apprentissage supervisé*, ou avec professeur. Elles nécessitent, comme il a été dit, de grandes quantités d'exemples étiquetés avec la bonne réponse associée (portion de signal vocal et phonème ou syllabe correspondant, description d'un cas de panne et diagnostic associé, image et sa description, etc.). La réalisation de telles bases d'exemples annotés est très coûteuse (les bases d'images utilisées pour entraîner un système d'identification d'objets comportent des millions d'images indexées ; les bases liées à la reconnaissance de la parole contiennent des centaines d'heures de parole étiquetées phonétiquement). Une tendance apparue avec Internet est de sous-traiter aux utilisateurs eux-mêmes l'étiquetage des données comme une tâche, parfois invisible, annexe à l'utilisation d'un système (*crowdsourcing*).

Une autre classe de méthodes concerne *l'apprentissage non supervisé*, dans lequel l'opération se fait de façon totalement autonome. Des données non étiquetées sont présentées en entrée sans indiquer les réponses attendues en sortie. C'est le mécanisme d'apprentissage lui-même qui propose des catégories pour regrouper les réponses possibles. Cette solution semble idéale dans la mesure où elle ne nécessite pas de grandes bases de données. En fait, les deux types d'apprentissage ne sont pas destinés aux mêmes tâches. L'apprentissage non supervisé cherche à partitionner de lui-même, sans intervention extérieure d'un « professeur », les données qui lui sont présentées en catégories homogènes au sens d'un certain critère ou d'un ensemble de caractéristiques communes. Les méthodes de regroupement, parfois appelées coalescence (*clustering*), ont également donné lieu à de nombreuses études et au développement de méthodes et de logiciels variés (partitionnement, hiérarchisation, etc.), notamment dans le domaine de la reconnaissance des formes. Le succès de l'apprentissage supervisé pour les réseaux neuronaux a quelque peu occulté l'importance de l'apprentissage non supervisé. Ce dernier est certainement appelé à se développer, par référence à l'être humain et à l'animal. Une autre tâche de cet apprentissage est la recherche de liens de causalité qui sont d'une grande importance dans de nombreuses activités telles que le diagnostic ou la prise de décision, comme il a été dit plus haut. Ce type d'apprentissage peut être utilisé simultanément avec un apprentissage supervisé dans lequel les ensembles de données étiquetées sont remplacées par des entités qui changent au cours du temps telles que des vidéos : la trame du temps *t* de la vidéo est utilisée comme un prédicteur pour la trame *t+1*, sans aucun étiquetage préalable [44].

Une solution intermédiaire a également été proposée, qualifiée de semi-supervisée. Comme l'étiquetage complet d'une base d'apprentissage est une tâche lourde et coûteuse, l'idée est de restreindre le volume de données étiquetées nécessaires à un bon apprentissage. Un exemple de tel apprentissage est le co-apprentissage, dans lequel deux classifieurs apprennent un ensemble de données, mais en utilisant chacun un ensemble de caractéristiques différentes, si possibles indépendantes. Le classifieur le mieux capable de bien traiter un exemple d'apprentissage va jouer le rôle de « professeur » pour l'autre. Des approches de ce type ont été utilisées pour la classification de documents HTML et en traitement de la langue naturelle. Une idée similaire se trouve dans les réseaux antagonistes génératifs (*Generative Adversarial Networks, GAN*). Dans un GAN, deux réseaux sont placés en compétition. Le premier réseau, le générateur, génère un échantillon d'apprentissage, tandis le second, le discriminateur, essaie de détecter si un échantillon est réel ou bien s'il est le résultat du générateur [13].

### 4.5 Apprentissage par renforcement

Parmi les méthodes d'apprentissage, l'*apprentissage par renforcement* occupe une place à part. Le principe est d'apprendre par essais et erreurs comment se comporter de manière optimale dans des environnements incomplètement connus, situation très commune dans la réalité. Cet apprentissage, très développé en robotique, comme l'algorithme de *Q-Learning* [14], s'est inspiré au départ de théories de psychologie animale et humaine ; il modélise le comportement d'apprentissage optimal par essais et erreurs permettant de

s'adapter à un environnement. Imaginons un système, ou un agent, situé dans un certain environnement, changeant et mal connu de l'agent, comme un robot se déplaçant dans un univers qu'il cartographie au fur et à mesure. L'agent peut décider de mener un ensemble d'actions qui lui apporteront éventuellement une récompense. Une action de l'agent conduit à un nouvel état de l'environnement dans lequel il peut mener une nouvelle action qui conduit à un nouvel état, etc. L'environnement est le plus souvent stochastique, ce qui signifie que le nouvel état est aléatoire. Ce processus d'essais et erreurs est conduit jusqu'à obtenir la meilleure réponse possible. Une façon efficace la plus utilisée pour raisonner dans de telles conditions est le formalisme des processus de décision markovien (*Markov Decision Process*, MDP) et celui des processus de décision markovien partiellement observables (*Partially Observable Markov Decision Process*, POMDP), ces derniers étant plus proches de la réalité des applications. Il existe de très nombreuses variantes de l'apprentissage par renforcement, en général formalisées dans un cadre probabiliste. Cet apprentissage est largement utilisé, y compris dans les réseaux neuronaux profonds (*cf.* par. 5), notamment dans le vaste domaine des jeux : backgammon (1994), Atari (2013) et, encore plus récemment go et poker : le programme *Libratus* développé à *Carnegie Mellon University* évoqué au par. 1 utilise aussi ce type d'apprentissage.

La combinaison de l'apprentissage profond et de l'apprentissage par renforcement semble également prometteuse. De très bons résultats ont déjà été obtenus en classification d'images, et aussi en traitement de la langue naturelle.

## 5. Les réseaux neuronaux profonds

Une avancée majeure de l'IA depuis 2010 s'est produite dans le domaine des réseaux neuronaux profonds (*Deep Neural Networks,* DNN) et des algorithmes d'apprentissage associés. Déjà, dès 2006, les modèles acoustiques de reconnaissance de la parole avaient été améliorés de façon importante grâce aux modèles neuronaux profonds. Depuis, les DNN ont montré leur efficacité dans des domaines très variés, non seulement en reconnaissance de la parole, mais aussi dans les jeux, en traitement de textes, en vision par ordinateur, en diagnostic, en robotique, etc. Comme on l'a vu au par. 3, un réseau neuronal est un classifieur capable d'apprendre des fonctions de décision. Le perceptron monocouche de Rosenblatt avait des performances très limitées puisqu'il ne pouvait apprendre que des fonctions linéaires. L'introduction de couches cachées en nombre restreint a permis d'augmenter les performances. Globalement, un réseau neuronal est capable d'apprendre une mise en correspondance (*mapping*) entre ses entrées et ses sorties, ce qui permet d'utiliser de tels systèmes pour des tâches de classification permettant d'identifier la classe d'appartenance de l'entité placée en entrée. Ces entités peuvent être de nature extrêmement diverse : mots prononcés par un locuteur, image, diagnostic d'un patient, place d'un pion sur un jeu de go, etc. Les DNN sont caractérisés par le fait que leur profondeur (c'est-à-dire le nombre de couches cachées de neurones) est augmentée de façon très importante pour atteindre jusqu'à un millier de couches, ce qui leur donne la capacité d'apprendre des fonctions de mise en correspondance beaucoup plus complexes, d'où leur succès actuel.

La rapide émergence des réseaux profonds est due à la conjonction de trois conditions :

- l'existence de très grandes bases de données étiquetées nécessaires à l'apprentissage de ces modèles. Il s'agit d'un exemple du phénomène récent de *Big Data*. En reconnaissance de la parole, les grands opérateurs du domaine disposent ainsi de millions d'heures de parole, ce qui permet de disposer de systèmes de reconnaissance dans de nombreuses langues (plus d'une centaine pour Google). Ces bases s'enrichissent quotidiennement. 80 % des données qu'elles renferment sont non structurées, ce qui nécessite une analyse sémantique préliminaire. Le RGPD (Règlement général sur la protection des données) mis en place en 2018 par l'Union Européenne, notamment dans la suite de la loi française de 1978 « Informatique et libertés » et des travaux de la CNIL, aura une influence sur l'exploitation de ces grands entrepôts de données. Il importe en effet d'assurer la protection des données personnelles, tout spécialement les données sensibles comme les informations médicales.

- la disponibilité de capacités de calcul en constante augmentation (notamment à l'aide de cartes additionnelles GPU et du calcul parallèle haute performance). Des exemples sont le processeur TPU (*Google's Tensor Processing Unit*) ou les puces NVIDIA qui accélèrent notablement le temps d'apprentissage d'un DNN. L'apprentissage d'un DNN peut également être implémenté sur un FGPA (*Field Programmable Gate Array*). Sur le plan des logiciels, citons Theano, bibliothèque logicielle écrite en langage Python pour le développement de systèmes d'apprentissage profond [15]. Ces moyens sont nécessaires en pratique, sachant que le nombre de paramètres à apprendre dans un DNN peut être de l'ordre du milliard.

  La loi de Moore, évoquée en introduction, ayant à peu près atteint ses limites, il ne faut plus compter seulement sur l'augmentation intrinsèque de puissance des circuits intégrés pour améliorer les performances d'apprentissage, mais aussi sur des solutions innovantes, plus créatives, tant logicielles que matérielles. Ceci constitue un changement important par rapport à ce que nous avons vécu lors des dernières décennies. Les algorithmes actuels d'apprentissage fonctionnent aussi bien sur le « nuage » (*cf.* les solutions *Open Source* proposées par Google avec son *Cloud Machine Learning*, ou d'autres) que localement là où s'effectue la capture des données. La technologie des circuits intégrés proposera certainement à l'avenir des architectures d'accélérateurs d'apprentissage, adaptées aux différents lieux de traitement des données.
- l'amélioration des algorithmes d'apprentissage profond (*Deep Learning* [16]). Initialement, les algorithmes d'apprentissage utilisaient des sigmoïdes comme fonctions non linéaires pour transmettre le résultat du traitement de l'erreur d'une couche à la suivante. L'utilisation de fonctions ReLU (*Rectified Linear Unit*) a permis un gain en temps très important. Par ailleurs, la technique de *dropout* qui

revient à éliminer des neurones inutiles durant l'apprentissage permet d'éviter l'écueil du sur-apprentissage, susceptible d'apparaître dans les réseaux neuronaux.

La complexité de l'apprentissage d'un DNN a conduit les chercheurs à essayer de réduire cette complexité. Une solution initialement proposée par K. He [45] est celle des réseaux résiduels, *ResNet*. L'idée est d'introduire des raccourcis dans les connexions entre les neurones de couches cachées, ce qui facilite grandement l'apprentissage.

L'apprentissage fondé sur la rétropropagation du gradient d'erreur peut aussi parfois être couplé à l'apprentissage par renforcement. Ainsi le logiciel *AlphaGo Zero* de *DeepMind* qui s'est révélé le meilleur au jeu de go combine DNN et recherche arborescente de type Monte Carlo. Son apprentissage comporte une phase d'apprentissage profond supervisé fondé sur une base de parties jouées par des experts (plus de 100 000) et une phase d'apprentissage par renforcement partant simplement des règles du jeu de go.

L'apprentissage de tels systèmes nécessite, comme on l'a dit, des quantités de données et des moyens de calcul considérables. Les systèmes résultants sont compétents dans la tâche unique sur laquelle a porté leur apprentissage. Une tendance actuelle est de concevoir des environnements d'apprentissage multi-tâches. Le système IMPALA (*Importance Weighted Actor-Learner Architecture*) de *DeepMind* en est un exemple [17].

Outre les perceptrons, il existe d'autres types de réseaux neuronaux profonds, avec leurs propres forces et faiblesses :

- *les réseaux récurrents* : l'état d'un réseau est ici fonction de l'entrée du réseau à un instant donné et de son état à l'instant précédent. Sa structure permet une mémoire des entrées précédentes qui persiste dans ses états internes et peut en conséquence avoir un effet sur ses sorties futures. Les réseaux récurrents sont ainsi les plus profonds des réseaux dits profonds. Ils permettent en outre d'exploiter de façon naturelle et efficace le parallélisme des programmes. Ces réseaux peuvent être entraînés par une variante de l'algorithme de rétropropagation, mais leur apprentissage est délicat. Ils sont bien adaptés au traitement de séquences temporelles telles que celles rencontrées en reconnaissance de la parole, reconnaissance de l'écriture, traduction automatique, etc. Le type de réseau récurrent le plus utilisé, tout spécialement en reconnaissance de la langue et en traduction, est le réseau de neurones récurrents à mémoire court et long terme (*Long short-term memory, LSTM)* [18]. Un réseau LSTM peut mémoriser une information à plus ou moins long terme, un peu comme le fait notre cerveau. Par exemple, lors de l'analyse d'une phrase, le réseau peut se souvenir du début lorsqu'il arrive à la fin ;

- *les réseaux convolutionnels* (*Convolutional Neural Networks, CNN ou ConvNet*) : ces réseaux dont un des initiateurs est Y. Le Cun [19], tirent leur inspiration du travail de Hubel et Wiesel sur le cortex visuel du chat. Un réseau convolutionnel profond est formé d'un ensemble hiérarchique de couches de cellules formant un tuilage couvrant l'ensemble du champ visuel. Chaque couche fournit un ensemble

de paramètres convolutionnels de niveau d'abstraction de plus en plus haut à partir des données d'entrée du réseau. Chaque cellule agit comme un filtre local permettant d'exploiter les corrélations spatiales présentes dans une image. L'extraction automatique et la hiérarchisation des paramètres représentatifs des données en entrée sont une des grandes forces de ce modèle. En effet, dans un système d'IA « traditionnel » (reconnaissance de formes, diagnostic), les données brutes présentées en entrée d'un système sont transformées en vecteurs de paramètres (*features*) sur lesquels est prise la décision finale. Une succession de niveaux de traitement permet de transformer les données brutes en représentations de plus en plus abstraites.

Par exemple, une image est présentée en entrée sous forme d'un ensemble de pixels. Le premier niveau de traitement va détecter et localiser des bords ou des angles, le second des arrangements particuliers de bords que le troisième niveau va identifier comme des objets, et ainsi de suite jusqu'à obtenir une représentation abstraite des données adéquate pour la tâche poursuivie.
De façon similaire, la parole se décompose en phrases, mots, syllabes, phones, traits acoustico-phonétiques (bruits, formants, etc.) jusqu'au niveau initial du signal acoustique. La définition de ces niveaux, ainsi que l'extraction automatique des paramètres associés, requiert une importante expérience du domaine d'application.
Le fait que, dans les CNN, la hiérarchie des niveaux ainsi que les paramètres associés sont déterminés automatiquement, sans intervention humaine lors de l'apprentissage, est un des gros avantages de ces modèles. Plus précisément, un CNN présente une structure comportant deux types de couches, directement inspirée des connaissances actuelles des neurosciences de la vision :
- des couches de convolution destinées à détecter les arrangements pertinents de paramètres provenant de la couche précédente,
- des couches de regroupement chargées de rassembler des paramètres semblables.
Les CNN se sont révélés très efficaces dans la représentation de données complexes structurées et dans leur traitement, tout spécialement dans les domaines de la classification d'images [23] (dans la compétition annuelle ILSVRC (*cf.* par. 1) dont le but est de localiser et identifier des objets dans des scènes naturelles, et récemment dans la reconnaissance de visages. Les CNN obtiennent les meilleurs résultats, de l'ordre de 2 % d'erreur de classification d'images pour ILSVRC, alors que l'erreur humaine sur les mêmes données est de l'ordre de 5 %). Il en est de même pour la reconnaissance de la parole [24] et de l'écriture. D'excellents résultats ont également été obtenus dans une grande variété de domaines [25] tels que la physique (analyse de résultats d'expériences d'accélération de particules), la biologie (traitement de mutations d'ADN), le traitement de la langue naturelle et la traduction automatique. Dans ces deux derniers champs d'activité, les traditionnels modèles de langage à base de n-grammes (courtes séquences de lettres et de symboles de longueur 2 à 5, voire plus) sont remplacés par des modèles de langage neuronaux.

Tous les réseaux neuromimétiques actuels présentent une structure plane bidimensionnelle, les différentes couches de neurones étant organisées horizontalement. Il n'en va pas de même dans le cortex humain qui possède une structure tridimensionnelle où les neurones sont organisés en colonnes. Nous avons réalisé à Nancy une modélisation informatique du modèle de colonne corticale d'Y. Burnod [33]. J. Hinton a récemment proposé un modèle similaire de *capsules* [34]. Ces modèles sont intéressants dans leur principe et doivent être améliorés pour devenir compétitifs avec les DNN actuels.

Malgré leurs propriétés remarquables ayant conduit aux spectaculaires performances évoquées ci-dessus, les réseaux profonds sont intrinsèquement limités et ne résoudront donc pas tous les problèmes qui se posent à l'IA. Les limitations les plus importantes concernent :

- la nécessité d'énormes quantités de données d'apprentissage. Contrairement à l'être humain ces réseaux ne possèdent pas encore de capacité d'abstraction,
- le manque de transparence, c'est-à-dire l'incapacité pour ces systèmes d'expliquer leurs résultats (aspect boîte noire). Des efforts sont faits en ce sens (cf. par. 7),
- la difficulté d'intégrer des connaissances explicites telles que celles utilisées en IA symbolique (cf. par. 2-3) : ainsi, un DNN peut apprendre aisément des corrélations entre ses entrées et ses sorties, mais sans pouvoir expliciter les relations de causalité éventuelles entre ces données. La causalité demeure un thème de recherche important en IA symbolique [42].

## 6. **Domaines d'application**

Depuis plusieurs décennies, l'IA a conduit au développement d'applications opérationnelles dans de nombreux domaines tels que le diagnostic et l'aide à la décision avec les systèmes à bases de connaissances ou les réseaux neuronaux, ainsi que la reconnaissance de formes et de la parole avec les modèles stochastiques. Mais l'introduction des DNN a permis d'atteindre des niveaux de performance inégalés qui ont valu une couverture médiatique sans précédent dans quasiment tous les secteurs d'activité. Comme toujours, le transfert de la recherche en laboratoire vers les applications réelles nécessite des investissements conséquents en temps et en argent. Le cheminement est long entre une première annonce et un véritable produit commercialisé. En médecine, notamment, le déploiement d'applications implique un mécanisme d'évaluation complexe, notamment clinique. Enfin, le succès d'une application implique une véritable confiance

dans l'IA ainsi qu'une claire identification des responsabilités (*cf.* la voiture autonome en cours d'expérimentation et de mise au point).

Sans prétention d'exhaustivité, car la liste s'allonge continuellement, on peut citer :

- Dans le domaine des jeux, avec l'évolution de la technologie et des algorithmes, les systèmes fondés sur l'IA sont devenus les meilleurs : morpion (1952), dames (1994), échecs (1997), go (2016), poker (2016), jeux vidéo (2017).
- En aide au diagnostic et à la recommandation d'actions (médical, spatial, industriel...) et aide à la décision (banques, assurances, conduite de procédés, domaine militaire).
- En aide à la conception, notamment de puces électroniques (*Electronic Design Automation*, EDA).
- En reconnaissance et synthèse de la parole : la conjonction HMM-DNN conduit à des taux de reconnaissance proches de l'humain (annonce récente de Google) et fait le succès des assistants personnels vocaux [41].
- En identification de locuteurs et détection d'émotions.
- En traitement de la langue naturelle : systèmes de questions-réponses, traduction, dialogue, description d'une image ou d'une scène. Le système Jeopardy d'IBM évoqué au par.1 est un bel exemple du niveau atteint dans le domaine.
- En interprétation de signaux (surveillance, conduite, cybercriminalité).
- En robotique : robots autonomes (exploration, intervention en milieu hostile), robots de compagnie (pour enfants malades, personnes âgées ou astronautes, *cf.* par. 7), voiture autonome sans chauffeur.
- En traitement d'images : diagnostic à partir d'images médicales (rayons X, IRM...), reconnaissance de l'écriture, reconnaissance de visages (avec les aspects éthiques associés et les dérives potentielles d'identification d'individus sans leur consentement…), télédétection, détection, classification et localisation d'objets, identification d'actions, etc.
- En biologie, notamment pour le traitement des séquences d'ADN.
- En finance : évaluation du risque, *trading, marketing* prédictif.
- En médecine : les applications citées ci-dessus préfigurent une évolution vers une médecine prédictive et personnalisée, ainsi qu'une évolution des métiers (radiologie, dermatologie, anatomo-pathologie...).

On assiste actuellement de ce fait à un accroissement considérable des investissements privés en intelligence artificielle par les industriels du domaine : Google, Facebook, IBM, Microsoft, Amazon, Adobe, Yandex, Baidu... Les startups ont également proliféré dans le

monde entier. Elles sont ainsi actuellement 14 fois plus nombreuses qu'en l'an 2000 en Amérique du Nord, selon l'index de l'IA de l'Université de Stanford. Leur nombre total est de plus de 1700 dans le monde. Le chiffre d'affaire de l'IA se situe autour de 2 milliards de dollars américains en 2018 (et pourrait atteindre près de 90 milliards en 2024, d'après le cabinet d'analyse Tractica), tandis que plus de 55000 brevets, relatifs à l'IA au sens large, ont été déposés et que plus de 130 000 articles scientifiques ont été publiés cette même année, selon l'Organisation Mondiale de la Propriété Industrielle, OMPI.

Parallèlement, les acteurs institutionnels de la recherche lancent de grands projets : en Europe, aux États-Unis (*cf.* le projet *Quest for Intelligence* du MIT tendant à coupler la compréhension de ce qu'est l'intelligence et la réalisation de nouveaux systèmes utiles à l'humanité, ou encore la création du *Joint Artificial Intelligence Center* par le Ministère de la Défense dans le cadre des *National Mission Initiatives*, grands projets s'attaquant aux grands défis actuels, intégrant les aspects éthiques et humanitaires des recherches).

La montée en puissance de l'IA pose la question des conséquences en matière d'emploi. L'IA permet d'automatiser de nombreuses tâches répétitives. Nous assistons à une évolution importante du travail car l'IA supprimera ou transformera très certainement des métiers dans de nombreux secteurs d'activité, tout en nécessitant une évolution forte des qualifications (dans la maintenance, la gestion des données, la création assistée, etc.) et des compétences, y compris pour des métiers qui sont encore à imaginer.

## 7. Perspectives et conclusion

Depuis 2010, la conjonction de l'évolution technologique, de l'amélioration des algorithmes d'apprentissage et de la disponibilité de grandes bases de données a conduit à des réalisations spectaculaires fondées essentiellement sur les réseaux neuronaux profonds, les DNN. L'enthousiasme pour l'IA, dernier outil en date de *homo faber*, est donc très grand et sans précédent dans l'histoire du domaine, avec toutes les exagérations médiatiques et les promesses (qui ne seront pas forcément toutes tenues) que cela implique. Cette inflation de promesses pourrait être préjudiciable à l'IA, comme cela s'est déjà produit dans le passé.

Même si aucun système n'est pour l'instant capable de réussir le test de Turing, les progrès réalisés sont importants. Le système d'IBM *Project Debater* est ainsi capable de débattre en temps réel avec un humain sur un sujet donné. Ce système élabore un argumentaire en puisant dans une base de données d'articles et de connaissances.

Par ailleurs, les recherches dans tous les domaines de l'IA se poursuivent, même si elles sont largement occultées par les réseaux neuronaux profonds. On peut considérer que les progrès récents de l'IA ont été surestimés ou du moins mal compris. Ces progrès portent essentiellement sur la reconnaissance de formes ou de concepts, et non sur la compréhension proprement dite : percevoir n'est pas comprendre. Des avancées importantes ont été faites en la matière mais les recherches fondamentales doivent être poursuivies, nos systèmes étant encore bien en deçà de l'intelligence humaine ou même animale. Ainsi, dans le domaine de la voiture autonome, des travaux tendent à doter le système de conduite de notions de physique et de psychologie humaine lui permettant

d'anticiper, ou du moins de gérer au mieux certains comportements inattendus de conducteurs.

Simultanément, les travaux se sont amplifiés dans le vaste champ de l'apprentissage automatique, supervisé ou non, car il s'agit d'un domaine clé pour l'avenir de l'IA.

Un problème lié à l'utilisation de DNN de très grande dimension est leur consommation électrique. Les DNN sont énergivores, contrairement aux systèmes biologiques dans lesquels les neurones communiquent par impulsions électriques très brèves, ou *spikes*, ce qui diminue grandement la consommation électrique. Des modèles informatiques de s*piking neurons* ont été proposés |22], de même que les algorithmes d'apprentissage associés. Les résultats obtenus en reconnaissance d'images et de contrôle de robots sont très encourageants, tout en demeurant inférieurs à ceux obtenus par les DNN classiques. Une voie prometteuse est l'utilisation de l'informatique quantique. Une implémentation d'un neurone de perceptron a déjà été réalisée [39], premier pas à confirmer vers un apprentissage quantique des DNN.

Les réseaux neuronaux profonds actuels vont encore progresser dans les années à venir, mais leurs limites vont apparaître, comme il a déjà été dit au par. 5. L'être humain a la capacité remarquable d'apprendre à porter son attention sur les parties pertinentes des objets qu'il observe. Les futurs DNN auront sans doute cette capacité, dans le prolongement des travaux antérieurs sur le développement de modèles d'attention sélective par renforcement. Une amélioration des modèles pourrait provenir d'équipes de recherche pluridisciplinaire regroupant des spécialistes d'IA, de sciences cognitives et de neurosciences. Cela n'est pas nouveau : l'idée initiale de l'algorithme d'apprentissage par rétropropagation vient ainsi de la psychologie. Un rapprochement de ces modèles avec les modèles symboliques « traditionnels » de l'IA pourrait aussi se révéler fructueux. Cela permettrait de remplacer avantageusement la manipulation symbolique de règles ou d'autres types de connaissances par diverses opérations menées sur des grands vecteurs de données numériques [21]. Par ailleurs, la conception de modèles neuro-symboliques hybrides est une voie qu'il convient de poursuivre [43].

Parmi les évolutions nécessaires pour aller vers une IA forte, aux capacités cognitives accrues, trois domaines me paraissent fondamentaux :

- l'*apprentissage* dans tous ses aspects (supervisé, non supervisé, par renforcement, etc.) et pas seulement relatif aux réseaux neuronaux, en lien avec la simulation des phénomènes considérés (cette technique est courante pour les jeux, indispensable pour les domaines à risques, tels que la voiture autonome). Un DNN est capable de reconnaître un chat après avoir analysé des milliers d'images de chat lors de la phase d'apprentissage. Il suffit pour un enfant d'avoir vu quelques chats pour accomplir la même tâche. L'apprentissage par renforcement va encore se développer dans de nombreux secteurs d'activité, notamment grâce à des progiciels tels que *Brain* acquis récemment par Microsoft. Par ailleurs, les progrès en la matière s'appuieront en partie, comme on l'a vu à plusieurs reprises, sur les apports des sciences cognitives et des neurosciences ;

- *la compréhension* de formes et de situations. Ce domaine nécessite des progrès en modélisation du bon sens (*cf.* par. 2-3) et en modélisation prédictive du monde, mécanismes d'interprétation fondés sur des ontologies et d'autres types de connaissances. L'Union Européenne devrait à l'avenir demander à toute société ou organisation d'être capable de fournir une explication à toute décision prise par un système automatique. Ce n'est encore qu'un projet pour tenter de briser le côté « boîte noire » des systèmes d'IA, déjà évoqué, plus précisément des réseaux neuronaux, les systèmes symboliques à bases de connaissances ayant la possibilité de tracer, dans une certaine mesure, le chemin de raisonnement ayant conduit à une décision. Cette tentative de l'UE n'est pas unique. Ainsi, l'agence américaine DARPA vient de lancer un programme de recherche intitulé *Explainable Artificial Intelligence.* Fournir une explication à une décision nécessite un niveau de compréhension du domaine d'activité ;

- *la coopération humain-machine*. La plupart des systèmes d'IA ne sont pas destinés à remplacer des humains, mais plutôt à coopérer avec eux pour optimiser leurs performances. La recherche doit tendre vers la création de modes d'interaction et de dialogue aussi efficaces que possible. L'IA ne doit pas conduire à la marginalisation de l'évaluation humaine dans un cadre applicatif donné, lorsque cette dernière est pertinente voire indispensable. Il s'agit de définir et de mettre en place le cadre permettant d'intégrer au mieux l'intelligence humaine et l'IA. De nombreuses expériences de ce type menées dans des domaines sensibles prouvent l'intérêt de telles approches, en particulier en médecine (*e.g*. le diagnostic de maladies mentales [35]). Un autre exemple de l'utilité, sinon la nécessité, de la collaboration entre l'IA et l'intelligence humaine concerne l'exploitation des énormes gisements de données de plus en plus disponibles. Ces données ne servent pas seulement à entraîner les réseaux neuronaux. Des techniques de fouille de données permettent aussi d'exploiter ces *big data* pour en extraire des pépites de connaissances. La complémentarité être humain-machine peut se révéler très fructueuse : une connaissance extérieure à ces données, notamment celle détenue par des experts du domaine considéré, en permet une exploitation optimale [36].

Un domaine connexe est celui des robots compagnons ou domestiques. Les expériences en la matière se multiplient, en particulier à destination des personnes âgées, comme les projets européens GIRAFFPlus et MOBISERV ou le projet français ROMEO2. De tels robots, comme aussi Pepper d'Aldebaran ou d'autres, se développent de plus en plus, au Japon mais aussi dans d'autres pays. En France, plusieurs EHPAD en sont équipés. Un autre exemple est CIMON (*Crew Interactive Mobile Companion*), un petit robot doté de parole conçu par Airbus pour assister les astronautes dans leurs tâches à bord de la station spatiale internationale.

La singularité (point hypothétique où l'IA dépassera l'intelligence humaine) paraît encore peu crédible et la crainte de voir les systèmes d'IA échapper au contrôle humain et prendre le pouvoir paraît infondée. En revanche, les travaux et les réflexions sur le transhumanisme,

l'homme augmenté, tout comme la singularité technologique, nécessitent d'amplifier la réflexion sur les aspects éthiques et sociétaux de la recherche en IA, ainsi que son enseignement. C'est l'être humain et son éthique qui doivent limiter les pouvoirs des systèmes d'IA, de même que leur utilisation frauduleuse : usurpation des données personnelles, pouvoir létal des robots, etc.

Nos valeurs fondamentales comme la liberté, la sécurité, l'intérêt collectif, la dignité de la personne, doivent être prises en compte pour harmoniser les différents systèmes de valeurs que l'on peut trouver dans le monde. Les technologies modernes nous ont rendus, parfois à notre insu, éminemment observables, voire manipulables. Les risques pour la vie privée et la liberté individuelle existent donc potentiellement.

La réflexion sur l'éthique des machines est également essentielle. Le projet *Moral Machine* du MIT est un premier exemple intéressant. L'irruption de l'IA dans le quotidien soulève des questions importantes relatives :

- à la nuisance des systèmes d'IA (*cf.* lois d'Asimov pour les robots. Des travaux récents portent actuellement sur la conception de robots éthiques, capables d'évaluer les conséquences de leurs actions),
- au statut moral des machines,
- à la détermination de responsabilités,
- aux propriétés requises d'un système du fait de son rôle social et médical potentiel : prédictibilité, transparence à l'inspection…

Il importe donc de définir ou de rappeler des règles éthiques et de mettre en place une charte déontologique pour cadrer aussi bien la recherche scientifique dans ce domaine, qu'elle soit publique ou privée, que les développements industriels. Les enjeux économiques et politiques rendent hélas très difficile la recherche d'un consensus mondial sur ce point. Les dérives possibles, évoquées au par. 6, d'utilisation de la reconnaissance de visages à des fins de surveillance des personnes en est un exemple parmi d'autres. Nous avons mentionné au par. 2 le RGPD européen relatif aux données personnelles. Ce règlement ne doit pas être une contrainte trop forte pour les sociétés européennes, mais il constitue une protection pour le citoyen. Son rôle est triple :

- renforcer les droits des personnes**,**
- responsabiliser les acteurs traitant des données,
- crédibiliser la régulation grâce à une coopération renforcée entre les autorités de protection des données.

Le Japon a rejoint l'Europe sur la protection des données et la Californie envisage une loi similaire dans un futur proche. Sur le même plan, Microsoft, Google, IBM, Amazon et Apple ont créé le *Partnership on AI* dont le but est de « *faire avancer la compréhension du public sur l'intelligence artificielle et de définir les meilleures pratiques sur les défis et les opportunités dans ce domaine* ». L'Europe joue un grand rôle sur le plan de l'éthique. Outre

le RGPD, l'UE a mis en place le *AI HLEG (High Level Group on Artificial Intelligence*) qui a publié un premier document sur les lignes directrices pour une IA de confiance (*Trustworthy AI*), utile et au service de l'être l'humain [46].

L'évolution des systèmes vers une IA de plus en plus forte pose par ailleurs la question de la conscience de ces systèmes. On considère généralement que la conscience est le propre d'un organisme vivant, par sa connaissance de lui-même et de son environnement. Bien que des modèles de conscience artificielle aient été proposés, aucun système d'IA n'est pour l'instant doté de conscience. Les agents conversationnels courants sur Internet et par téléphone, ou de nouvelles générations en cours de conception, sont capables de simuler des émotions. Ils peuvent aussi dans une certaine mesure détecter l'état émotif de l'humain (par les gestes, le visage, la voix), répondre en simulant une émotion (*cf.* le robot Pepper d'Aldebaran), voire influer sur l'état émotif de l'interlocuteur. En revanche, ils ne peuvent en aucun cas pour l'instant ressentir des émotions.

Nous manquons de recul pour examiner sereinement l'évolution de l'IA, et plus généralement de notre société, car les échelles de temps sont très courtes. Il ne s'agit pas de sacraliser la technique, mais simplement d'éviter des peurs irraisonnées et de ne pas oublier que l'IA permet de sauver des vies, d'améliorer nos conditions de vie et de travail ou de mieux comprendre ce que nous sommes.

Il faut donc sensibiliser, former et informer, sans occulter les risques réels, tout en combattant les utilisations malveillantes aussi bien que les craintes non fondées, de façon à établir autant que faire se peut des relations de confiance avec l'IA. Les risques peuvent et doivent être identifiés, anticipés et maîtrisés. Le rôle du monde politique et de la loi, tant pour éviter les dérives « à la Orwell » que pour codifier les utilisations frauduleuses (commerciales, militaires...), est également très important. C'est à ces conditions que l'IA, maîtrisée et démythifiée, pourra apporter tous ses bénéfices potentiels à l'humanité.

# 8. Références


[1] J.-P. Haton et M.-C. Haton, *L'intelligence artificielle,* Coll. Que-sais-je?, PUF, 3e éd., 1993.

[2] J.-P. Haton *et al. Le raisonnement en intelligence artificielle - Modèles, techniques et architectures pour les systèmes à bases de connaissances*, InterEditions, 1991.

[3] L. Rabiner and B.H. Huang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[4] A. Turing, "Computing Machinery and Intelligence", *Mind*, 49, pp. 433-460.

[5] F. Rosenblatt, "The Perceptron: a probabilistic model for information storage and organization in the brain", *Psychological review*, vol. 65, 6, pp. 386-408, 1958.

[6] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry. MIT Press,* 1969.

[7] D.E. Rumelhart *et al.*, *Parallel Distributed Processing*, vol. 1 and 2, MIT Press, 1988.

[8] E.H. Shortliffe, *Computer-based Medical Consultation: MYCIN*, Elsevier, 1976.

[9] T.M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[10] E.A. Feigenbaum and B. G. Buchanan, "DENDRAL and Meta-DENDRAL: roots of knowledge systems and expert system applications", *Artificial Intelligence*, 59, pp. 233-240, 1993.

[11] L. Breiman *et al.*, *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.

[12] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.

[13] I.J. Goodfellow *et al.*, "Generative Adversarial Networks", *Advances in Neural Information Processing Systems 27*, 2014.

[14] C.J. Watkins and P. Dayan, P. (1992), "Q-learning", *Machine Learning*, 8, pp. 279-292, 1992.

[15] R. Al-Rfou *et al.*, "Theano: A Python framework for fast computation of mathematical expressions", arXiv, 2016.

[16] G. Hinton *et al.,* "A fast learning algorithm for deep belief nets", *Neural Comp*. 18, pp. 1527–1554, 2006.

[17] L. Espeholt *et al.*, "IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner", *DeepMind Publication*, 2018.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, 8, pp. 1735–1780, 1997.

[19] Y. Le Cun *et al*., "Gradient-based learning applied to document recognition", *Proc. IEEE*, vol. 86, 11, 1998.

[20] A. Newell, "The Knowledge Level", *Artificial Intelligence*, vol. 18, n°1, pp.87-127, 1982.

[21] L. Bottou, "From machine learning to machine reasoning", *Machine Learning*, vol. 94, pp.133-149, 2014.

[22] F. Alexandre, "Les *spiking neurons* : une leçon de la biologie pour le codage et le traitement des données", F. Ghorbel, J-P. Haton et L. Ben Youssef, editors, *Traitement et Analyse de l'Information : Méthodes et Applications* - TAIMA, 2005.

[23] A. Krizhevsky *et al*., "ImageNet classification with deep convolutional neural networks", *Proc. Advances in Neural Information Processing Systems*, vol. 25, pp. 1090-1098, 2012.

[24] G. Hinton *et al*., "Deep neural networks for acoustic modeling in speech recognition", *IEEE Signal Processing Magazine*, vol. 29, pp. 82-97, 2012.

[25] Y. Le Cun *et al*., "Deep learning", *Nature*, vol. 521, pp. 436-444, 2015.

[26] D. Lenat and R.V. Guha., *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*, Addison-Wesley, 1990.

[27] A. Newell, J.C. Shaw and H. A. Simon, N. Wiener, "Report on a General Problem Solving Program", *Proc. Int. Conf. on Information Processing*, UNESCO, 1959.

[28] J.-L. Laurière, Un langage et un programme pour énoncer et résoudre des problèmes combinatoires, Thèse de doctorat d'état, Université Pierre-et-Marie-Curie, 1976.

[29] N. Wiener, C*ybernetics or control and communication in the animal and the machine*, Wiley, 1948.

[30] T. Winograd, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. PhD Dissertation, MIT Jan. 1971

[31] J. Weinzenbaum, “ELIZA-A Computer Program For the Study of Natural Language Communication Between Man and Machine”, *Com. ACM*, Jan. 1966.

[32] W.S. McCulloch and W.H. Pitts, “A Logical Calculus of the Ideas Immanent in Nervous Activity”, *Bulletin of Mathematical Biophysics*, 5, pp. 115-133, 1943.

[33] F. Alexandre, Y. Burnod, F. Guyot and J.-P. Haton, "The Cortical Column: A New Processing Unit for Multilayered Networks", *Neural Networks*, 4, pp. 15-25, 1991.

[34 J. Hinton *et al*., “Transforming auto-encoders", *Proc. Int. Conf. on Artificial Neural Networks*, pp. 44–51, Springer, 2011.

[35] M. De Choudhury and E. Kiciman, “Integrating artificial and human intelligence in complex, sensitive problems domains: experiences from mental health”, *AI Magazine*, 39, n°3, pp. 69-80, 2018.

[36] J. Pearl, “Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution”, *Technical Report R-475*, UCLA, 2018.

[37] C. Khatri *et al*., “Alexa Prize - State of the art in conversational AI”, *AI Magazine*, 39, n°3, pp. 40-55, 2018.

[38] J. Ferber, *Les systèmes multi-agents,* InterEditions, 1995.

[39] F. Tacchino *et al*., "An Artificial Neuron Implemented on an Actual Quantum Processor", *arXiv: 1811.02266v1 [quant-ph]*, 2018.

[40] G. Polya, *How to solve it*, Doubleday, 1957.

[41] J-P. Haton, *La parole numérique*, Académie Royale de Belgique, Collection Poche, n° 79, 2016.

[42] J. Pearl, *Causality: models, reasoning, and inference,* Cambridge University Press, 2000.

[43] T.R. Besold *et al.*, “Neural-Symbolic Learning and Reasoning: A Survey and Interpretation”, *arXiv*, *cs.AI*, 2017.

[44] P. Luc *et al.*, “Predicting Deeper into the Future of Semantic Segmentation”, *Int. Conf. on Computer Vision,* 2017.

[45] K. He *et al.*, “Deep residual learning for image recognition”, *arXiv preprint arXiv: 1512.03385*, 2015.

[46] https://ec.europa.eu/digital-single-market/en/news/have-your-say-european-expert-group-seeks-feedback-draft-ethics-guidelines-trustworthy

# TROISIÈME PARTIE

## INTELLIGENCE, CONSCIENCE ET IMPACT DE L'IA SUR LES ACTIVITÉS HUMAINES

## Troisième Partie

## Intelligence, conscience et impacts de l'IA sur les activités humaines

### Présentation

La troisième partie rassemble, par-delà les avancées scientifiques et techniques des deux parties précédentes, différentes réflexions sur l'intelligence, la conscience et l'impact de l'IA sur les activités humaines. Avant la conscience, il y a l'intelligence. Le premier texte expose d'abord une approche statistique des différents termes associés à la notion d'intelligence, pour aboutir par le biais des mots les plus fréquemment associés à une sorte de définition « faible » de l'Intelligence, faible mais consensuelle. L'auteur se pose alors la question de la différence entre Intelligence et Intelligence Artificielle, et propose une approche de cette différence, basée sur des références philosophiques. La conscience a été longtemps considérée comme une exclusivité humaine, mais la conviction grandit - et s'est pour ainsi dire pratiquement établie - que les animaux ont également une forme de conscience, à des degrés dépendant de leur place dans l'arbre phylogénétique. Mais peut-on, nous humains, avoir une expérience de - trouver au fond de nous-même - ce que peuvent être ces consciences animales, sachant que notre expérience immédiate de la conscience est langagière ; c'est à cette question difficile que s'attaque le second texte, en s'appuyant sur de nombreuses références philosophiques. Les deux derniers textes traitent de l'utilisation des robots, et du contrôle de cette utilisation. Le premier se place dans le contexte militaire, mais lui aussi se pose la question de la distinction entre Intelligence et Conscience, des différents types de conscience, de l'utilité et de la possibilité de robots conscients. Le second traite de l'éthique entre Humains et Robots, et du degré d'intelligence voire de conscience – mais l'auteur ne croit pas à cette dernière possibilité-souhaitable pour ces derniers dans de telles interactions.

### Définir l'intelligence

La définition des mots « Intelligence », « Intelligence artificielle » relève de la gageure. Non pas qu'il soit difficile d'en trouver une définition ; au contraire, il y a profusion. Ernesto Di Mauro part ainsi à la recherche de définitions couramment données du mot « intelligence », dans le but d'en extraire les traits les plus fréquents, et d'aboutir ainsi à une formulation commune et consensuelle. Il base son analyse sur 70 définitions de l'intelligence parues dans la littérature anglophone spécialisée. Au terme d'une analyse comparative, il aboutit à rassembler les termes majeurs dans la définition suivante :

*« L'intelligence est la capacité de s'adapter à de nouveaux environnements au service d'un but, d'apprendre et de comprendre. »* Ernesto Di Mauro réfléchit ensuite à ce qui sépare l'Intelligence de l'Intelligence Artificielle. La démarche adoptée s'appuie sur la distinction, que les philosophes stoïciens de l'antiquité ont fait leur[1], entre le *logos endiathètos* et le *logos prophorikos*. Le premier est le discours intérieur, le « dialogue de l'âme avec elle-même » selon l'expression de Platon, s'identifiant à la pensée consciente, telle que l'homme l'expérimente. Le second est le discours externe, le discours proféré par la voix. L'intelligence humaine ressortirait ainsi, dans son essence, du *logos endiathètos.* L'IA se situerait, à contrario, du côté du *logos prophorikos.* Ces distinctions se rattachent en fait aux discussions très anciennes, mais réactualisées par les recherches en IA, sur les rapports entre langage et pensée.

## Une conscience animale ?

L'intelligence et/ou la conscience sont-elles réservées, dans le monde vivant, aux seuls êtres humains ? La réponse à la question a évolué dans le temps entre une refus net de toute forme d'intelligence et de conscience au sein des animaux et une reconnaissance tout aussi globale, largement empreinte d'anthropomorphisme. Conclure est difficile, car d'une part, les animaux ne peuvent pas « rapporter » leurs éventuelles expériences conscientes et d'autre part les humains, dans leur essai de compréhension de ce que peut être une conscience animale, sont bloqués par leur propre expérience. A partir de ce constat Franck Cosson, en s'appuyant sur sa connaissance de l'éthologie, développe son analyse philosophique. Il prend parti en mettant en avant l'idée d'une différence de degré et non de nature entre les diverses potentialités conscientes. Ce principe de continuité l'autorise à affirmer que les humains peuvent accéder, par une introspection bien conduite dans leur propre animalité, à ce que peut vivre un animal, et à théoriser les moyens d'approcher ainsi ce que peut être son niveau de conscience. Franck Cosson poursuit en tentant de cerner ce qu'est ce niveau de conscience primaire, dans ses rapports avec la perception de l'environnement, la perception de son propre corps et de ce qui l'affecte : besoin, plaisir, douleur, ressenti du danger, ressenti de l'effort ; dans ses rapports avec les émotions, les interactions avec les semblables ; dans ses rapports enfin avec le parcours de vie, l'expérience acquise, la mémoire, éléments pouvant conduire à l'émergence d'une « singularité » de chaque individu, base de l'apparition d'une conscience d'un niveau supérieur.

## Conscience artificielle et monde militaire

Un tel sujet peut surprendre ; moins cependant quand on sait la place que prennent les drones, et d'une façon plus large toutes sortes de techniques liées à l'IA, dans l'armement

[1] Cf Curzio Chiesa, Le problème du langage intérieur dans la philosophie antique de Platon à Porphyre. *Histoire Epistémologie du Langage, tome 14, fascicule 2, 1992, (partie d'un numéro thématique « Théories linguistiques et opérations mentales »)*

contemporain. Aussi n'est-il pas étonnant qu'un spécialiste de ces questions, Gérard de Boisboissel, vienne devant nous s'interroger sur ce qu'est la conscience, sur la possibilité ou non de disposer de machines conscientes, et sur l'intérêt qu'il y aurait à le faire dans le contexte militaire. Il commence par bien distinguer intelligence et conscience : s'il existe des machines auxquelles sont attribuées sans trop de discussion le qualificatif d'intelligentes, aucune d'entre elles ne peut être vue, pour l'instant, comme « consciente ». Il examine ensuite les différentes catégories de conscience, en reprenant la typologie séparant conscience mesurée par un état de vigilance et conscience d'accès à telle ou telle information ; il y rajoute la conscience des conséquences de ses choix d'action. Il prend parti en posant qu'on ne pourra pas doter une machine d'aucun de ces types de conscience. Si tant est qu'on puisse cependant un jour le faire, en complément d'algorithmes d'IA, serait-ce vraiment utile ou souhaitable ? Pour réfléchir à cette question G. de Boisboissel analyse les différentes facettes de la conscience (humaine) nécessaires à (et mises en œuvre par) un combattant sur un champ de bataille. Cette analyse lui permet de déterminer quelles seraient les facettes qui pourraient avantageusement être intégrées - lorsqu'elles ne le sont pas déjà - dans des systèmes robotisés. Il termine en posant qu'un programme, aussi évolué soit-il, n'aura jamais l'intuition ou l'instinct d'un être humain. Et qu'une machine, militaire ou non, est et restera (ou devra rester) un outil au service et sous le contrôle de ceux qui l'utilisent.

## Relation émotionnelle entre humains et robots : quelle éthique ?

Le dernier chapitre concerne l'application de l'IA et de la robotique au secteur de l'aide à la personne. Le choix de ce thème résulte d'un constat tout aussi dramatique que le constat sur la course aux armements sous-tendant le thème précédent : l'explosion de la longévité et la rupture, dans la plupart des civilisations, du tissu familial, avec des fins de vie solitaires. C'est, pour beaucoup de personnes, une situation tragique et on n'a rien trouvé de mieux que de fournir aux vieillards des aides, à la fois matérielles et morales : des robots affectueux… Les bénéficiaires en seront-ils comblés ? Avec ces questions en arrière plan, Laurence Devillers aborde ce qu'on appelle le thème de la « robotique émotionnelle », entrant dans le cadre de la relation entre humain et robots. Elle commence par rappeler les avancées, apportées par la psychologie et les neurosciences, sur la compréhension des émotions, de leur rôle, de leur rapport avec la conscience ; elle revient également sur cette dernière notion, ses différentes définitions et composantes. Elle retrace ensuite le parcours de l'IA sous l'angle de son propos, savoir l'intégration des machines dans la société, l'utilisation de plus en plus fréquente de ces machines par les individus, les travaux concernant, dans le cadre de ces interactions, la détection des émotions individuelles par les machines et leur simulation en retour. Elle s'interroge enfin sur les bénéfices et les risques de toutes ces avancées. Doter des robots de simulation émotionnelle ou de capacité d'interprétation est notamment réalisé dans le domaine de la santé. Ils peuvent alors apporter une aide précieuse : jeu avec les enfants malades, surveillance des personnes hospitalisées, intervention en cas de dépression ou d'autisme, aide au diagnostic… Mais d'un autre côté, « les robots sont aussi porteurs de risques importants d'isolement, de déshumanisation et de manipulation des humains ». La co-adaptation humain-machine

s'étend rapidement mais demande réflexion tant la possibilité d'être manipulés par ces engins est grande. Laurence Devillers incite donc à la vigilance et assure que cette co-adaptation « devra sur le long terme être un axe de recherche et de surveillance important dans les prochaines années. »

Pour le comité de lecture de l'AEIS[2]

[2]Gilbert Belaubre, Jean Schmets, Jean-Pierre Treuil

# 11

# A la recherche de définitions

**Ernesto Di Mauro**

Dipartimento di Scienze Ecologiche e Biologiche,
Università della Tuscia, Viterbo, Italie.

## Abstract

The definition of the words "Intelligence", "Artificial Intelligence" is a challenge. Not that it's hard to find a definition; on the contrary, there is a profusion of definitions. Ernesto Di Mauro thus sets out to find commonly given definitions of the word "intelligence", with the aim of extracting the most frequent features, and thus arriving at a common and consensual formulation. He bases his analysis on 70 definitions of intelligence published in specialized English-language literature. At the end of a comparative analysis, he ends up gathering the major terms in the following definition: "Intelligence is the capacity to adapt oneself to new environments to achieve a goal, to learn and to understand". Ernesto Di Mauro then thinks about what separates Intelligence from Artificial Intelligence. The approach adopted is based on the distinction that the Stoic philosophers of antiquity made between the *endiathètos logo*s and the *prophorikos logos*. The first one is inner discourse, the "dialogue of the soul with itself" identifying with conscious thought, as man experiences it. The second is external speech, speech uttered by voice. Human intelligence would thus emerge, in its essence, from the *endiathètos* logos. The AI would, on the contrary, be on the side of the *prophorikos* logos. These distinctions are in fact linked to very old discussions, updated by AI research, on the relationships between language and thought.

La recherche des signatures de la conscience est un problème énorme qui n'a pour le moment aucune solution univoque. Il suffit de lire les titres des chapitres de ce livre pour se rendre compte qu'on est en train de fouiller dans les concepts d'intelligence, de mémoire et de conscience. En lisant le texte de n'importe lequel des auteurs, on s'aperçoit immédiatement que la façon d'entendre ces mots n'est pas toujours la même. N'est pas identique non plus l'importance donnée à chacun de ces phénomènes en tant que composant de l'*esprit*. Le fait que ce mot français se traduise différemment dans la plupart des langues voisines *( mens* en latin, *mente* en italien, espagnol et portugais, *mind* en anglais, etc.) souligne la nécessité d'un effort de formalisation.

Ce colloque traite de ces trois aspects (intelligence, mémoire, conscience) de l'esprit en retrouvant et actualisant une dichotomie ancienne ; dichotomie qu'aujourd'hui on pourrait traduire par la question : *Quelles sont les différences et les points de contact, plus ou moins intimes, entre* ***intelligence*** *et* ***intelligence artificielle*** *?* Peut-on en attendre des définitions ?

## 1. Intelligence

Peut-on définir l'intelligence ? Un premier effort de formalisation pourrait se fonder sur l'analyse structuraliste comparative (à la Lévi-Strauss, pourrait-on dire) de 70 définitions parues dans la littérature internationale spécialisée (en anglais) du mot « intelligence » [1]. Souvent les définitions existantes ressemblent plus à des descriptions qu'à de vraies définitions. Cependant, certains mots sont plus représentés que d'autres. Ce qui permet d'atteindre une définition consensuelle quoique « faible ». La méthode consiste simplement en une analyse grammaticale initiale, un regroupement, et une analyse de fréquence.

*Analyse initiale*. Les termes présents dans les 70 définitions choisies ont été d'abord séparés en noms, adjectifs, verbes, pronoms, adverbes et locutions ; puis regroupés selon un critère grammatical (exemple : les noms avec les adjectifs et les adverbes de même racine, comme *Adequate* et *Adequately*) ; ensuite on insère dans le même groupe les verbes et les participes correspondants (exemple : *to Reason* et *Reasoning*).

*Regroupements sémantiques.* Les termes ont été ensuite regroupés par homogénéité de signification (exemple : *Ability, Capability, Capacity* …) et par ordre alphabétique. Dans chaque groupe le terme le plus fréquent a été choisi comme représentatif (exemple: *Ability*, plus fréquent que *Capability*, etc.). Les termes présents moins de 10 fois n'ont pas été considérés. Les autres ont été organisés dans la phrase de définition finale en fonction de leur fréquence. Je reporte ci-dessous une partie de l'analyse :

***Ability****, abilities,* présent **44** fois dans les 70 définitions. **/** ***Capability (-ies), Capacity (-ies),*** présent **21** fois sur 70. **/** *Faculty (of),* présent 3 fois. Au total, en sommant la fréquence de ces mots regroupés par ressemblance de signification, on trouve que les termes corrélés à "*Ability*" sont présents **68** fois sur 70. Entre les différents mots du groupe "*ability, capability-capacity, faculty of "* on choisit "***Ability***" parce que le plus fréquent.

*Abstraction, Abstractness* 3*./ Abstract, abstractly* 5./ Total 8. // Non choisi parce que présent moins de 10 fois.

***to Act*** 3./ *Action* 1./ *Activity (-ies)* 2./ *Agent* 1./ *to Deal with* 3./ *to Do* 1./ *to Engage* 1./ Total **12. //** Ce regroupement est sélectionné car il totalise plus de 10 présences et, à l'intérieur du groupe, on choisit "**to act**" parce que le plus fréquent.

*Achievement* 1./ *to Achieve* 3/ *Advancement* 1./ *to Carry on* 2./ *Performance* 3./ Total 10.//

***to Adapt,*** *to Adjust* **12**./ *Adaptation, adaptness* 4./ *Adaptive* 3./ Adjustment 4./ *Aptitude* 1./ *Accommodation* 1./ *Adequate, adequately* 2./ *Adaptive* 3./ *Appropriate, appropriately* 5./ Total **35.**

*to* ***Understand, understanding*** **9**./*Alertness* 1./ *Attention (span of)* 1./ *Aptitude* 1./ *to Comprehend* 4./ *to Realize* 1./ Total **14.**

Autres exemples de regroupement, plus avant dans la liste :

*to Create* 1./ *to Generate* 1./ *to Demand* 1./ *to Differ* 1./ *to Enable* 1./ *to Encompass* 1./ *to Exercise* 1./ *to Enter* 1./ *to Find (one self in)* 1./ *to Function, functioning* 1./ *to Generalize* 1./ *to Give, given* 1./ *to Grasp* 1./ *to Have* 1./ *to Incorporate, incorporating* 2./ *to Include* 1./ *to Inhibit, inhibited* 1./ *to Increase* 1./ *to Involve* 1./

***to Learn, learning, learned*** **22**./ ***Knowledge*** **10**./ Total **32.**

A coté de pages dans lesquelles on rencontre des mots répétés et en quelque sorte partagés, on rencontre de longues listes de mots présents très rarement, ou seulement une fois. Par exemple : *Experience, experienced* 1./ *Extent (to which)* 1./ *Fact(s)* 3./ *Things* 1./ *Concrete*1./ *Facility* 1./ *Factor* 1./ *Flexibility* 1./ *Forces* 1./ *Form(s)* 1./ *Formation* 1./ *Framework* 1./ *Function(s)* 1./ *Get (better)* 1./ *Ideas* 3./ *Imagination, imaginally* 4./ *Impulse(s)* 1./ *Industry* 1./ *Information* 1./ *Initiative* 1./ *Intelligence test* 1./ *Interest* 1.

On poursuit cet inventaire alphabétique, avec beaucoup de mots dispersés, apparemment dépourvus d'un apport réel à la définition de l'intelligence, jusqu'au W de *Will.*

On arrive ainsi à la définition consensuelle : [X/70] **Intelligence [70] is the mental [17] ability [68] to adapt [35] to new [12] environment(s) [34] for a goal [24], to learn [32], and to understand [14].** L'intelligence est la capacité mentale de s'adapter à de nouveaux environnements au service d'un but, la capacité d'apprendre et de comprendre.

Au delà de cette définition consensuelle, et en considérant toutes les 70 définitions spécifiques formulées pour "intelligence", on note au fil des listes de fréquence une grande dispersion des termes employés. Les mots rencontrés dans les définitions de l'intelligence sont tellement différents, dispersés et diffus qu'ils couvrent un très grand nombre

d'activités humaines, comme on peut s'y attendre. Les définitions existantes sont souvent relativement informelles (ce qui laisse penser que le sujet n'a pas encore atteint un niveau suffisant de maturité). Ce qui est plus intéressant est peut être ce qui manque: les termes "mémoire", "conscience", créativité sont incroyablement absents ou très rares.

## 2. Intelligence artificielle

La définition courante de intelligence artificielle est : l'IA *est une discipline qui appartient à l'informatique. Elle étudie les fondations, les méthodologies et les techniques qui permettent la réalisation de systèmes hardware et la programmation de systèmes software qui fournissent à l'ordinateur les fonctions qu'un observateur humain pourrait attribuer exclusivement à l'intelligence humaine.*

Plus en détail, on pourrait formuler des définitions spécifiques de l'IA en focalisant sur les processus de raisonnement intérieur, ou sur le comportement extérieur d'un système intelligent en utilisant comme mesure d'efficacité la ressemblance avec le comportement humain, ou la ressemblance avec un comportement idéal, défini comme "rationnel".

L'intelligence artificielle est donc définie par rapport à l'intelligence humaine. Cette dichotomie n'est pas nouvelle. Montaigne (II, 2, *Apologie*) la remonte aux Stoïciens. Plus précisément à Philon, qui avait introduit les termes : *Logos prophorikòs*, discours proféré, modus inveniendi et proferendi, et *Logos endiathètos*, discours intérieur, modus intelligendi [et on revient donc là à notre discours exactement dans les mêmes termes de *intelligence*]. Le logos prophorikòs est l'intelligence artificielle, le modus inveniendi et proferendi est l'intelligence, celle qui appartient aussi aux animaux (et aux ordinateurs). Le logos endiathètos, le discours intérieur, le modus intelligendi, correspond à la conscience, fonction typiquement humaine, point de référence de tout discours, au moins jusqu'à présent.

Dans un autre domaine (tout en restant dans la sphère biologique), on pourrait se confronter avec une autre dichotomie similaire : celle existant entre Génétique et Epigénétique ; la Génétique, qui constitue le hardware de base, les gènes, le mécanisme, l'intelligence mécanique de l'ordinateur ; tandis que l'Epigénétique est le vécu, les connexions (*inter-legere*, encore l'intelligence) entre les gènes et leurs fonctions. La Génétique est prophoriké, l'Épigénétique est endiàthike.

La dichotomie est donc toujours la même. Historiquement elle a eu une grande importance parce qu'on voyait là une possible solution au problème du libre arbitre: "*Celui qui est seulement prophorikòs n'est pas responsable, celui qui est endiathètos oui*". Jusqu'au moment où David Hume a mis en lumière que le libre arbitre est une invention des théologiens pour résoudre leurs problèmes de logique à eux. Hume identifie (qu'on me passe la simplification brutale) *conscience* et *volonté*, et définit avec précision uniquement cette dernière :

**Will**: " … that by the will, *I mean nothing but the internal impression we feel and are **conscious of,** when we **knowingly** give rise to any **new** motion of our body, or new perception of our **mind** ". D. Hume. A Treatise of Human Nature II, 3, 1* (1740) (v. aussi *Enquiry concerning Human Understanding,* 1748). Le thème de l'ouvrage et du colloque qui en a été la source est, selon moi, essentiellement la recherche d'une possible solution de cette dichotomie. dont Hume paraît s'être bien approché.

## Références

**[**1**]** Shane Legg and Markus Hutter, *A collection of Definitions of Intelligence*, 2017, IDSIA-07-07 Technical Report.

# 12 | Recherches sur l'apparition des niveaux de conscience chez l'animal

**Franck Cosson**

Docteur en Philosophie, Université de Lorraine

« I would even suggest that the greatest riddle of cosmology may well be neither the original big bang, nor the problem why there is something rather than nothing (…), but that the universe is, in a sense, creative : that it creates life, and from it mind - our consciousness - which illuminates the universe, and which is creative in its turn. » (Karl Popper, The Self and its Brain. An Argument for Interactionnism, Springer international, p. 61)

« L'individualité psychologique se surimpose à l'individualité biologique sans la détruire, car la réalité spirituelle ne peut être créée par une simple négation du vital. » (Gilbert Simondon, L'individuation à la lumière des notions de forme et d'information, Millon, « Collection Krisis », p. 276.)

« La réciproque affinité nous façonne. » (Edmond Jabès, Le parcours, Gallimard, p. 16.)

## Abstract

Are intelligence and / or consciousness reserved, in the living world, only for human beings? The answer to the question has evolved over time between a clear rejection of all forms of intelligence and consciousness within animals and an equally global recognition, largely marked by anthropomorphism. Concluding is difficult, because on the one hand, animals cannot "relate" their possible conscious experiences and on the other hand humans, in their attempt to understand what animal consciousness can be, are blocked by their own experience. From this observation Franck Cosson, drawing on his knowledge of ethology, develops his philosophical analysis. He takes sides by putting forward the idea of a difference in degree and not in nature between the various conscious potentialities. This principle of continuity allows him to assert that humans can have access, by a well-conducted introspection in their own animality, to what an animal can experience, and to theorize the means of approaching thus what can be its level of consciousness. Franck Cosson goes on to try to define what this primary level of consciousness is, in its relationships with the perception of the environment, the perception of its own body and what affects it; in its relationships with emotions, interactions with its fellows; finally in its relationships with the life course, the acquired experience, memory, elements which can lead to the emergence of a "singularity" of each individual, the basis of the appearance of a consciousness of a higher level.

# 1. Introduction

Auguste Comte dénonçait, dès 1838, la responsabilité des métaphysiciens qui imaginèrent « une vaine démarcation » « entre l'homme et les animaux », s'efforçant de « (…) conserver, par un principe unique ou du moins souverain, ce qu'ils ont appelé l'unité du moi, afin de correspondre à la rigoureuse unité de l'âme. »[1] Des capacités cognitives démultipliées par le langage permettent en effet à notre espèce l'énonciation pure et « métaphysique » de l'ego cogito par lequel la pensée du sujet s'énonce selon l'ordre d'une perfection intelligible indépendante, *comme conscience pure et abstraite*, de ses conditions matérielles d'énonciation et de tout rapport à une quelconque réalité autre que pensée. Cette conscience existe comme conscience de soi, à travers un acte différencié révélant le sujet à lui-même comme être unique, identique et stable, lui permettant d'accéder, sur le mode d'une connaissance intellectuelle, à l'ensemble de ses opérations ainsi qu'à ses principaux attributs.

Dans un tel contexte, la principale difficulté, d'ordre phénoménologique et épistémologique, vient de ce que notre expérience consciente masque la possibilité pour d'autres formes d'expériences de conscience ou d'états conscients d'exister selon des modalités qui leur sont propres. Péchant alors par simplification et homogénéisation des données, en assimilant l'altérité animale à ce que nous sommes et expérimentons nous-mêmes, deux tendances contraires nous animent. Soit en effet nous nions que les animaux puissent disposer d'une forme quelconque de vie consciente soit, prenant le contre-pied de cette première tendance pour des raisons tout aussi fallacieuses, nous nous laissons aller au travers anthropomorphe et assimilons alors à l'excès l'inconnu au connu et l'animalité à l'humanité. Dans les deux cas évoqués, nous refusons de nous confronter à ce que l'altérité relative des animaux peut avoir de problématique pour nos investigations. Or, nous savons aujourd'hui que tout état conscient renvoie − sans nécessairement s'y réduire − à la convergence de fonctions biologiques interagissantes qui culminent dans une sensation ou un sentiment global, signe du fonctionnement de l'organisme considéré comme totalité qui s'éprouve dans un rapport vécu à l'environnement. C'est alors seulement, comme nous le verrons, que peut être envisagée la question d'un être − sujet de ce qui l'affecte − susceptible d'accéder à une notion de ce qu'il est en tant qu'il vit et agit. Nous retiendrons ainsi dès à présent que, si l'animal n'a pas nécessairement conscience de ce qu'il est ni de qui il est, il sait vraisemblablement qu'il est et que cette connaissance première constitue un vecteur pour poser dans toute sa complexité la question de l'apparition des divers niveaux de conscience ainsi que de leur différenciation et des propriétés respectives qu'il faut leur reconnaître.

Dernière remarque introductive en forme de préalable méthodologique. S'interroger sur ce qu'éprouvent et vivent les animaux implique une part d'interprétation dont on pourrait craindre qu'elle nuise à la rigueur des investigations engagées sur un thème aussi philosophiquement connoté et difficile à traiter − notamment sur le plan des redoutables obstacles épistémologiques qu'il comporte − que celui de la conscience. Nous rappellerons

1 Comte A., 1975 (1re édition 1838), Cours de philosophie positive, 45e leçon, Hermann, p. 856

ici l'interrogation d'Elisabeth de Fontenay formulée dès 1998 : « Un malaise, de toute façon, persiste écrivait-elle alors : de quel droit s'autorisera-t-on pour rendre justice à ceux qui se taisent ? A quel titre, se mettant à leur place, les fera-t-on parler ? »[2] Or, *un principe de continuité associé à l'idée d'une différence de degré et non de nature entre les diverses potentialités conscientes*, implique d'adopter une méthode d'investigation fondée sur la possibilité d'inférer certains états éprouvés par les animaux en combinant des données d'expérience qui divergent quant à leur mode de manifestation et à la manière dont nous pouvons y accéder mais qui convergent, chez l'homme et les animaux, du point de vue de leur origine biologique et organique.

Cette méthodologie « mixte » est fondée sur une analyse phénoménologique rapportée à ce que nous éprouvons de nos propres états « animaux » − c'est-à-dire corporels et sensibles, vécus en intériorité − associée à une approche objectivée et circonstanciée des animaux vivants considérés dans leur environnement en exploitant, dans ce dernier cas, les observations et les connaissances acquises relevant de la biologie, de l'éthologie, des systèmes sensoriels en jeu ou de données neurophysiologiques accessibles. L'analyse combine ainsi notre expérience d'états internes subjectivement « vécus » et l'observation de corps animaux « étrangers » et séparés, extériorisant des comportements dotés de propriétés que nous pouvons à la fois « comprendre » − sur la base d'une signification partagée et accessible qui, dans une certaine mesure, les rend intelligibles − et qui peuvent, concomitamment, être interprétés sur la base de la connaissance des données biologiques et éthologiques déjà évoquées. Certes, incapables de décrire l'expérience d'un animal « en original » nous ne savons pas, pour reprendre la formule célèbre de Thomas Nagel, « quel effet cela fait » d'être ce qu'il est[3]. L'objectivation approximative de cette expérience n'est pas inaccessible pour autant si nous disposons d'une connaissance du système perceptif de l'animal et pouvons établir *des corrélations* entre les conduites observées et ce qui se déroule dans l'organisme en termes de pulsions, de sensations, de tendances, et de motivations. Nous pourrons alors imaginer quel effet cela fait d'être un animal quelconque en rapportant ses réactions, mouvements, comportements, attitudes et conduites à des états qui pourraient lui correspondre en prenant en compte ce qu'il est susceptible de percevoir ou non. A un moment ou à un autre il faut bien que, par imagination et réflexion, nous nous mettions à la place de l'animal. Le regard d'un épistémologue du vivant éclairera notre propos : « Nous pensons quant à nous, écrit Georges Canguilhem, qu'un rationalisme raisonnable doit savoir reconnaître ses limites et intégrer ses conditions d'exercice. L'intelligence ne peut s'appliquer à la vie qu'en reconnaissant l'originalité de la vie. La pensée du vivant doit tenir du vivant l'idée du vivant. "Il est évident que pour le biologiste, dit Goldstein, quelle que soit l'importance de la méthode analytique dans ses recherches, la connaissance naïve, celle qui accepte simplement le donné, est le fondement principal de sa connaissance véritable et lui permet de pénétrer le sens des événements de la nature". Nous soupçonnons que, pour faire des mathématiques, il nous suffirait d'être anges, mais

2 De Fontenay (E.), 1998, Le silence des bêtes. La philosophie à l'épreuve de l'animalité, Fayard, p. 21.

3 « Pour autant que je pourrais avoir l'apparence extérieure d'une guêpe et me comporter comme elle, ou comme une chauve-souris, sans changer ma structure fondamentale, mes expériences ne ressembleraient en rien à celles de ces animaux. D'un autre côté, il est douteux que l'on puisse attacher une signification quelconque à la supposition que je pourrais posséder la constitution neurophysiologique d'une chauve-souris. Même si je pouvais par degrés successifs être transformé en chauve-souris, rien dans ma constitution présente ne me permet d'imaginer ce à quoi ressembleraient les expériences d'une telle incarnation future de moi-même ainsi métamorphosé. » (Nagel (T.), « Quel effet cela fait d'être une chauve-souris ? » dans Hofstadler (D.) et Dennet (M.), 1987, Vues de l'esprit. Fantaisies et réflexions sur l'être et l'âme, Interédition, p. 395).

pour faire de la biologie, même avec l'aide de l'intelligence, nous avons besoin parfois de nous sentir bêtes.»[4]

Pour appréhender cette expérience, nous mettrons à profit la méthode « hétérophénoménologique » proposée par le philosophe Daniel C. Dennet pour tenter de connaître ce qu'un autre sujet éprouve grâce à des descriptions mettant en évidence « (…) un portrait neutre et exact de l'*effet que cela fait* d'être ce sujet − dans les termes mêmes du sujet, à partir de la meilleure interprétation possible en la circonstance. »[5] Cette méthode « implique l'extraction et la purification des *textes* venant de *sujets* (…) parlants, et l'utilisation de ces textes pour engendrer une fiction théorique, le *monde hétérophénoménologique* du sujet. Ce monde fictionnel est peuplé de toutes les images, évènements, sons, odeurs, intuitions, pressentiments et sensations que le sujet croit (…) sincèrement exister dans son flux de conscience.»[6] Si la capacité de décrire les états vécus sur la base objectivée d'une retranscription par l'écriture manquent pour les animaux, il reste possible selon nous de leur transposer cette méthode en acceptant un travail d'interprétation et de relativisation de notre propre point de vue. Objective d'un côté, elle s'appuie sur la connaissance de la vie des animaux, de leurs systèmes perceptifs et des préférences qu'ils induisent ; subjective de l'autre, elle utilisera des capacités d'interprétations applicables sur la base de grandes catégories d'actions signifiantes à forte implication comportementale que nous avons en commun avec la plupart des animaux comme la répulsion, la fuite, l'attirance, le rapprochement, le plaisir et la douleur qui constituent généralement des signes interprétables de crainte, de désagrément, de désirs ou de satisfactions qui, s'ils sont vécus différemment, renvoient à des expériences comparables. Sans de telles expériences notre relation aux animaux serait faite d'incommunicabilité radicale et d'une ignorance nous condamnant à un solipsisme anthropologique allant de pair avec un monde singulièrement appauvri où l'homme, isolé et souverain, culminerait dans l'affirmation métaphysique d'une conscience exclusive et abstraite.

## 2. Phénomènes organiques et connaissance corporelle. Une conscience incorporée

### 2.1 Apercevoir sa propre existence. Phénoménologie de l'animalité en soi

Nous partirons, pour amorcer cette réflexion, de notre propre animalité, c'est-à-dire de *l'animalité phénoménale* que nous connaissons le mieux par expérience directe. S'efforçant d'être attentif aux états corporels qui l'affectent en suspendant son jugement, sans rien connaître de ses organes et en se gardant de projeter des connaissances sur ce qu'il ressent, l'homme se trouve en présence de données qui constituent une expérience « animale » entendue au sens d'une vie organique, partie constituante de son être corporel et éprouvée à travers divers états. Cette expérience pré-linguistique − « antéprédicative » comme le dit Husserl dans *Expérience et jugement* −, ne comportant ni identification conceptuelle ni dénomination, est fondée sur une forme d'indifférenciation constitutive − trame sensible de tout être auto-affecté.

4 Canguilhem (G.), 1985, La connaissance de la vie, Vrin, « Problèmes et controverses », pp. 12/13.

5 Dennet (D.C.), 1993, La conscience expliquée, Odile Jacob, « philosophie », p. 130.

6 Ibid.

La connaissance porte alors sur des sensations globales, le sujet n'identifiant pas d'organe ou de partie corporelle comme tels. Il s'agit d'accéder originairement, grâce à une phénoménologie de l'être manifesté dans l'expérience du corps organique tel qu'il est appréhendé, au « sentiment de ce qui est » et à la manière dont nous en sommes affectés. Cette démarche rejoint la préconisation du neurophysiologiste Antonio Damasio qui nous demande de regarder « dans notre esprit conscient » pour observer à quoi il ressemble « à la base des couches qui le composent et débarrassé de tout le bagage que lui apporte l'identité, le passé vécu, l'anticipation du futur, l'esprit conscient.»[7] Dans cette expérience préconsciente, spontanée et presque immédiate, les objets qui nous affectent relèvent d'expériences qui nous mettent au contact de strates d'une vie primordiale débarrassée de toute référence renvoyant à une conceptualisation ou à une énonciation. « Certains [objets] sont ainsi placés, explique-t-il, dans une certaine perspective par rapport au moi matériel que (…) je peux localiser non seulement dans mon corps mais, plus précisément, dans un espace situé derrière mes yeux et entre mes oreilles. Fait tout aussi remarquable poursuit-il, certains objets au moins s'accompagnent d'un sentiment qui les relie sans ambiguïté à mon corps et à mon esprit. Et ce sentiment m'apprend – sans qu'aucun mot ne soit prononcé – que je possède ces objets, pour une certaine durée et que je peux agir sur eux si je le souhaite. C'est littéralement le " sentiment de ce qui est " (…). Un sentiment plus profond se dessine et se manifeste dans les profondeurs de l'esprit conscient. C'est le sentiment que mon corps existe et est présent, indépendamment de tout objet avec lequel il interagit, tel un roc solide, telle l'affirmation brute que je suis vivant (…). Ce sentiment fondamental (…) me semble (…) être un élément essentiel du processus du soi. Je l'appelle *sentiment primordial* (…). »[8] Plonger ainsi au cœur de notre propre animalité nous permet de comprendre ce qu'un animal peut lui aussi éprouver quant au sentiment primordial de sa propre existence, sentiment qui prend la forme, concrète et éprouvée, de son propre corps en tant qu'il l'affecte et le fait accéder à une réalité qui est, pour lui, à la fois ultime et première. Ce sentiment primordial, d'un point de vue qualitatif et tel qu'il est vécu, permet en outre d'éprouver un état interne global et d'en prendre connaissance sans que celle-ci ne soit mise en forme par un acte distinct de prise de conscience qui entrainerait une dualité du corps et du sujet pensant.

Cette connaissance de soi ou, plus exactement, de *ce qui est soi en tant que vécu à partir d'une expérience qui affecte le vivant*, s'accomplit sur la base d'une réalité matérielle et sensible. Celle-ci renvoie à une *sensation globale indistincte manifestée comme présence au corps propre d'une vie autonome accaparante et irrépressible dans sa manifestation.* L'animal accède à ce qui se produit en lui, à ce que lui « enseigne » la nature ; il éprouve alors les manifestations phénoménales de sa propre vitalité avec des régularités, des pulsations, et surtout des sensations qui lui sont propres. Celles-ci peuvent être localisées différenciées sous formes de plaisirs et de douleurs ou globales et indifférenciés comme dans l'effort exigé par la mobilité du corps ou l'effort organique involontaire impliqué par exemple dans les phénomènes de digestion ou de respiration. L'ensemble des fonctions vitales en activité et une physiologie continuellement « vécue » permettent de parler de la présence d'états organiques qui font « corps » avec l'animal[9]. Or, l'une des grandes fonctions biologiques du phénomène conscient apparaît dans la nécessité pour l'organisme

7 Damasio (A.), L'autre moi-même. Les nouvelles cartes du cerveau, de la conscience et des émotions, Odile Jacob, 2010, p. 226.

8 Ibid. pp. 226/227.

9 Damasio (A.), Ibid., p. 223.

d'accéder à la multiplicité de ces états sous forme unifiée et opérationnelle, prédisposant ainsi à la réaction adéquate ou à l'action coordonnée.

### 2.2 Sensation, « intelligence organique » et émergence d'états conscients

Nous partirons des analyses profondes et éclairantes de Maurice Pradines : « La sensation, écrit-il, est une opération de l'intelligence en ce qu'elle est toujours interprétation d'un signe (…). La sensation est donc une impression interprétée en matière et en extension. Toute sensation est *in-tuition* au sens étymologique du terme, c'est-à-dire appréhension d'une chose à travers un médium interposé. »[10] Le processus organique apparaît travaillé par une dualité constituante qui tend à se résorber dans son propre approfondissement : « La sensation (…) apporte ainsi une connaissance qui *pénètre* ce qu'elle connaît et consiste justement à *s'en pénétrer*. La saveur, l'odeur, toutes les prénotions sexuelles ne peuvent ainsi devenir des anticipations conscientes d'objets complémentaires sans nous fournir sur la nature de ces objets une lumière spécifique qui est comme l'éclairement des appétitions mêmes qui nous portent vers eux. C'est une connaissance paradoxale des choses *extérieures* par leur *intérieur*, qui est la nôtre. »[11] Pour autant, la présence d'une information ne correspond pas nécessairement à son objectivation. C'est le cas notamment de « (…) l'activité d'appropriation alimentaire ou sexuelle. Par nature, elle tend à l'appropriation de quelque objet complémentaire ; sa source est donc dans l'être vivant lui-même (…).»[12] De fait, l'activité appropriative trouve sa source « dans une sensation anticipée, une espèce de prénotion organique. »[13]

Ainsi, nous assimilons trop vite une sensation à un objet auquel elle renvoie et dont nous serions conscients. Chez l'animal, c'est l'opération de l'intelligence elle-même qui sert de vecteur à une forme de conscience inséparable du travail d'interprétation qui permet de mettre en évidence les significations immanentes au sensible. Une conscience sensible qui se manifeste dans la sensation apparaît confuse en raison de la mixité qui rend les phénomènes indistincts en termes de connaissance. C'est pourquoi « les odeurs et les saveurs nous font connaître, non pas ce qui est vraiment *distinct* de nous dans les choses, comme le sens tactile (…) mais, au contraire, ce qui nous est complémentaire, et, par conséquent, congénère, donc un excitant qui est encore *nous*. Il ne peut donc être parfaitement *objectif*. Le sujet ici pénètre l'objet et, en conséquence, ne nous en laisse prendre qu'une fausse idée. (…) La saveur et même l'odeur, comme liées au besoin, ne nous font connaître des choses que ce qui nous *convient* plus ou moins, et, de ces choses mêmes, que leurs éléments apparentés à notre propre nature.»[14] On conçoit aisément en effet que l'animal, en lien immémorial avec des propriétés de l'environnement intégrées à son organisme sous formes d'adaptations, reste sous l'emprise d'une histoire évolutive qui modèle ses organes et ses comportements. Il porte et assume les contraintes et nécessités du monde extérieur au titre de sa condition biologique et accomplit d'immémoriales adaptations qui donnent sens et densité à ses actions. C'est pourquoi, « L'individu (…) naît obligé à des relations extraordinairement intimes avec des choses dont il n'a encore aucune *expérience*. C'est que (…) ces choses font partie de sa nature même, de sorte

10 Pradines (M.), 1932, Philosophie de la sensation, 2, la sensibilité élémentaire (les sens primaires), 1, Les sens du besoin, Les belles lettres, « Publication de la faculté des lettres de l'université de Strasbourg », Fascicule 61, p. 1.

11 Pradines (M.), 1954, L'aventure de l'esprit dans les espèces, Flammarion, , p. 157.

12 Pradine s (M.), 1946, Traité de psychologie générale,1, Le psychisme élémentaire, PUF, « Logos. Introduction aux études philosophiques », p. 294.

13 Ibid.

14 Ibid., p. 500.

qu'il les connaît déjà simplement en apprenant à se connaître dans l'exercice de ses activités.»[15] Cette tendance encore obscure de la sensation vitale s'accompagne d'une forme d'intelligibilité dont la raison d'être est précisément la constitution organique en tant qu'elle s'active dans une progression interprétative qui l'éclaire sur son objet. C'est pourquoi, « (…) comme tout sens est *représentation*, signification, le besoin en a tiré une lumière, c'est-à-dire une conscience (…). Le besoin, puisqu'il devenait un sens, c'est-à-dire une intellection, s'éclairait pour un esprit, et l'esprit, en affinant la stimulation dont sortait le sens la rendait plus forte et plus impérieuse. »[16] De même, le sens psychologique qui s'accompagne d'une connaissance en prise sur les significations visées et exprimées est lié à ces notions originaires intégrées au processus de la sensation animale. Il existe alors, ajoute Pradines des « aspirations conscientes, munies de prénotion et d'orientation, qui constituent un besoin *psychologique*. »[17]

D'une manière plus générale, nous dirons qu'une appréhension de soi comme corps affecté surgit alors au cœur de la relation que l'animal entretient avec le monde physique.

Le sentiment éprouvé est d'autant plus prégnant qu'il en trouve l'écho dans l'expression du besoin qui prend sa source dans la physiologie de l'organisme. Ce besoin lui fait connaitre une présence à soi immédiate c'est-à-dire une affinité avec sa propre constitution organique telle qu'elle lui apparaît dans l'irréductibilité d'une expérience *qui le fait naître à ce qu'il est*, amorçant ainsi un phénomène de prise de connaissance de soi. On peut imaginer que ce dernier marque déjà une forme originaire et de distanciation à l'égard d'une vitalité qui n'est plus considérée, comme chez le végétal, dans sa seule fonctionnalité organique. Odeur et saveur en particulier révèlent au vivant non pas « (…) la présence de l'objet au sens banal du terme, c'est-à-dire de choses ou d'êtres étrangers à sa nature propre (indifférentes ou hostiles), mais celles de choses ou d'êtres complémentaires ou appropriables (aliment ou partenaire sexuel).»[18] Ainsi, ces phénomènes s'accompagnent toujours d'une « visée » liée à la représentation encore confuse et indéterminée d'une fin. Le sens du tact a une importance particulière dans l'émergence de phénomènes de conscience car, en introduisant un « jeu » dans et à l'égard de la matière, il rend possible une libération de l'esprit comme le montre Pradines : « Le toucher écrit-il dans *L'aventure de l'esprit dans la matière*, est (…) une parade (…). La menace est connue comme distante, si minime que soit la distance (*connaissance d'espace*) ; et cette distance en mesure l'imminence (*connaissance de temps*). La menace d'un mal n'est pas le mal même (…) : tout contact est donc *interprétation* ; il nous impose de comprendre en affection éventuelle ce qu'il nous donne en qualité de pression, de dureté, de mollesse, de rugueux, de poli. Le mal en question n'est pas présent ; il ne peut donc être que *représenté en images* (*imagination*) mais cette représentation n'est pas une divination ; elle ne peut donc être qu'un souvenir provoqué par l'analogie des cas (*mémoire et association*). La réaction doit être composée avec tous les éléments documentaires (*calcul et dessein*). Je laisse à penser les *sentiments* multiples qui peuvent se greffer sur ces connaissances. Il est donc bien vrai que toute la vie de l'esprit s'introduit dans l'action avec la parade tactile. »[19] On voit à quel point l'animal, auquel pourrait s'appliquer cette analyse, est l'être qui « mesure » et évalue, ouvrant ainsi la possibilité d'une conscience incarnée déployant ses potentialités au plus près de la matière touchée tout en ne s'y laissant jamais réduire. Il tend, au contraire, à développer une vie de l'esprit qui intègre des éléments absents à la présence matérielle du tactile pour en faire

15 Ibid.

16 Ibid., p. 298.

17 Ibid.

18 Pradines (M.), L'aventure de l'esprit dans les espèces, Flammarion, « Bibliothèque de philosophie scientifique », pp. 166/167.

19 Ibid., p. 94.

émerger de nouveaux niveaux de signification. Cette spiritualisation est un affinement dans l'ordre des relations ; il conditionne l'accès à un monde de nuances et de subtilités. Minkowski fait ainsi une remarque instructive qu'il réserve aux humains mais qui pourrait s'appliquer à presque tous les animaux sensibles et « tactiles » : « Avoir du tact ou du " doigté ", écrit-il, est une particularité essentiellement humaine. Avoir le sens des nuances du toucher, toucher sans heurter, sans blesser, tels sont ses caractères. Avoir du tact est presque une vertu. Et il est surprenant au premier abord de voir le sens du toucher, cet " infirme ", donner naissance à cette " métaphore " destinée à désigner une délicatesse toute exceptionnelle de l'être humain dans ses rapports avec ses semblables. »[20]

A un autre niveau enfin, la perception tend de plus en plus à dégager la représentation pour elle-même, dans sa valeur plaisante ou esthétique apparaissant alors comme « distraction » à l'égard de l'utile. Déjà la sensation, « impression pourvue de sens », n'agit pas sur le mode de la causalité immédiate : elle est à l'origine de « plaisirs nouveaux » détachés du besoin. Un univers plus désintéressé émerge dans l'environnement de l'animal. Certains vertébrés peuvent-ils alors atteindre le « seuil de l'art » ? Pradines montre que l'être vivant peut s'attacher à des sensations inutiles intégrées à la perception qui alors « (…) n'est qu'un jeu de plus en plus étendu avec des images. La perception, en progressant, s'apprêtait donc à servir son propre contraire fonctionnel. (…) Les couleurs et les sons qu'elle nous a donné pour atteindre les objets vont nous devenir plus importants que les objets mêmes. L'artiste est cet être que la couleur, la lumière, les formes et les sons enivrent à ce point qu'il en devient aveugle et sourd aux *choses* qui les produisent, et pour qui le monde n'existe plus que dans ses symboles. »[21] Ainsi, le jeu avec les sons libère des simples bruits de la nature comme ce peut être le cas dans les chants d'oiseaux. De même, une « représentation sensible » « ludique et spectaculaire » créée pour l'être vivant « jusqu'au sein de la perception, une tentation de distraction à l'égard de son propre objet, c'est-à-dire des choses représentées et préavisées »[22]. On retrouve de tels exemples dans les jeux des jeunes Faucons pèlerins en vol, l'objet que l'oiseau cherche à saisir, une feuille ou un insecte par exemple, ne correspond ni en taille ni en forme aux espèces qu'il chassera adulte. Il joue alors, s'entraîne et « s'illusionne » tout en appréciant cette activité pour elle-même sa fonction et son utilité étant décalées[23].

Dans une combinaison de qualités sensibles valant en elle-même naît une associativité nouvelle fondée sur des principes attrayants et esthétiques − au sens de la perception plus ou moins unifiée d'un ensemble de combinaisons harmonieuses visuelles, sonores et mobiles − comme cela s'observe chez certains Paradisiers pour lesquels les parades nuptiales et l'architecture des nids ne semblent plus guidées par la seule utilité. Chez ces oiseaux, l'attrait pour des formes spectaculaires et distrayantes est souvent couplé à une sensibilité à des formes colorées évoquant des figurations à forte connotation symbolique comme des cercles, des angles, ou encore des lignes brisées ou continues. Au cours des parades nuptiales, l'oiseau vit d'ailleurs un état d'excitation et de stupéfaction devant le spectacle qui s'offre à lui, semblant saisi et interloqué par l'aspect extraordinaire que prend l'allure de son partenaire − exhibant par exemple des plumes démesurées, des tâches plus ou moins géométriques de couleurs vives, bigarrées, présentant même une désarticulation

20 Minkowski (E.), 1967 (1re édition 1936), Vers une cosmologie. Fragments philosophiques, Aubier-Montaigne. p.184.

21 Pradines (M.), L'aventure de l'esprit dans les espèces, Flammarion, p. 172.

22 Ibid., p. 119.

23 Cf. Monneret (J.R.), 2006, Le faucon pèlerin, Delachaux et Niestlé, « Les sentiers du naturaliste ».

de la structure du corps devenu méconnaissable. Ces parades démonstratives relèvent d'une sorte de théatralisation et de dramatisation allant jusqu'à inclure une forme de violence symbolique : comme l'explique un spécialiste dans le *Handbook of the Birds of the World* : « Courtship (…) typically includes surprise elements of vigorous advance at females, including dramatic changes of size and appearence, attitude or sound, and it may even involve the pecking of females or the beating of them with the open wings.»[24]

### 2.3 Dualité organique et aperception de soi. L'expérience de l'effort

La vitalité active exige un effort qui engendre une connaissance première déterminée par l'expérience du corps qui nous affecte. Or, « du point de vue de la physiologie, la volonté se trouve identifiée à la sensibilité, écrit Maine de Biran ; il est donc impossible de lui trouver une origine autre que celle de la vie : l'être organisé ne se meut que parce qu'il sent et il ne sent que parce qu'il se meut, conclut-il (…).»[25] La volonté naît d'abord insensiblement de l'effort, lui-même indissociable d'une mobilité liée au fonctionnement des parties organiques, comme lors de la digestion par exemple, et de "l'auto"-mouvement de l'organisme ainsi que de l'affection qui l'accompagne. Ce processus d'effort implique un dynamisme interne tourné vers ce qui constitue une opposition indifférenciée mais objectivée sous forme d'une résistance organique rencontrée lors de l'accomplissement de l'action. Une logique de la vie liée au sentiment de l'effort mène ainsi à l'engendrement d'une notion du temps interne vécu et permet l'émergence d'un *sujet organique* disposant d'une notion de ce qu'il est. Engendrant un acte indépendant lui permettant de se reconnaître, l'animal accède à une connaissance de ce qu'il est, à une manière de fonctionner qui l'affecte. Multiplicité cohérente d'organes finalisés, le corps animal est constitué de parties interdépendantes et « communicantes ». Le sens de l'effort continu, global et unifié, intègre cette multiplicité que l'animal ne se représente pas. Au principe de l'animation corporelle se greffe alors une aperception interne qui renvoie l'agent à la genèse de sa propre activité en tant qu'il en est l'auteur par ses activités et comportements. Cette aperception est le « résultat sensible pur de l'action de l'âme sur le corps » ; ce résultat permet de concevoir une « affection simple sans personnalité. »[26] Le sens de l'effort permet à l'animal de saisir intrinsèquement le fait même de sa propre existence dans la mesure où ce sens « (…) n'a besoin de se composer avec aucun autre pour qu'existe le sujet capable d'apercevoir sa propre existence »[27]. Celle-ci n'est pas représentée car le sens de l'effort est « (…) immédiat, c'est-à-dire exempt de médiations visuelles et tactiles » et que « l'espace intérieur dans lequel le corps propre est aperçu est strictement irreprésentable (…) »[28] On comprend alors qu'« au déploiement unique de cet effort commun, à l'uniformité ou à la continuité de résistance organique, doit correspondre le sentiment d'une sorte d'étendue intérieure d'abord vague et illimitée. »[29] En se localisant,

24 Hoyo (J.), Elliot (A.), Christie (D.), 2009, Handbook of the Birds of the World, Lynx edicions, Volume 14, p. 429. Le sens de l'harmonie dépend de l'exercice des facultés animales ; il semble antérieur à la beauté proprement dite qui suppose la contemplation d'une œuvre achevée. Sur ces questions d'esthétique animale voir : Cosson (F.), 2016, Animalité et humanité. La frontière croisée, Ovadia, « Chemins de pensée », troisième partie : « La beauté des mondes. Esthétique croisée. »

25 Azouvi (F.), 1995, Maine de Biran. La science de l'homme, Vrin, « Histoire de la philosophie », p. 161.

26 Ibid., p. 229

27 Ibid.

28 Ibid.

29 Maine de Biran, 1805, Mémoire sur la décomposition de la pensée (version remaniée), dans Œuvres, 2000, Vrin, « Bibliothèque des textes philosophiques », tome 3, p. 432.

le processus se précise ensuite dans la mesure où « chaque effort particulier, chaque acte exprès de la volonté va localiser un terme d'action ou marquer des points de séparation dans le contenu résistant ; à mesure que le mode fondamental se reproduit sous des formes plus variées, l'existence personnelle s'affermit et se développe. Le sujet moteur s'individualise plus complètement dans le rapport senti des termes mobiles. En se mettant hors de chacun, le moi apprend à distinguer et leurs limites communes et les siennes propres. »[30] On échappe ainsi à l'affection simple dans laquelle « (…) les impressions quelconques demeurent générales, absolues et diffuses (…) »[31] et, surtout, l'être vivant commence à s'apercevoir lui-même en tant qu'il s'oppose par l'effort à des parties localisées et, par cette amorce d'objectivation, entame un processus de dédoublement qui préfigure, bien que de manière encore organique, l'émergence d'un être conscient de son activité et d'une présence au corps propre médiatisée par l'effort. Si nous appliquons ces analyses à l'animal, nous dirons qu'il est, pour lui-même, une réalité éprouvée et qu'en raison de la dualité et de l'objectivation naissantes dont il est question, il accède au sentiment d'une existence propre difficile à appréhender et à décrire dans la mesure où elle ne passe pas par un « je » référentiel et par des actes de langage.

Quoi qu'il en soit de ces difficultés qui tiennent à l'impossibilité de transcrire cette expérience silencieuse, nous retiendrons que le sentiment du moi est lié à la réalité éprouvée d'une étendue corporelle plus ou moins différenciée et en cours d'objectivation qui permet à l'animal d'accéder à la notion d'une individualité unifiée inscrite dans la continuité d'une durée vécue. La genèse corporelle du moi s'opère finalement grâce au contact spatialisé de la volonté qui fait effort, et fait émerger la notion d'un moi inscrit tout à la fois dans l'immanence de l'action et dans « l'espace intérieur » qui lui est opposable et constitue la prémisse d'une dualité naissante.

## 3. Emotion, communication et conscience

### 3.1 Emotion animale et connaissance du soi

Mais des états de conscience peuvent-ils apparaître sur la base d'un rapport plus caractérisé, plus spécifique, avec le monde extérieur en intégrant des sentiments émotionnels qui reflètent l'environnement vécu ? Des émotions individuelles induites par l'expérience passée exprimeraient alors un profil spécifique pour l'organisme réaffecté.

Sur le plan neurophysiologique, l'existence d'émotions d'arrière-plan reflètent l'état de l'organisme : « Les processus de régulation biologiques eux-mêmes peuvent provoquer des émotions d'arrière-plan, mais des conflits mentaux incessants aussi, qu'ils soient explicites ou implicites, lorsqu'ils suscitent une satisfaction soutenue ou une inhibition de pulsion ou de motivations (…). Certaines conditions de l'état interne engendrées par des processus physiologiques continus ou par les interactions de l'organisme avec l'environnement, ou bien encore les deux, provoquent des réponses qui constituent des émotions d'arrière-plan. Ces émotions nous permettent d'avoir, entre autres, les sentiments d'arrière-plan de tension ou de relaxation, de fatigue ou d'énergie, de bien-être ou de malaise, d'anticipation ou de crainte. Dans les émotions d'arrière-plan, les réponses constitutives sont plus proches du for intérieur vital, et leur rôle est plus interne qu'externe. »[32] L'émotion renvoie ainsi à *une qualité*

30 Ibid., p. 433.

31 Ibid.

32 Damasio (A.), 2002, Le sentiment même de soi. Corps, émotion, conscience, Poches Odile Jacob, p. 73.

*d'expérience éprouvée de manière globale par l'animal qui en est affecté* et non plus seulement à un mécanisme d'induction physiologique ou cérébrale transcrit sous forme d'expressions ou de comportements.

Ces émotions sont en outre liées à la manière dont l'animal est affecté en raison de sa constitution et de son histoire. Elles sont, en particulier, façonnées par les attentes et la structure de l'environnement plus ou moins changeant ; elles sont alors « (…) inséparables de l'idée de récompense ou de punition, de plaisir ou de peine, d'approche ou de retrait, d'avantage ou de désavantage personnel. Inévitablement, les émotions sont inséparables de l'idée de bien ou de mal. »[33] Celles-ci permettent de ressentir la mise à disposition d'une énergie physiologique, la force affective induite par les situations ainsi que leur qualité distinctive : vecteurs d'états psychologiques chez les espèces les plus évoluées, elles constituent un mode de *connaissance auto-affective,* autrement dit un mode de connaissance par compréhension de ses propres états. Enfin, les émotions induites sont elles-mêmes inductrices en un sens plus général car « (…) elles fournissent automatiquement à l'organisme des comportements orientés vers la survie. Dans des organismes équipés pour sentir des émotions, c'est-à-dire pour avoir des sentiments, les émotions ont également un impact sur l'esprit, telles qu'elles se présentent dans l'ici et le maintenant. Mais dans des organismes équipés de conscience, c'est-à-dire capable de savoir qu'ils ont des sentiments, c'est un autre niveau de régulation qui est atteint. La conscience permet aux sentiments d'être connus et promeut ainsi l'impact de l'émotion de façon interne ; elle permet à l'émotion d'imprégner le processus de pensée par l'entremise du sentiment. »[34] Que l'animal ne puisse rendre compte de cette qualité d'expérience n'a pas d'importance ; l'essentiel est qu'il s'agisse d'un état unifié qui constitue *en lui-même* le signe de l'état d'un être qui *s'anime* sur cette base émotionnelle et constitue alors une connaissance qu'il a de son propre mode de réaction et une motivation pour une action ainsi éclairée par l'émotion qui l'instruit. C'est dans cette autonomisation de la conduite qu'il faut chercher l'émergence d'états internes qui, dans leur spécification individuelle et existentielle, expriment la singularité de l'animal. Ajoutons que cette commotion physiologique et affective − et peut être « psychologique » − provoque des changements dans la perception de l'environnement affecté alors d'une sorte de coefficient d'ambiance, d'une coloration subjective déterminée par le type d'émotion en jeu avec des variations d'intensité indissociables de l'individualité et de son expérience étho-écologique.

Enfin, de par le rapport problématique au réel qu'elle induit et du fait de son intensité, toute émotion est une mise en question de l'organisme ou de son environnement. Renvoyant à une réaction globale qui l'affecte au point de le troubler dans ses conduites, elle constitue un mode de connaissance si la commotion initiale devient une impulsion propice à la réalisation d'une d'action biologiquement normée ou si elle constitue un mode d'apprentissage utile. Le phénomène émotionnel met en jeu une puissance qui excède, dans ses effets, la norme des conduites habituelles et constitue un saisissement, un temps d'arrêt et de surprise[35]. C'est pourquoi l'animal peut apprendre, comme l'enfant, à se connaître dans le fait émotionnel : « les émotions en effet (…) s'imposent à l'attention de l'enfant et lui fournissent un thème d'intérêt d'autant plus accessible qu'il est lui-même l'auteur des effets qu'il observe ; (…) ces démonstrations sont " plus proches de la représentation que de l'action " et, par-là, elles

33 Ibid., p. 76.

34 Ibid., pp. 77/78.

35 On retiendra la distinction entre étonnement, admiration et surprise : « A. Smith distingue l'étonnement, réaction à l'insolite et à l'étrange, de la surprise, réaction à l'inattendu dans l'ordre du connu, et de l'admiration, réaction au beau et au grand, même dans l'ordre du familier. » (Canguilhem (G.), 2002, Etudes d'histoire et de philosophie des science, Vrin, « Problèmes et controverses », p. 91).

deviennent le support de la conscience à ses débuts. »[36] L'émotion peut se développer en outre de manière individualisée lorsqu'elle s'inscrit dans le cadre d'une expressivité reconnue dans une extériorisation « publique » ou destinée aux autres. Ajoutons que, dans sa manifestation organique, elle apparaît lorsque la représentation d'une chose ou la situation vécue est indissociable d'une valeur vitale qui ne fait qu'un avec sa signification. Elle constitue alors une préfiguration de notions normatives de « bon » et de « mauvais » ou d'autres de même nature et laisse entrevoir la possibilité de l'émergence d'une conscience morale naturelle qui serait l'expression des nécessités organiques.

### 3.2 Émotion et altérité : la conscience empathique de l'autre

L'émotion s'accomplit ainsi comme possibilité cognitive prolongée par une action qui l'accomplit en lui donnant une signification biologique observable et communicative. De ce point de vue, le sens profond du phénomène émotionnel est interrelationnel. « L'émotion, explique Henri Wallon, a besoin de susciter des réactions similaires ou réciproques chez autrui et, inversement, elle a sur autrui une grande force de contagion » ce qui explique que, dans un premier temps au moins, elle tend à effacer « la distinction de soi et d'autrui. »[37] En outre, d'après le phénoménologue Max Scheler, les sentiments animaux peuvent être distingués de deux manières : « Alors que les sentiments sensoriels sont étendus et localisés, le sentiment vital participe au caractère globalement extensif du corps propre, mais sans contenir en lui une extension et un lieu déterminé. »[38] De plus, « Le sentiment vital et ses modes propres constituent déjà une couche émotionnelle originale et irréductible à celle des sentiments sensoriels.»[39]. Les animaux peuvent-ils éprouver des « sentiments vitaux » les mettant à distance d'eux-mêmes ? il semble que la réponse soit positive si nous suivons les analyses de Scheler : « Dans le sentiment vital, nous sentons notre *vie* elle-même ; en d'autres termes, *dans* ce sentir même, quelque chose nous est " donné ", ce qui est son " accroissement " ou sa " diminution ", sa maladie ou sa santé, ses " dangers " ou son " avenir ". Et la chose vaut aussi bien pour le sentiment vital orienté vers notre propre vie que pour celui dont la fonction s'applique (…) à d'autres êtres vivants (…) dans la sympathie vitale »[40]. Ce sentiment est donc « (…) accessible à notre sympathie et à notre compréhension, quelles que soient ses manifestations et dans l'étendue du monde vivant, bien que les *qualités* particulières qu'il affecte chez les représentants les plus inférieurs du monde animal nous échappent (…).»[41] Dès lors, poursuit-il, « Nous comprenons en " sympathisant " l'angoisse mortelle d'un oiseau, sa " fraîcheur " ou sa " lassitude ", etc., mais ses sentiments *sensuels*, qui dépendent de l'organisation spéciale de ses organes des sens, nous sont inaccessibles.»[42]

Dans ce contexte d'un monde partagé sur le mode du vécu interspécifique, l'animal, communiquant avec des êtres zoologiquement ou « psychologiquement » proches, tisse la structure émotionnellement vécue d'un monde commun fondée sur une co-compréhension

36 Martinet (M.), 1972, Théorie des émotions. Introduction à l'œuvre d'Henri Wallon, Aubier Montaigne, « Analyses et raisons », p. 114.

37 Ibid., p. 105.

38 Scheler (M.), 1955. Le formalisme en éthique et l'éthique matériale des valeurs. Essai nouveau pour fonder un personnalisme éthique, Paris, Gallimard, « Bibliothèque de Philosophie », p. 345.

39 Ibid.

40 Ibid., pp. 347-348.

41 Scheler M., 1971. Nature et forme de la sympathie. Contribution à l'étude des lois de la vie affective, Paris, Payot, coll. « Petite bibliothèque Payot », p. 73.

42 Ibid.

d'origine empathique. Celle-ci laisse entrevoir la possibilité d'une reconnaissance de l'autre ainsi que l'établissement d'un rapport cognitif fondé sur la similarité des émotions communiquées et peut-être de comportements plus élaborés. La possibilité d'induire l'existence de sentiments chez d'autres individus est impliquée dans cette relation. Les comportements d'apaisement chez les primates indiquent par exemple que l'animal, empathique, éprouve quelque chose de ce qu'éprouve son congénère. En cas de malaise ou de souffrance notamment, il accède à la signification de certaines de ses émotions dans une relation qui s'apparente à une compassion qui reste globale et non objectivée quant à sa nature et à sa localisation. Les prémisses d'une conscience de relation distinguant l'utile du nuisible en particulier a dû prendre naissance dans la révélation émotionnelle et affective d'états considérés comme défavorables et « mauvais » (comme la douleur, l'insatisfaction, la tension du désir inaccompli, la frustration, l'inquiétude et la crainte) ou favorables et « bons » (le plaisir et la jouissance, la satisfaction, le désir accompli, la satiété, et la quiétude) éprouvés par d'autres que soi.

Chez certains primates ces capacités prennent parfois une tournure anticipatrice et stratégique montrant que les relations s'inscrivent dans une temporalité qui, tenant compte du passé, intègre l'avenir et génère un décalage par rapport au présent. L'hypothèse d'une *conscience de situation* peut alors être proposée dans la mesure où les relations de groupe peuvent être fondées sur l'observation d'attitudes et la compréhension d'états émotionnels ou de comportements socialement connotés. Cette compréhension mutuelle, à la base de l'établissement de relations interindividuelles complexes et évolutives, s'avère indispensable à une organisation cohérente et durable fondée sur la nécessité d'une stabilisation ou d'une pacification des rapports entre membres de la communauté. Les stratégies de coopération et de réconciliation prennent une grande importance et la connaissance mutuelle des individus, évolutive et spécifiée en fonction de contextes en devenir, valorisent l'observation et la connaissance des autres dans le but d'apprécier leurs dispositions physiologiques mais aussi leurs motivations psychologiques, leurs intentions et d'éventuelles possibilités d'alliances. Certains comportements attestent d'une grande subtilité en terme de mise en place de stratégies conciliatrices qui intègrent la nécessité de mettre en œuvre des attitudes échappant à l'immédiateté des réactions. Ces attitudes, individuelles, construites et réfléchies − sorte de sagesse pratique permettant l'optimisation de l'action et de la coordination sociales s'expriment dans l'attente du moment opportun ou dans l'évaluation « mesurée » de tentatives tenant compte des circonstances et d'un passé mémorisé qui pourrait interagir avec la situation présente. En outre, l'existence d'un sens de la réciprocité et de relations fondées sur la mémoire des services mutuels atteste d'une capacité à intégrer les états vécus par les congénères. Chez des Chimpanzés, des relations proches de la gratitude s'instaurent grâce à la mémorisation de faits passés et déterminent un échange de bons procédés. Ajoutons que l'émergence de relations de plus en plus subtiles pourrait donner naissance à des notions moralement connotées impliquant une conscience empathique naturelle assez proche de celle du « bien » et du « mal ». Ces relations relèveraient d'un mode de compréhension intégrant une diversité d'histoires individuelles transmises sous la forme d'une culture de groupe dans laquelle chaque individu est identifié par les autres dans le cadre de la position relative qu'il occupe et du rôle qu'il joue. Cette compréhension pourrait donner naissance à une forme de conscience collective distinguant ces catégories héritées, et intégrées à la dynamique de la vie sociale.

L'émotion est enfin, au point de vue cognitif, indissociable d'une extériorisation corporelle. Trans-individuelle et méta-individuelle elle affecte les autres individus ou s'adresse à eux. Dès lors, l'animal, sur la base d'une compréhension réciproque d'origine empathique, peut communiquer avec des êtres ayant des capacités de ressentir et de se rapporter au monde

comparables aux siennes. Chez des animaux interactifs et sociaux, l'expression d'un « sentiment vital » représente une part importante de la communication. La communication n'est alors plus « froide », fondées sur des échanges chimio-tactile ou de simples cris d'alerte par exemple, mais procède par expression communicative d'états internes ; participative, expansive et « contagieuse », elle laisse entrevoir la possibilité d'une prise de *conscience collective* d'un événement suscitant une réaction émotionnelle communicable affectant tout un chacun et générant une modalité commune d'existence. La conscience s'intègre en effet à *la structure émotionnellement vécue d'un monde commun qui en constitue le fondement. Le caractère empathique de cette compréhension réciproque fixe les propriétés des significations partagées*. Tout individu, affecté et affectant au regard de la transmission de l'information empathique, se distingue des autres en accédant à une perception de lui-même. Cette forme de conscience empathique fonctionne comme un prisme, le vivant s'apparaissant à lui-même à travers l'autre dans une relation médiatisée dans laquelle les individus sont co-donnés dans une expérience où n'apparaît pas encore une dualité constituée qui permettrait de se distinguer comme être indépendant et « isolé » de celui ou de ceux qui nous font éprouver un état déterminé.

Nous poserons pour terminer la question de savoir s'il existe des échanges permettant à certains animaux d'établir des relations intersubjectives avec l'homme et de se référer à des situations partagées et à un monde commun −monde anthropo-zoologique « mixte » comprenant des éléments appartenant aux deux espèces. L'existence de ces relations s'avère décisive pour établir la possibilité de l'émergence d'une conscience animale se développant au contact de l'homme. Dans le cas de mammifères sociaux hautement interactifs comme le chien, ces relations − fondées sur la familiarité d'une vie partagée où les signes manifestés enrichissent l'environnement de l'animal ainsi « humanisé » − lui permettent d'accéder à des niveaux de significations qui favorisent l'émergence d'une empathie interspécifique et d'une psychologie de situation à connotations « mixte », à la fois humaines et canines. Elles impliquent la compréhension de conduites interindividuelles instituant un être en commun. Cet être en commun, composante essentielle de la familiarité domestique, permet l'émergence d'une compréhension trans-spécifique de la part de l'animal ainsi que d'une souplesse psychologique permettant un optimum adaptatif [43]. Ces facultés nouvelles s'expliquent sans doute par le fait que « Le chien et le chat sont des hybrides sociaux : ils se considèrent à la fois comme des individus de leur espèce et de notre espèce.»[44] La longue coexistence de l'éthologue Konrad Lorenz avec des chiens l'amène à d'intéressantes conclusions à la fois humaines et canines. Ainsi note-t-il, certaines « mimiques démonstratives » « (…) sont pour une grande part indépendantes des caractères innés, et c'est l'animal en tant qu'*individu* qui les a apprises ou librement inventées. Aucune règle préétablie de l'instinct ne force un chien à exprimer son affection en posant la tête sur le genou de son maître. »[45] Mais, « Quand ils " parlent " à leurs semblables, les chiens les plus riches en expressions individuellement acquises reviennent aux mimiques de l'état sauvage.(…). Les chiens les mieux domestiqués sont donc les plus riches et les plus adaptables, conclut Lorenz. »[46] De même, sans cesse accompagné et aidé dans son

43 Pour une discussion détaillée de ces questions concernant la domesticité voir Cosson (F.), Animalité et humanité. La frontière croisée, 1997, Ovadia, « Chemins de pensée », en particulier Cinquième partie, § 9 : « Au-delà de la domesticité ? Institutions de communautés anthropozoologiques. »

44 Jouventin (P.), 2014, Trois prédateurs dans un salon. Une histoire du chat, du chien et de l'homme, Belin, p. 129. P. Geluck souligne, sur un mode humoristique, cette ambivalence relationnelle : « J'ai connu un type qui était très humain avec son chien. Et le chien le lui rendait bien. Il était très canin avec son maître. » (Geluck (P.), 2005, Le chat a encore frappé, Casterman, p. 16.)

45 Lorenz (K.), 1970, Tous les chiens. Tous les chats, Flammarion, p. 180.

46 Ibid., pp. 180/181.

existence sensible et affective, l'animal accède probablement à des notions « compréhensives » et intuitives liées à la présence et à la sollicitude du maître qui s'expriment à travers la bienveillance par exemple. Ces notions pourraient fort bien constituer les prémisses de sentiments moraux encore obscurcis par l'intensité émotionnelle et les comportements d'appétence innés qui sont en jeu.

### 3.3 Action finalisée, conscience incarnée et liberté

Le philosophe Ruyer, s'intéressant de près à la biologie, nous propose, dès les années 50, une notion de conscience non anthropocentrique étendue à tous les vivants animaux : « Si une étendue " absolue " ou vraie, implique, dans sa texture même, conscience, conscience signifie tout simplement " existence ici ", et non " existence connue par un Je " comme " troisième œil ", comme œil philosophique écrit-il. (…) " Existence ici maintenant " ou "présence de conscience" n'implique pas que la conscience " se montre ", et dise " présent " à quelqu'un. À qui se montrerait-elle, sinon à un œil mythique, redoublement inexistant et inutile de l'œil réel ? L'œil réel, dont le bon fonctionnement et condition d'existence de la vision, n'a pas à être remis en scène une fois la vision présente. »[47] Le corps en effet, coextensif à l'*Umwelt*, ne permet pas d'accéder à une conscience-substance indépendante et permanente, à l'énoncé d'un « je suis » qui culminerait dans le concept d'un « moi » fictif détaché dans une abstraction et une autonomie cognitive métabiologique rendant le sujet transparent à lui-même. En tant qu'organisme en effet, *il vit le « suis » de manière active sans « je » (l'être qui est en acte)* et ne peut distinguer clairement *ce* qu'il exécute de lui-même *qui* exécute, ce qui revient à reconnaitre qu'il n'y a pour lui ni objet ni sujet indépendants de l'action en cours et que le « qui » est dans le « ce ». Il éprouve sa vitalité telle qu'elle se manifeste à travers l'être sensible aux prises avec le monde perçu. Soulignons qu'un « je suis » pensé ou énoncé comme cogito aurait un coût biologique exorbitant. S'affirmer sur la base d'une conscience de soi reviendrait en effet pour l'animal à agir en gardant contact avec le réel dans la présence engagée de sa structure corporelle tout en s'en abstrayant par un acte de conscience de soi.

Faisant prévaloir ces thèmes qui s'associent aux éléments de perception, l'animal dispose alors d'une « conscience organique primaire. »[48] Tout comme l'homme, il agit en fonction d'un thème unifié par l'action finalisée inscrite dans une durée. Le phénomène conscient plonge très loin ses racines car « la conscience sensorielle, (…) n'est jamais que l'auxiliaire d'une activité instinctive ou visant une réalisation de valeur pressentie. L'animal cherche une proie. L'homme cherche à préciser le sens de la forme vue, ou à " reconnaître " le personnage entrevu. Une sensation vient toujours dans un contexte d'effort en cours. »[49] C'est pourquoi ce phénomène conscient, avec ses différences de degrés, se concrétise comme présence finalisée d'un thème qui oriente la conduite en lui donnant une signification. Tout état conscient *présent à l'animal*, est alors *traversé* par une intentionnalité qui fait de l'activité finalisée une réalité immanente à l'organisme pour lequel tout est donné dans un présent vécu et signifiant qui prolonge un passé − immémorial/inné ou mémorisé/acquis − finalisé par l'avenir.

47 Ruyer R., 1966. Paradoxes de la conscience et limites de l'automatisme, Paris, Albin Michel, coll. « Les savants et le monde », pp. 23-24. Voir également la définition du « champs de conscience » dans Ruyer (R.), La conscience et le corps, 1937, Félix Alcan, « Nouvelle bibliothèque de philosophie », p. 52.

48 « Nous n'avons pas – à part nos préjugés humains – la moindre raison de refuser à un protozoaire ou à la colonie amibienne, quand elle manifeste un comportement unitaire, la conscience de ses mouvements, comme schémas de comportement, de même sorte exactement que la conscience volontaire du gymnaste qui organise la forme et le rythme d'une performance musculaire. » (Ruyer R., 1964. L'animal, l'homme, la fonction symbolique, Paris, Gallimard, coll. « L'avenir de la science.», pp. 61/62.

49 Ibid., p. 84-85.

En termes darwiniens, ce thème constitue le soubassement organique capitalisé par l'espèce sous forme de « traits d'histoire de vie » au cours de la phylogénèse. Ces derniers se manifestent dans les phénomènes de la vie dès l'apparition de conduites déterminées par des tendances instinctives. Celles-ci donnent à l'animal *la notion de ce qu'il éprouve c'est-à-dire la connaissance sensible et intuitive de ce qu'il cherche alors même qu'il le cherche et en est affecté comme manque consubstantiel à son être ou comme simple déficience organique à combler*. Cette connaissance, bien que confuse − car l'animal ne peut l'objectiver comme telle −, est accaparante. Elle constitue une modalité ontologique fondamentale du vivant animé. Affecté par le fond organique qui le constitue, à la fois affecté et affectant, l'être vit cette présence somatique sur la base de *ce* qu'il éprouve de lui-même en tant qu'*il* éprouve dans une sorte de circuit de récurrence fermé − expérience primordiale qui renvoie à ce que nous appellerons la *sphère animale d'auto-affection première* [50]. Poussée par une pulsion migratoire ou de reproduction par exemple, l'animal ne se représente pas la migration ou la reproduction dans ce qu'elles impliquent. Le déterminisme ne comporte vraisemblablement pas d'images précises et rien n'est objectivé comme tel. Les stimuli environnementaux constituent en effet le déterminisme des impulsions lorsque l'organisme se trouve dans des conditions physiologiques spécifiques. Hanté par ces impulsions au plus profond de sa constitution, il se connaît comme affecté par cette expérience pulsionnelle confuse mais thématisée qui constitue un *comme je suis* qui correspond à un *ce que je suis* c'est-à-dire à une manière d'être à la fois donnée et actualisée par l'animal. Il est alors un simple sujet « (…) caché à lui-même (…) – il n'a que son corps-vécu et il se fond dans la centralité spatialisante-temporalisante de la vie subjective sans vivre cette vie, il est à partir d'elle pur " me " (moi-objet), non pas " je " (moi-sujet) (…).»[51]

Certes, un ordre organique et instinctif conditionne la conscience primaire pour l'animal habité par les finalités qu'il poursuit. Le dehors et le dedans constituent alors deux facettes d'une même réalité organique qui s'appellent l'une l'autre dans un dialogue dont l'animal est le médiateur. Un prédateur visuel se confondra ainsi avec l'étendue visible en dégageant une perspective thématique inséparable du type de proies recherché. Ce dernier, intégré comme reconnaissance de formes, de couleurs ou de mouvements associés, constitue une image directrice. Il est alors présent à la totalité de l'étendue concrète d'où émergent des significations qui impliquent des conduites orientées par la fin qui l'habite, le stimule et le motive. Dans un contexte où la réalité mise en forme devient structure signifiante, cette totalité acquiert une signification *topologique*[52]. L'espèce agissante éprouve des états assimilables à des états subjectifs dans la mesure où ils produisent des effets sur l'animal qui les éprouve et agit selon leurs significations.

---

50 Au sens chronologique et ontologique.

51 Plessner (H.), 2017, Les degrés de l'organique et l'homme. Introduction à l'anthropologie philosophique, Gallimard, « Bibliothèque de philosophie », p. 421.

52 La reconnaissance n'est en effet pas seulement topographique et empirique. Elle implique une mise en forme signifiante avec des lignes de forces émergeantes et comporte un caractère de généralité rappelant la forme (*Geschtalt*) dans le sens que donne René Thom à ce terme : « J'appelle " forme saillante " écrit-il toute forme qui frappe l'appareil sensoriel d'un sujet par son caractère abrupt ou imprévu (…). Une forme peut être saillante par une irrégularité de rythme, une brisure de symétrie, aussi bien que par une discontinuité sensorielle (…). Une forme saillante peut saturer momentanément l'appareil sensoriel du sujet, elle s'inscrit dans la mémoire à court terme (…). Par opposition, certaines formes ont pour le sujet une importance biologique immédiate ; telles sont chez les animaux, les formes des proies, des prédateurs, des partenaires sexuels. De telles formes seront dites « prégnantes », *prégnance* désignant la qualité associée. Elles induisent des modifications importantes dans le comportement moteur ou affectif du sujet, avec des changements hormonaux à longue durée dans sa physiologie. Ces formes peuvent être attractives ou répulsives. » (Thom (R.),1990, Apologie du logos, Hachette, « Histoire et philosophie des sciences », pp. 55/56.)

Ainsi donc cette conscience de premier niveau – connaissance de soi manifestée comme totalité organique dans l'acte finalisé – se manifeste comme *préscience actualisée et manifestée* : l'animal « sait » ce qu'il cherche en l'éprouvant, en étant affecté par la déficience organique et la pulsion qui le poussent à s'extérioriser en vue d'un accomplissement dont lui seul connaît, au plus profond de son organisation somatique, la raison d'être. Sans cette conscience première, réinitiée à chaque instant, il devrait être remonté comme un automate ayant à l'extérieur de lui-même la cause matérielle et le principe intellectuel de son animation. Cette option serait contraire à la logique même de la vie qui ne s'accomplit que par le dialogue de moyens donnés et de fins poursuivies en vue de la constitution d'actions adéquates. Dès lors, un profil d'existence transcrit quelque chose de la manière dont les animaux appréhendent le monde tout en s'éprouvant eux-mêmes. Cette présence est, selon la terminologie de Gilbert Simondon, « transductive ». En effet, « (…) l'être individuel est transductif, non substantiel, et la tendance de l'être à persévérer dans son être cherche l'équivalence d'une substantialisation, même si l'individu n'est fait que de modes. »[53] L'individu, toujours en cours d'individualisation, est processus en devenir et tension différenciatrice en cours de singularisation. Ce processus « transductif » signe une présence originale au monde dans la tension vitale finalisée qui l'habite et qu'il actualise sans cesse en assumant une relation constructive entre le passé incarné par l'espèce et l'avenir qui donne sa signification concrète au présent des situations vécues.»[54]

Nous proposerons ici une conclusion nuancée qui engage malgré tout sur la voie du développement d'une vie consciente chez les animaux : « les gestes du comportement, les intentions qu'il trace dans l'espace autour de l'animal ne visent pas le monde vrai ou l'être pur, mais l'être-pour-l'animal, c'est-à-dire un certain milieu caractéristique de l'espèce, ils ne laissent pas transparaître une conscience, c'est-à-dire un être dont toute l'essence est de connaître, mais une certaine manière de traiter le monde, d' " être au monde " ou d' " exister ". » écrit Merleau-Ponty. Une conscience est, selon le mot de Hegel, un " trou dans l'être " et nous n'avons encore ici qu'un creux. »[55] « Creux » dont nous pourrions ajouter qu'il ne peut que s'approfondir dans le cours des phénomènes évolutifs menant à une individualisation et à une singularisation toujours croissantes se manifestant sous la forme d'états conscients de plus en plus prégnants.

### 3.4 Finalité immanente à l'acte de vision : perception et émergence « consciente » de niveaux de signification

L'analyse des systèmes perceptifs est essentielle pour toute investigation portant sur la conscience animale. La manière dont l'environnement apparaît aux animaux constitue une véritable « signature » qui permet d'identifier des profils d'existence rapportés à l'environnement. Or, l'hypothèse défendue ici, est que l'animal a nécessairement *une notion de ce qu'il est* à travers les propriétés perceptives qui l'affectent en lui permettant de s'ancrer dans une vitalité en connexion avec le milieu. Plus précisément, l'aperception que l'animal a de lui-même correspond à une connaissance de son propre mode de fonctionnement qui établit lui-même une relation avec l'environnement.

53 Simondon (G.), 2013 (1958, thèse de l'auteur correspondant à l'ouvrage cité), L'individuation à la lumière des notions de forme et d'information, Millon, « Collection Krisis », p. 215. La très grande richesse de cet ouvrage nécessiterait d'amples développements que nous ne pouvons mener à bien dans le cadre de cet article.

54 Ibid.

55 Merleau-Ponty, (M.), 1990, (1ère édition 1942) La structure du comportement, PUF, « Quadrige », p. 137.

Dans la focalisation visuelle par exemple, l'organisme est présent non seulement à ce qu'il voit mais en même temps à *lui-même* qui voit, ces deux aspects de la réalité étant indissociables dans l'acte de vision qui unifie le sujet et l'objet dans l'indivisibilité de l'acte perceptif.

Dans la théorie écologique de la perception visuelle proposée par Gibson, l'animal est en contact de flux optiques indissociables de la manière dont il est conditionné pour percevoir le monde : « l'animal et l'environnement sont considérés comme intimement liés : le produit final de la perception n'est pas conçu comme une représentation interne du monde visuel – un " percept " – c'est plutôt l'animal qui est vu comme détectant des offrandes. L'offrande d'une surface ou d'un objet dans l'environnement consiste en ce que ce dernier offre à l'animal – que ce soit être capturé, mangé, foulé ou pris comme support pour s'assoir. On peut retracer les origines de la notion d'offrande en remontant aux gestaltistes, en particulier Koffka et sa notion de " caractère de demande " d'un objet : " Pour l'homme primitif, chaque chose dit ce qu'elle est et ce à quoi elle sert : (…) un fruit dit ' mange-moi ' ; l'eau dit ' bois-moi ' ; le tonnerre dit ' crains-moi ' (…). Le tronc d'un arbre abattu, de surface et taille adéquates, offre à l'homme de ' s'y asseoir ', à la grenouille de ' sauter dessus ' (…). »[56] Les structures optiques pertinentes constituent de véritables patterns incitatifs et l'animal, sollicité par le monde extérieur *et* la manière dont il l'appréhende, accède simultanément à des flux optiques circonstanciés et à ce qu'ils expriment en termes de signification vitale. Percevoir c'est donc pour l'animal déjà agir et agir c'est, réciproquement, continuer à percevoir. Dans cette relation double se joue l'aperception première que *tout* animal a de lui-même en tant qu'il ne peut qu'être simultanément agissant-percevant. C'est pourquoi, « (…) le concept d'offrandes fournit un moyen puissant de combler la brèche qui existe, dans les théories à orientation plus cognitive, entre " perception " et " action " : dans le cadre de la théorie des offrandes, la perception est une invitation à agir et l'action est une composante essentielle de la perception. »[57]

L'animal se rapporte ainsi à son environnement optique pour élaborer des conduites dont il a la notion interne dans la mesure où son autonomie permet d'orienter les mouvements. Imaginons un insecte ou un oiseau évoluant à travers des objets de texture variable, volant entre les branches d'un arbre. Dans ces circonstances, « à chaque instant, le champ visuel contient une masse d'angles lumineux solides reflétant des éléments de texture tels que des veines ou des taches de couleur sur les feuilles. Le champs optique stationnaire ne précise pas la limite des objets ni leur distance relative ; cependant, dès que l'insecte bouge, cette information du champ de flux optique lui devient accessible. Les bords des feuilles sont définis par des différences de rapport entre flux de texture optique : le flux optique reflété par une feuille lointaine se déplace à une vitesse moindre que celui provenant d'une feuille proche. Selon Gibson, le balayage et le découpage de texture permettent de spécifier l'espace à trois dimensions. »[58] C'est pourquoi, en adhérant à son environnement optique, l'animal adhère à son être actif-perceptif et cette expérience lui donne la notion *de ce qu'il vit de la manière dont il le vit* ; il en a donc une connaissance sur la base d'un système sensoriel. Cette aperception de sa propre réalité comme système sensoriel et complexion biologique et corporelle caractérise tout animal, la conscience étant alors intégrée au corps percevant « finalisé » par l'action

56 Bruce (V.) et Green (P. R.), 1993, La perception visuelle. Physiologie, psychologie et écologie, Presses universitaires de Grenoble, « Sciences et technologies de la connaissance », pp. 284/285.

57 Ibid. On sait ainsi que certains insectes disposent d'une capacité à modifier leurs mouvements de marche ou de vol en fonction des mouvements de rotation d'une surface englobantes assimilable à un flux optique, montrant la parfaite synchronicité entre le mouvement des flux optiques et les mouvements réactionnels de l'animal et attestant de l'existence d'un lien intégré avec le monde extérieur.

58 Ibid., p. 305.

entreprise dans une relation circonstanciée à l'environnement.

Dans le même ordre d'idées, une analyse intégrée de la vision nous est proposée par David Marr pour lequel il faut s'interroger sur les systèmes visuels en s'intéressant à leurs objectifs : ainsi, « (…) Comme les vertébrés passent une grande partie de leur temps à se mouvoir et à rechercher de la nourriture, l'identification et la description de formes tridimensionnelles, afin de les éviter sans encombre ou de s'en emparer pour les examiner plus facilement, constituent une des tâches essentielles de leur système visuel. Par conséquent, un des objectifs du système visuel de la grenouille est de repérer des proies en vol dont l'image bidimensionnelle se forme sur la rétine, alors que le système visuel d'une mouche sera à la recherche d'une surface suffisamment vaste où se poser, et que les animaux supérieurs tireront de l'image rétinienne bidimensionnelle les descriptions d'objets à trois dimensions. »[59] Les systèmes visuels dépendent donc du style de vie et de l'activité principale de l'animal dans la relation à un environnement médiatisé par les caractéristiques intrinsèques de l'acte de vision. Le type de recherche visuelle et les caractéristiques de l'activité de l'espèce sont inséparables. Si l'on considère l'activité de l'animal « voyant » et se mouvant dans l'environnement, cette recherche et cette activité sont co-données et imbriquées au point qu'il est impossible de les séparer, ce qui revient à dire que l'animal se meut d'une manière spécifique parce qu'il voit comme il voit (système visuel spécifique) et voit comme il voit parce qu'il est conditionné par une finalité globale qui lui confère son originalité organique d'espèce (« objectif »). Bien que déterminé par le système visuel, cet accomplissement bénéficie d'une marge d'interprétation fonction de la variabilité et de l'imprévisibilité relative de l'environnement. Ainsi, Pour un animal visuel tout se joue au niveau des formes[60] extraites de l'environnement sur la base d'un système visuel auto-conditionnant qui ne dépend ni d'un apprentissage ni de conditionnements antérieurs. Ces formes constituent autant de repères saillants ayant valeurs de *signes biologiques premiers* qui permettent à l'individu de « savoir » ce qu'il a à faire dans des contextes changeants.

# 4. Cognition et réflexivité

### 4.1 Connaissance empirique et subjectivité

L'univers perceptif, formant une succession d'éléments s'enchaînant par associations continues, *anime* l'animal qui accède alors à la profondeur d'un monde « double », les perceptions étant associées à d'autres impressions non immédiatement perceptibles mais présentes et vives en tant que réalités motivationnelles, subjectives ou idéelles associées.

Interprétant un monde devenu signifiant par l'état ou la représentation qui accompagne la perception, éprouvant des passions et sensations déterminées par ceux-ci, il inscrit l'univers perceptif dans sa propre perspective, dans un cycle de phénomènes dont il est *partie constitutive* et, paradoxalement, *détachée* en tant qu'il le génère. En fonction des situations, crainte et espoir sont ainsi activés. En-deçà de toute question explicite sur les phénomènes en jeu, l'animal en reste en effet à l'expérience charnelle d'une rationalité qui l'accapare et le seul recul cognitif en jeu provient de l'emboîtement de la cause et de l'effet qui provoque l'impression sensible et l'idée qui lui est concomitamment associée.

---

59 Rosenfield (I.), 1989, L'invention de la mémoire. Le cerveau, nouvelles donnes, Eshel, p. 113.

60 Rappelons que D. Marr, inspiré par les travaux de la psychologue Warrington, amorça ses recherches sur la base de l'idée qu'il était possible de reconnaître des formes d'objets *pour elles-mêmes*, indépendamment du rapport qu'elles entretiennent avec la fonction et l'utilité d'un objet. Ces formes, ayant une valeur de reconnaissance cognitive en elles-mêmes peuvent jouer sans l'intervention de traces mnésiques et de processsus de remémoration.

Impression et lien de causalité étant co-donnés surgissent dans une expérience où passé et présent s'emboîtent sans qu'aucune objectivation puisse être mise en évidence. Cette pensée liée aux effets d'une « impression intérieure » est rendue possible par *la ressemblance* répétée au cours d'associations similaires d'éléments de perceptions. Si tout animal connaît ainsi des signes, il ne parvient pas nécessairement au niveau de la connaissance de causes pour expliquer ce qu'il observe et lui donner une valeur cognitive. Comme l'écrit P. Jouventin, « (…) les chiens et les chats sont un peu comme des gens superstitieux qui tentent d'établir des liens signifiants pour interpréter le monde et avoir prise sur lui. Nos compagnons sont, eux, perpétuellement à la recherche de corrélations qui leur permettront, non pas de comprendre, mais de prévoir ce qui va arriver. Ils ne se préoccupent pas des mécanismes, seulement du résultat immédiat. »[61]

À certains égards l'animal devient ce qu'il a expérimenté : son individualité, considérée dans sa singularité, est liée à la vie qu'il a menée qui est intégrée aux structures corporelles, aux comportements et aux modes de réactivité qui le caractérisent.

### 4.2 Cognition et individualisation. Singularité et conscience

Ce « recul », manifesté dans la structure corporelle associée au comportement et à l'acte qu'il implique, ne permet pas de prise de distance suffisante pour permettre un détachement ou une généralisation à l'égard de son corps. La structure spatio-temporelle de l'expérience permet à l'animal de se singulariser et de s'individualiser mais pas d'accéder au niveau du jugement abstrait. Sur la base d'une modalité du raisonnement déterminé par des inférences vécues, l'animal, assimilant des données de son environnement, ne peut pour autant considérer ce dernier comme une réalité propre en face de laquelle il subsisterait comme sujet connaissant. Ainsi écrit Hume, « Les bêtes ne perçoivent certainement aucune connexion réelle entre les objets. C'est donc par expérience qu'elles infèrent un objet d'un autre. Elles ne peuvent jamais établir par des arguments la conclusion générale que les objets dont nous n'avons pas eu l'expérience ressemblent à ceux que nous avons expérimentés. C'est donc au moyen de l'accoutumance seule que l'expérience agit sur elles »[62]. Des liens de contiguïté mémorisés s'établissent alors. Une forme de rationalité immanente au réel vécu prend forme grâce à la mémoire[63]. Cette rationalité naturelle est mise en œuvre sans opération de la raison considérée comme faculté distincte projetant sur la réalité ses propres structures *a priori* [64]. L'animal s'en

61 Jouventin (P.), 2014, Trois prédateurs dans un salon. Une histoire du chat, du chien et de l'homme, Belin, p. 88.

62 Ibid.

63 « La simple conscience de relations écrit Cassirer ne peut donc être considérée comme un caractère spécifique de la conscience (…) humaine. Cependant, il existe bien chez l'homme un type particulier de pensée relationnelle sans équivalent dans le monde animal. Chez l'homme s'est développée l'aptitude à isoler les relations – à les considérer dans leur signification abstraite. Pour saisir cette signification, l'homme ne prend plus appui sur les données sensibles concrètes, sur les données visuelles, auditives, tactiles, kinesthésiques. Il considère ces relations " en elles-mêmes " ». (Cassirer E., 1975. Essai sur l'homme, Paris, éd. de minuit, coll. « Le sens commun », p. 62.)

64 Certaines catégories de la raison humaine correspondent peut-être chez les animaux à des notions acquises et à des modalités du connaître comparables sur le plan de la mise en forme du donné perceptif. Ainsi, certains pourraient « pratiquer », sans nécessairement les objectiver, deux des trois catégories de la Relation énoncées par Kant : celle de la causalité ou encore celle qui distingue le substantiel persistant de l'accidentel contingent et changeant dans les phénomènes. Les catégories de la Quantité portant sur les concepts d' « unité », de « pluralité » et de « totalité », ou encore celles de la Modalité portant sur les concepts d' « existence/non existence », de « possibilité/impossibilité », et de « Nécessité/Contingence », et enfin celle de Qualité portant

tient au contraire à la structure d'une réalité qui émerge d'une rationalité expérimentée, construite dans la durée et sur la base de rapports circonstanciés à l'espace et au temps.

L'influence de la coutume sur l'imagination est à l'origine d'un « acte de l'esprit » que Hume appelle « croyance. » Tout comme les humains, les animaux projettent des croyances sur leur environnement et leurs perceptions. Leur univers est celui d'une *doxa* fragile et souvent évanescente et non d'une *épistémè* sûre de son objet. Il y a une naïveté cognitive de l'animal et une forme de « fétichisme » au sens de Comte, dans la mesure où le monde matériel est sans cesse animé par des croyances et, en particulier, des attentes et des craintes. Ce monde devient habité de fictions curieusement proche de ce que Ernst Bloch appelle « attente » dont il dit, dans *Le principe espérance*, qu'elle est une « crainte dans l'espérance. »[65] Le monde est essentiellement constitué de croyances lui conférant une densité et une complexité vécues qui le rendent digne de curiosité et d'attention et semble ainsi préfigurer certains aspects d'une attitude réfléchie et consciente surtout si l'on songe au surgissement d'un sens interrogatif, d'expériences décalées et problématiques liés à cet univers de croyances animées.

### 4.3 Schèmes et pensées animales

Piaget appelait « *schème* d'action » « ce qui, dans une action, est (…) transposable, généralisable ou différenciable d'une situation à la suivante, autrement dit ce qu'il y a de commun aux diverses répétitions ou applications de la même action.»[66] Dans le prolongement de cette définition, nous appellerons schème un ensemble organisé et signifiant de représentations, d'affects et de traces motrices ou kinesthésiques amalgamés susceptibles d'être réactivés et d'affecter à nouveau l'animal. Des fragments de représentations et d'affects d'intensité variable liés à des contextes passés déploient une cohérence « schématique » d'ensemble associée à une signification ou à des significations convergentes et « intelligibles » pour l'animal qui en est affecté dans des circonstances où le schème prend une valeur cognitive déterminante et utile à l'action. Il y a alors reconnaissance d'amalgames signifiants. Dans sa manifestation, ce schème est antérieur à tout acte de conscience lié au cogito réflexif. Enfin, cet ensemble d'expériences amalgamées plus ou moins unifiées exprime une signification supérieure à la somme de ses constituants : la totalité schématique manifestée ne peut être reproduite par la simple addition de ses éléments constitutifs.

---

sur les concepts de « réalité », de « négation », de « limitation » ne semblent pas faire partie de notions cognitives assimilables par les animaux en raison de la nécessité de les concevoir a priori, de manière entièrement abstraite c'est-à-dire indépendamment de tout support concret. Certains animaux semblent malgré tout comprendre la signification opérationnelle d'une catégorie proche de celle qui, pour nous, oppose « existence » et « non existence » sans parvenir au niveau d'une conceptualisation complète qui permettrait de saisir la valeur du concept signifiant en lui-même et dans toutes les occurrences possibles. Ainsi, « une femelle chimpanzé ainsi que le perroquet Alex se sont montrés capables d'utiliser un signe algébrique ou le vocable « none » pour désigner un ensemble vide dans des tâches de dénombrement d'objets. Le chimpanzé est en outre capable de résoudre correctement des additions avec utilisation du nombre 0 (2 + 0 = 2 par exemple) une fois la signification du symbole acquise. Cependant, Alex s'est avéré incapable de généraliser le sens de " none " dans de nouvelles situations et le chimpanzé n'a pas pu placer le symbole 0 avant le symbole 1 sans entrainement supplémentaire lorsqu'on lui a demandé de classer les différents symboles correspondant aux nombres appris par ordre croissant. Il semble donc que le 0 signifiait tout simplement " rien " plutôt qu'un nombre. » (Darmaillacq (A.-S ) et Dickel (L.) (sous la direction de), 2018, Cognition animale. Perception, raisonnement et représentations, Dunod, « Sciences Sup », pp. 135/136)

65 Bloch (E.), 1976, Le principe espérance, 1, Gallimard, « Bibliothèque de philosophie », p. 297. « (…) La crainte aussi bien que l'espérance, la crainte dans l'espérance, l'espérance dans la crainte, sont au départ quand l'homme n'intervient pas, aussi bien justifiable l'une que l'autre, face à cet " en-suspend " réel. » (Ibid., pp. 297/298).

66 Piaget (J.), 1973, Biologie et connaissance. Essai sur les relations entre les régulations organiques et les processus cognitifs, Gallimard, « Idées », p. 23.

Cette signification schématique pose le problème biologique et épistémologique de savoir comment une forme générique peut être rapportée à une matière changeante d'une part, et à des expériences toujours singulières d'autre part. Le schème, bien que réactualisé sur une base matérielle, constitue en effet une structure dotée d'un *potentiel de reconnaissance* qui place, en les valorisant, les éléments de perception dans une perspective déterminée. Du schème émerge l'unité d'une signification ayant valeur de connaissance qui apparaît à l'animal agissant. Au moment de son actualisation celui-ci apprend en effet quelque chose du réel perçu. Il établit une relation signifiante avec une réalité différenciée qui prend une forme déterminée. Est-il l'expression d'une forme de vérité ? Peut-on l'assimiler à une activité intentionnelle ? Comment caractériser sa fonction cognitive ? Sa signification atteste de l'existence d'un monde privé et vraisemblablement d'une subjectivité susceptible d'être actualisée au moment de l'action « dirigée » par les perceptions et stimuli. La question est de savoir s'il s'agit uniquement d'une subjectivité passive. Ce que l'animal éprouve en présence de ses schèmes relève nécessairement d'une singularité d'expérience. A la différence de l'homme cependant, il ne peut revenir intentionnellement sur les constituants de son propre système subjectif. La puissance évocatrice des composants du schème trouve en effet son origine dans la concomitance ou la succession d'événements associés et mémorisés. Sa structure de signification est fondée sur la ressemblance et la reconnaissance et ne semble pas pouvoir faire surgir une unité de signification indépendante.

Ces capacités fonctionnelles acquises expliqueraient l'émergence de phénomènes conscients, autrement dit d'une expérience singulière dont le vivant est affecté en situation et dont il a la capacité de saisir, en tant qu'il la vit, la teneur propre et la spécificité. Il éprouve alors des sentiments originaux qui le révèlent à lui-même et l'identifient grâce au contact évocateur d'expériences mémorisées et « cartographiées ». Sans cette capacité, l'être vivant n'accéderait pas à certaines informations biologiques qui le concernent et dont dépend sa survie[67]. N'y a-t-il pas phénomène d'émergence lorsque la complexité d'un processus atteint un certain degré et que le mode de gestion dudit phénomène ne peut plus être exclusivement local? L'animal affecté par ses propres pensées apparaît paradoxalement sujet de *ce* qui l'affecte tout en étant objet de *ses* pensées[68]. Phénomène biologique émergent[69], il s'indétermine de par sa propre activité et prend des distances à l'égard d'une matière dans laquelle il s'incarne et dont il continue à dépendre ; le conscient pénètre alors les processus matériels.

En conclusion nous dirons de l'animal qu'il *est* ses pensées plutôt qu'il n'*a* de pensées. Ces dernières ne sont pas au pouvoir de l'individu car il ne peut se les (ré)approprier sous forme de représentation indépendante ou de jugement porté sur elles. Ces pensées prendraient en

67 Les inter-relations entre aires différentes feraient naître des états spécifiques et nouveaux (émergence) affectant l'animal en transcendant leurs conditions cérébrales « locales » de réalisation et leur dépendance à l'égard du milieu externe. On entrevoit ici la difficulté pour établir des frontières étanches entre cerveau, esprit et conscience : « (…) si l'esprit est bien inhérent à l'activité du tissu cérébral, il n'en mérite pas moins une description propre, du fait du caractère privé de l'expérience qu'on en a et parce que c'est précisément ce phénomène-là que nous souhaitons expliquer. » (Damasio (A.), 2010, L'autre moi-même. Les nouvelles cartes du cerveau, de la conscience et des émotions, Odile Jacob, p.126.)

68 Voir Cosson (F.), 2017, Animalité et humanité. La frontière croisée, Ovadia, « Chemins de pensée », Quatrième partie : « Une subjectivité sans ego. »

69 « Il semble (…) difficile qu'un agrégat connecté de matière dense ne présente pas de propriété émergentes ; c'est pourquoi les théories de ces propriétés constituent un lien naturel entre les différents niveaux de description des phénomènes naturels et cognitifs. » (Varela (F.), Thomson (E.), Rosch (E.), 1993, L'inscription corporelle de l'esprit. Sciences cognitives et expérience humaine, 1993, Seuil, « La couleur des idées », p. 137.)

effet la forme d'une contemplation de représentations imagées animées par un sujet accédant à un contenu ayant une valeur ontologique *indépendante de lui et de toute expérience présente*. Bien qu'elles l'identifient, il ne peut s'identifier à ces pensées qui l'orientent et structurent ses conduites. *L'animal est ainsi déterminé par des pensées qui le mobilisent.* Cette présence de pensées différenciées fait de l'animal un sujet accédant à des états internes dont il prend connaissance et qui constituent une modalité du phénomène conscient. Immanente à l'organisme affecté, cette conscience reste confuse. Tendant en effet à se confondre avec sa pensée, celui-ci n'a pas pleinement connaissance de ce qu'il est en tant que sujet pourvu d'une identité lui apparaissant comme telle. Sans savoir qu'il est, nous dirons alors *qu'il est sachant* c'est-à-dire doté d'un savoir immanent de lui-même ordonné aux actes nécessaire à la vie. Ce cogito animal, incarné, est indissociable de l'être agissant. C'est pourquoi comme l'explique Merleau-Ponty, « L'appréhension de moi par moi est coextensive à ma vie. (…) Le *cogito* comme expérience de mon être est pré-réflexif, il ne le pose pas en objet devant moi ; par position, et avant toute réflexion, je me touche à travers ma situation, c'est à partir d'elle que je suis renvoyé à moi (…). »[70] Quoi qu'il en soit, des données expérimentales sur les catégorisations animales suggèrent l'existence de capacité d'abstraction développées chez certaines espèces. Anne-Sophie Darmaillacq et Ludovic Dickel expliquent ainsi que, « Au-delà des catégories perceptives basées sur une ou plusieurs caractéristiques physiques communes, les concepts ne sont pas définis par des similarités perceptives et impliquent donc un plus grand degré d'abstraction. »[71] L'existence de catégories abstraites semblent de mise lorsque l'on a affaire à un « transfert entre modalités sensorielles » qui attestent d'une capacité d'association entre des évènements de natures différentes. Ainsi, « (…) des babouins sont capables de faire un lien direct entre des photographies de babouins ou d'humains et des cris de babouins ou un enregistrement de la voix d'un homme. ce résultat, ajoutent-ils, suggère que les babouins " possèdent " le concept d' " humain " ou de " babouin ".»[72] De même, « Dans le cas de concepts relationnels, des relations définissent à elles seules l'appartenance à une catégorie sans intervention des caractéristiques perceptives des objets.»[73]

Insistons pour terminer sur l'importance d'éventuelles méta-représentations animales qui nécessiteraient une forme de dédoublement interne proche d'un acte de conscience : en effet, « (…) Je dois être conscient de mes représentations pour que mes représentations soient conscientes. La question cruciale qui se pose alors est de comprendre la nature de l'état méta-représentationnel. S'agit-il d'une réelle pensée sur ses propres états, c'est-à-dire d'une véritable méta-représentation cognitive, ou bien d'une forme plus simple de perception interne ? Si l'un des prérequis de la conscience est en effet une capacité méta-cognitive complexe, alors sans doute ni les jeunes enfants, ni certains animaux ne sont conscients. »[74] Une capacité métacognitive de premier degré, permettant un recul à l'égard des schèmes constitués, serait la préfiguration animale d'une forme plus élaborée de vie consciente dont l'espèce humaine est l'exemple le plus sophistiqué. De telles pensées − ce point est décisif pour qui s'interroge sur l'existence d'une forme de conscience animale − n'impliqueraient pas de prise en charge comportementale spécifique et immédiate. Cette hypothèse, si elle était avérée, conditionnerait la possibilité d'être en présence de pensées

70 Merleau-Ponty (M.), 1964, Le visible et l'invisible, Gallimard, « Tel », pp. 82/83.

71 Darmaillacq (A.-S ) et Dickel (L.) (sous la direction de), 2018, Cognition animale. Perception, raisonnement et représentations, Dunod, « Sciences Sup », p. 146.

72 Ibid.

73 Ibid., pp. 146/147.

74 De Vignemont (F.) et Sackur (J.), chapitre « Conscience » dans Collins (T.), Andler (D.) Tallon-Baudry (C.), 2018, La cognition. Du neurone à la société, Gallimard, « Folio Essais », p. 493.

qui affectent l'animal *en l'absence de situation de déclenchement*. L'animal a-t-il parfois affaire à des pensées décontextualisées dont la valeur, s'éloignant de la vie, ne serait plus directement liée à l'efficacité de l'action présente ? Dans ces moments privilégiés, sans doute rares et fugaces, ces pensées gratuites et « inutiles » ouvriraient la possibilité d'une forme *d'idéation* consistant à considérer le schème pour la signification et la valeur propres du contenu qu'il exprime en tant que détachées de l'expérience éprouvée. L'animal, si l'on excepte de vagues représentations d'origine instinctive ou des images spécifiques de recherche pré-orientant son action, ne dispose pas d'idées innées, de formes intelligibles pures, mais, partant de l'expérience, élabore ses propres idées par étape, parvenant ainsi à saisir des réminiscences idéelles le faisant adhérer à son histoire et à sa propre identité[75].

# 5. Ecologie comportementale, individualisation et singularisation

## 5.1 Subjectivité écologique, profil d'existence et phénomènes de conscience

Qu'il s'agisse de migrations de grande ampleur, de simples dispersions ou de mouvements locaux liés à la topographie, ou encore de la répartition des ressources dans l'habitat, la localisation d'un animal et ses déplacements orientés dans des espaces configurés constituent un chapitre passionnant de l'écologie comportementale qui permet de reformuler la question de la conscience animale. En effet, « La question évolutive qu'a, à tout moment, à résoudre un individu est : « " *Ai-je ce qu'il me faut à cet endroit et à ce moment ?* "»[76] La connaissance d'une topographie connotée s'avère ainsi déterminante pour l'émergence d'états émotionnels et de conduites préfigurant un phénomène d'individualisation incarnée dans la façon d'appréhender les réalités de l'environnement. De fait, certaines préférences, craintes, stratégies de fuites ou d'évitements, hésitations, mais aussi des attentes « mesurées » et des immobilisations attentives sont autant de comportements complexes − composés et donc mesurés et « maîtrisés » − impliquant une évaluation pour l'animal qui tente de composer avec ce qu'exprime pour lui l'environnement et les situations rencontrées.

Des conduites suspensives insèrent ainsi une dualité naissante au cœur de l'expérience montrant qu'il n'existe pas de collusion instantanée entre un lieu ou une situation et l'organisme attentif à ce qui s'y déroule. Le présent en effet est investi par le passé sous forme d'expériences réactualisables. Des conduites comme la prudence, l'hésitation, l'espoir, et toute conduite différée en général, expriment une tension mettant en jeu un conflit entre deux options. Tout espoir est attente c'est-à-dire anticipation associée à une connaissance préalable ; il est en même temps tension structurée vers l'avenir au titre d'une projection temporelle. On peut alors parler d'une conscience de ce qui est éprouvé. Elle correspond à l'accès à des sentiments émotionnels et à des états de choses environnantes qui sont autant de *modulation*s de la vie concrète d'animaux devenus des sujets affectés capable de s'appréhender eux-mêmes dans un environnement vécu. Le monde extérieur

75 La conception des « assemblées de cellules » proposées dès 1949 par Donald Hebb dans *L'organisation du comportement* semble aller dans le sens d'un processus d'autonomisation de certaines configurations de neurones associés constituant peut-être des schèmes. Ainsi, les assemblées de cellules sont, rappelle S. Dehaene, « (…) des groupes de neurones reliés entre eux par des synapses excitatrices, et qui ont donc tendance à rester actif longtemps après que la stimulation a disparu. " Toute stimulation qui se répète fréquemment, conduira au développement progressif d'une assemblée de cellules, une structure diffuse comportant des cellules du cortex et du diencéphale (…), capable d'agir brièvement comme un système fermé ". » (Dehaene (S.), 2014, Le code de la conscience, Odile Jacob, « Sciences », p. 242. La citation est de Donald Hebb).

76 Danchin (E.), Giraldeau (L.-A.) Cézilly (F.), 2005, Écologie comportementale. Cours et questions de réflexion, Dunod, « Sciences Sup », p. 199.

s'organise donc en perspectives individuelles jamais abstraites et acquiert une qualité en terme de vécu et d'expérience possible c'est-à-dire de crainte et d'espoir. Crainte, espoir, déception n'ont de réelle valeur vitale qu'en tant qu'ils se rapportent à un environnement circonscrit. S'ils n'étaient que des passions éprouvées toujours semblables dans leur modalité, ils se réduiraient à des réactions schématiques impraticables pour l'animal confronté à des situations réelles. D'un point de vue éco-éthologique, les catégories ambivalentes hésitation/décision, espoir (déception/satisfaction, prudence/prise de risque), jouent ainsi un rôle prépondérant dans l'approche de l'environnement. Les valences et polarisations expriment l'ambivalence de toute existence au cours de laquelle les comportements sont des compromis évaluatifs et des prises de décision liés à l'exercice d'une *subjectivité écologique* s'exprimant par le biais de sentiments émotionnels. C'est pourquoi, toute existence animale−nécessairement « spéciste » et perspectiviste−progresse par intégration de données extérieures pour les rechercher, les maîtriser, les éviter ou s'y adapter. Dès lors, cette existence, comme le montre Florence Burgat, est façonnée par une inquiétude face à l'environnement de la part d'organismes qui tendent, à travers l'affirmation d'une subjectivité écologique et d'un profil existentiel, à gagner toujours plus en liberté : en effet, « Au caractère clos de l'animalité, exprimé par le *détachement* (au sens littéral du terme) à l'égard du sol et par le mouvement spontané, appartiennent la diversité des relations avec l'environnement externe. D'où il suit que l'animal est habité par un sentiment d'inquiétude : plus grande est la clôture, plus étendu le champ des possibles, plus grande l'inquiétude. C'est bien la liberté qui s'ouvre devant les animaux, l'indétermination, l'immensité des possibles, le toujours nouveau départ, la contrainte de choisir une direction de déploiement plutôt qu'une autre. »[77] Gilbert Simondon voit, dans l'inquiétude l'avènement d'une « individualité psychologique » qui « (…) fait intervenir des normes qui n'existent pas au niveau biologique ; tandis que la finalité biologique est homéostatique et vise à obtenir une satisfaction de l'être dans un état de plus grand équilibre, l'individualité psychologique existe dans la mesure où cet équilibre, cette satisfaction, sont jugés suffisants. L'inquiétude dans la sécurité vitale marque l'avènement de l'individualité psychologique, ou tout au moins sa possibilité d'existence. L'individualité psychologique ne peut se créer par une dévitalisation du rythme vital, ou par une inhibition directe des tendances, car cela ne conduirait alors qu'à une intériorité et non à une spiritualité. L'individualité psychologique se surimpose à l'individualité biologique sans la détruire, car la réalité spirituelle ne peut être créée par une simple négation du vital. »[78] L'adaptation « psychique » peut d'ailleurs compléter et même suppléer l'adaptation vitale : « En fait, le véritable psychisme apparaît lorsque les fonctions vitales ne peuvent plus résoudre les problèmes posés au vivant, lorsque cette structure triadique des fonctions perceptives, actives et affectives n'est plus utilisable, ajoute Simondon. Le psychisme apparaît ou tout au moins est postulé lorsque l'être vivant n'a plus en lui-même assez d'être pour résoudre les problèmes qui lui sont posés. On ne doit pas s'étonner de trouver à la base de la vie psychique des motivations purement vitales (…). »[79]

En raison même de cette inquiétude qui traduit l'émergence d'un psychisme spécifique, nous admettrons que les conduites évoquées expriment des qualités individuelles différenciées relatives à des occurrences locales d'événements et, plus généralement, à une structuration spécifique de l'espace et du temps renvoyant à l'émergence de territoires singularisés. Inscrit dans son propre univers territorial, l'animal s'en détache ainsi

77 Burgat (F.), 2005, Liberté et inquiétude de la vie animale, Kimé, p. 190.
78 Simondon (G.), 2013 (1958, thèse de l'auteur correspondant à l'ouvrage cité), L'individuation à la lumière des notions de forme et d'information, Million, « Collection Krisis », p. 276.
79 Ibid., p. 166.

progressivement en se l'appropriant sur la base d'une expérience qui devient partie intégrante de ce qu'il est. Cet écart, qui procède d'une confrontation au réel et d'un apprentissage de données qui lui donnent une signification biologique et subjective, permet d'évoquer la nécessité d'une éthologie différenciée, attentive à l'émergence de singularités chez des animaux biologiquement individués mais surtout *individualisés, en quête d'un optimum éco-éthologique qui leur corresponde et signe une présence au monde originale telle une signature extériorisant la personnalité acquise du signataire*. Ajoutons que ce déconditionnement s'accompagne de l'émergence d'une vie subjective unique − capacité d'être affecté par des circonstances intégrées d'une manière unique − donnant à l'animal un profil existentiel spécifique et « reconnaissable » qui ouvre de la manière la plus directe au questionnement éthique[80]. Cette subjectivité écologique qui s'accompagne de l'émergence de nouveauté et d'originalité rend possible des processus cognitifs eux-mêmes utiles aux transformations évolutives des organismes. Concernant l'apparition de la conscience, l'un des plus grands paléontologues du 20ème siècle, Georg Gaylord Simpson, propose ainsi une perspective continuiste des plus intéressantes : « Un organisme qui a une conscience accrue de son milieu et qui réagit à une gamme plus étendue de stimuli, remarque-t-il, devient plus adaptable et souvent aussi moins rigide dans ses réactions. Cet élargissement de la portée et de la variété possibles des réactions d'un organisme isolé augmente l'indépendance de chacun en tant qu'unité et rend plus distinctes ses réactions et ses relations particulières avec les autres. En un mot, il peut être accompagné par un progrès dans l'individualisation. (…) Le progrès dans la diversité des réactions individuelles et dans l'individualisation à l'intérieur de l'espèce a été poussé par l'évolution de l'homme jusqu'à des niveaux jamais atteints auparavant.»[81]

### 5.2 Perception, signification et émergence d'états conscients

Certains animaux disposent-ils de la capacité de s'abstraire de leur environnement ? Comme l'ont montré les recherches du prix Nobel Gerald Edelman, le fonctionnement cérébral met en œuvre des « scènes » qui contribuent à faire émerger une sphère où se manifestent des états conscients particuliers nécessaires à la reconnaissance de lieux et d'affects qui y sont liés lors de la réitération de l'expérience. Ce dialogue permanent entre les aires cérébrales, qui inclut un retour vers les aires d'origine, a été nommé « réentrée ».[82] Sur le plan neuronal, « La modélisation de réseaux de neurones montre que la réentrée permet des calculs sophistiqués qui convergent vers l'interprétation statistique la plus probable d'une scène visuelle. Chaque groupe de neurone agit comme un statisticien expert, qui collabore avec ses collègues afin d'expliquer les données sensorielles dans leurs

80 L'émergence d'états internes attestant de l'existence de formes de conscience démultipliera à l'avenir l'émerveillement devant l'inventivité du vivant. Elle nous conviera à une meilleure compréhension des profils existentiels exprimés par les animaux tant sauvages que domestiques. Les processus en jeu nous invitent ainsi à *une réflexion transdisciplinaire* intégrant le fait majeur d'une diversité qui en fait naître d'autres dans un cycle indéfini et ouvert qui intègre l'espace, le temps et une matière vivante transformable. De même, se pose la question ontologique et éthique de la valeur donnée à des êtres *affectés par des états conscients singularisés qui les renvoient à eux-mêmes*. Nous retiendrons alors avec le biologiste Jacques Blondel que « La valeur d'un élément de diversité est incommensurable, dès lors qu'il existe en lui-même et qu'il se situe dans un enchaînement spatio-temporel de causes et d'effets qui nous échappe à peu près totalement, surtout dans sa dimension de différenciation évolutive se projetant dans un futur imprévisible (…).» (Blondel (J.), 2012, L'archipel de la vie. Essai sur la diversité biologique et une éthique de sa pratique, Buchet Chastel, « Écologie », p. 174).

81 Simpson (G. G.), 1951, L'évolution et sa signification, Payot, « Bibliothèque scientifique », p. 228.

82 Dehaene (S.), 2014, Le code de la conscience, Odile Jacob, « Sciences », p.245.

moindres détails. »[83] Ce phénomène donne à l'individu une manière d'être singulière dans son rapport à l'environnement − forme d'idiosyncrasie assez proche de la notion de personnalité. L'animal puiserait ainsi des informations pertinentes pour le présent sur la base d'indices fragmentaires : dans « la structure neuronale du réseau hippocampique, des propriétés qui permettraient de retrouver un épisode ou une combinaison de sensations avec seulement une partie de l'information initialement mémorisée sont ainsi mises en évidence. »[84] (…) Un mécanisme particulier, celui de « la structure auto-associative », permettrait même « (…) l'utilisation d'épisodes anciens comme modèles de ce que l'action commencée peut produire. »[85] Dans une perspective plus large, Edelman explique que la « conscience primaire » est un « présent remémoré » : « (…) selon la TSGN en effet (théorie générale des groupes de neurones), la conscience primaire contribue à abstraire et à structurer les modifications complexes qui surviennent dans un environnement faisant intervenir des signaux multiples et parallèles. Et même si certains de ces signaux n'ont aucun lien de causalité direct entre eux dans le monde extérieur, ils peuvent constituer, *pour l'animal*, des indicateurs significatifs d'un danger ou d'une récompense. Cela est dû au fait que la conscience primaire relie leurs caractéristiques en fonction de ce qui *compte* pour l'animal, ce qui à son tour dépend de l'histoire passée et de l'histoire de cet animal.»[86] Cette modalité du phénomène conscient correspond à l'émergence d'un phénomène cérébral d'auto-production *singularisé* dans ses composantes anatomiques et son fonctionnement physiologique. Ce lien entre fonctionnement individualisé et singularisé montre qu'une conscience peut émerger pour l'individu « comportemental » tout en étant dépendante, dans sa genèse, d'états cérébraux originaux liés à une histoire intégrée.

L'émergence du conscient doit donc s'analyser en terme d'écologie cérébrale jetant ainsi les bases d'une écologie de la conscience reflétant les conditions matérielles de la vie auxquelles, pourtant, elle ne se réduit pas. Cette conception pourrait être étayée par la théorie de l'espace de travail neuronal global proposée par Stanislas Dehaene, Jean-Pierre Changeux et Lionel Naccache[87] : « (…) La conscience n'est rien d'autre que la diffusion globale d'une information à l'échelle de tout le cerveau écrit S. Dehaene. Tout ce dont nous prenons conscience, nous pouvons le garder à l'esprit longtemps après qu'il a disparu de nos organes sensoriels. Une fois l'information acheminée vers l'espace de travail, elle y reste stable, indépendamment du moment et du lieu où nous l'avions initialement perçue. Nous pouvons alors l'utiliser de mille manières, et en particulier l'expédier aux aires du langage, donc la nommer (…). Nous pouvons également la stocker dans notre mémoire à long terme ou l'intégrer à nos plans d'action, quels qu'ils soient. La dissémination flexible de l'information caractérise l'état conscient.»[88] Il est remarquable qu'un état conscient émergeant puisse être en corrélation avec les nécessités évolutives : « Grâce à l'espace de travail neuronal global, nos modules cérébraux peuvent partager librement certaines informations. Leur disponibilité globale est précisément ce que nous appelons " conscience ". Les avantages évolutifs que confère cette organisation sont évidents. La modularité est utile parce que chaque domaine de connaissance nécessite des

83 Ibid.

84 Berthoz (A.), 1997, Le sens du mouvement, Odile Jacob, « Sciences », p. 141.

85 Ibid.

86 Edelman (G. M.), Biologie de la conscience, 1992, Odile Jacob, « Sciences », pp. 160/161.

87 Dehaene (S.), 2014, Le code de la conscience, Odile Jacob, « Sciences », p. 227. « Le psychologue Bernard Baars l'appelle " espace de travail global " : un système interne, découplé du monde extérieur, au sein duquel nous sommes libres de créer nos propres images mentales et de les transmettre à n'importe quel processeur cérébral spécialisé. » (Ibid., p. 228)

88 Ibid., p. 228.

microprogrammes dédiés : la programmation d'un geste exige des calculs différents de la reconnaissance d'un paysage ou de la récupération d'un souvenir en mémoire. Prendre une bonne décision, par contre, nécessite souvent de briser cette modularité en recueillant un maximum d'informations les plus diverses.»[89] L'exemple donné par S. Dehaene est particulièrement éclairant pour notre propos. Il montre en effet à quel point certains animaux doivent rendre disponible et associer des informations de nature différentes en lien avec ce qu'ils cherchent et éprouvent au contact de leur environnement : « Imaginez écrit-il, un éléphant assoiffé, seul dans la savane. Sa survie dépend de la découverte d'un point d'eau. Sa décision de se mettre en route vers telle ou telle destination lointaine doit se fonder sur un usage optimal de toutes les informations disponibles : cartographie de l'espace environnant, reconnaissance des sentier et des arbres, souvenir des succès et des échecs antérieurs… (…) La conscience aurait ainsi évolué, il y a des millions d'années, afin d'extraire et de diffuser au reste du cerveau un maximum d'informations pertinentes initialement confinées dans des circuits cérébraux spécialisés. » [90] La fonctionnalité cérébrale ainsi acquise, susceptible de changer, constitue, pour l'animal, une première couche de singularité et une part essentielle de son identité relationnelle vivante autrement dit de sa subjectivité. La façon dont il s'appréhende, au moment où il éprouve une sensation en percevant, a une dimension ontologique propre dans la mesure où *l'être est révélé à sa subjectivité unique* qui saisit, met en forme et agit en conséquence de ces opérations. Se percevoir « animal », être conscient qu'on est – dans la trame même de sa propre existence sans savoir ni *ce* qu'on est ni *qui* on est –, c'est éprouver une expérience spécifique dépendant de l'espèce à laquelle on appartient *et* de sa propre histoire perceptive et écologique.

Cette réflexion peut être prolongée par des considérations ayant trait à l'*émergence neuro-écologique d'un monde privé*. On sait que certains éléments de perception sont associés à d'autres qui, mémorisés, servent de repères et constituent une opération cognitive de reconnaissance pour l'animal qui, dès lors, « sait » où il se trouve. Ces systèmes de repérage, sont « égocentrés » ou « allocentriques », ces derniers entraînant une « décentration perceptive. »[91] Des comportements induits par des processus de mémorisation impliquent en effet certaines parties du cerveau : «(…) l'hippocampe serait ainsi utilisé par l'animal pour réactualiser des cartes spatiales de l'environnement fondées sur des cellules de lieu et calculer les tours et détours nécessaires pour atteindre la destination. »[92] Ces neurones «  déchargent chaque fois que l'animal passe par un endroit particulier de l'arène, quelle que soit son orientation (…). Ils couvrent un domaine spatial qui varie mais ne dépasse pas quelques dizaines de centimètres de rayon. L'ensemble des cellules de lieu serait donc susceptible de constituer " une carte cognitive de l'espace ". » [93] De plus, ils sont sensibles *à un point de vue spatial particulier* dans la mesure où certains neurones sont activés en relation avec *la direction du regard* vers un point particulier de l'espace. Ce repérage peut être lié à la pratique potentielle de l'animal qui se projette ainsi dans un espace topographique chargé de significations pratiques : « (…) le mouvement du corps est placé dans l'espace allocentrique et, surtout, référé à l'usage que pourrait faire l'animal de telle ou telle partie de l'espace. En d'autres termes, le mouvement du corps est étiqueté sur l'espace. C'est le cerveau qui étiquette ses perceptions en fonction de ses

89 Ibid., p. 234.
90 Ibid.
91 Ibid, p.140.
92 Ibid.
93 Ibid., p.139.

intentions et de ses buts (…). »[94] Ces potentialités qui s'expriment sous forme de conduites sophistiquées et autonomes renvoient à des fonctions cérébrales précises. Du fait de cette décentration, une prise de distance s'opère à l'égard du monde extérieur ; l'animal investit l'espace perçu et se connaît lui-même en tant qu'il connaît l'environnement sous forme de repères topographiques réactualisables. Le monde extérieur ainsi codé lui donne les éléments qui lui permettent de reconnaître des contextes signifiants et, à travers eux, de se reconnaître lui-même et d'éprouver les sensations et les états propres liés à cette reconnaissance. Comme chez l'homme, la manière d'habiter des espaces connus ne peut manquer de générer des sentiments individuels. Ajoutons que certains animaux disposent de la capacité d'objectiver l'environnement sur la base d'une mise en perspective surplombant les éléments discriminés. Ces distinctions relèvent de données dépendantes d'une activité de réactualisation de ses conduites et d'une histoire accessible sous forme de *traces* ayant valeur de *signes* induisant un phénomène de reconnaissance et des comportements adaptatifs. Cette activation, apparentée à l'acte révélateur d'une configuration assimilée par l'animal qui en prend connaissance par une opération intérieure, peut être assimilée à un acte de prise de conscience autrement dit à la présence à soi d'une connaissance constituée réactualisable. Chez le singe, les neurones ne coderaient pas « le lieu où se trouve l'animal dans la pièce, mais une zone de l'espace où le neurone serait activé par un indice pertinent pour l'animal. L'activité des neurones de lieu ne ferait que révéler une zone de l'espace où ceux-ci seraient impliqués dans une plus grande variété de comportements.»[95] Dans ce contexte où se joue une complexité nouvelle, l'objectivation se fait bien sur la base de traces coordonnées à des éléments moteurs qui traduisent l'emprise de l'animal sur les réalités perçues et éprouvées et expriment, dans leur émergence cérébrale, un rapport *constitué* – et donc *objectivé* – au monde environnant, rapport susceptible d'être « recyclé » c'est-à-dire réactivé en induisant un acte ou un phénomène de connaissance cérébrale original affectant subjectivement l'individu.

Une approche cognitiviste de la perception étayera notre propos. Dans le phénomène de la perception visuelle considérée comme « structure hiérarchique », trois niveaux fonctionnels interagissent pour intégrer les phénomènes perçus. Comme l'explique Marion Luyat, le niveau sensoriel se réfère aux « tout premiers traitements effectués par le système visuel sur les images. »[96] Les éléments du niveau sensoriel « sont sous-tendus par des mécanismes neuro-sensoriels » et seraient liés « aux caractéristiques physiques du stimulus. »[97] Dans un deuxième temps, se met en place un « niveau perceptif, ou niveau configurationnel » qui met en correspondance les éléments du niveau sensoriel. Enfin, « Un niveau cognitif ou niveau conceptuel qui permettrait de donner une signification aux différentes organisations avec leurs positions relatives les unes aux autres indépendamment de l'observateur. Les objets seraient alors reconnus, identifiés même s'ils sont partiellement vus. Ils seraient représentés dans un référentiel allocentré de l'espace avec leurs positions relatives les unes aux autres indépendamment de l'observateur. Ce niveau interagirait avec nos représentations en mémoire et nos connaissances antérieures. »[98] Un comportement perceptif intégré est alors en jeu car ces niveaux interagissent entre eux ainsi qu'avec un « contexte émotionnel et motivationnel particulier ». La cognition permet ainsi un décalage

94 Ibid., p. 140.

95 Ibid., p. 139.

96 Luyat (M.), 2009, La perception, Dunod, « Les topos », p. 46.

97 Ibid., p. 46.

98 Ibid., p. 47. Sur ce sujet voir également Darmaillacq (A.-S.), Dickel (L.) et al., Cognition animale. Perception, raisonnement et représentions, Dunod, « Sciences Sup ».

entre la situation présente et la mémorisation du passé à travers une reconnaissance propice à la saisie d'états physiologiques ou émotionnels révélant l'organisme à lui-même dans l'immanence et l'efficacité propre des perceptions et, plus généralement, d'une relation orientée et structurée à l'environnement. Cette conception peut être rapprochée de l'interactionnisme de Karl Popper développé dans ***The Self and its Brain*** écrit en collaboration avec le neurophysiologiste John Eccles. Les entités du monde 3, explique-t-il, sont acquises, « objectives » et relativement durables ; elles ont une structure et des propriétés propres indépendantes du monde physique (monde 1) et de celui des émotions, des sentiments et de la subjectivité (monde 2). Ces entités sont saisies chez l'homme par un acte de l'esprit susceptible de les réactualiser et de réinvestir leurs significations. Elles ont une fonction biologique manifeste dans la mesure où elles constituent un cadre d'appréhension fixe d'une réalité diverse dont les fluctuations sont un obstacle à la cohérence comportementale, à l'unité des attitudes et à l'élaboration de conduites autonomes. Si l'on suit les analyses poppériennes, l'animal pourrait fort bien disposer de connaissances acquises et intégrées appartenant au monde 3. En effet, on peut penser que l'animal objective ses catégories acquises ainsi que des éléments de connaissance liés aux régularités de l'environnement − aux conjonctions régulières d'événements observées notamment − qui appartiennent à ce troisième monde. Tout animal disposerait ainsi de catégories « théoriques » tirées de l'expérience et constituant des anticipations hypothétiques de la même manière que nos relations au monde sont le reflet de théories et, en particulier, d'attentes et d'anticipations : « A mes yeux écrit Popper, la perception d'une forme est pour l'essentiel la même chose qu'une hypothèse (…). Et la notion d'attente est de la plus haute importance pour moi. Je vois dans l'attente, l'attitude de l'animal qui anticipe sur ce qui va arriver, le début d'une hypothèse, l'amorce de la théorie. »[99] « Quelle est alors, demande Popper, la différence entre la conscience animale et la conscience humaine ? La différence est le langage humain qui permet la critique. Tout le reste est secondaire. »[100] La différence ne semble être que de degré et pas de nature. En effet, les entités du monde 3 se rapportent à des éléments stables – simples images réactualisables, percepts et concepts « théoriques » par exemple – intégrés comme structures de signification liées à une donnée d'expérience. Ces entités émergeantes s'ancrent dans la réalité biologique : comme l'écrit le neurologue Franz Seitelberger, « ce monde des produits de l'esprit marqués par le langage correspond au monde 3, autrement dit à la somme et à l'ordre d'équivalents de la réalité, autonomes, extériorisés, produits par le cerveau et qui, coupés de l'organe qui les a portés, constituent une couche particulière de la réalité, une couche de réalité fonctionnelle *méta-organique*, qui engendre ses organisations propres, présente ses propres lignes d'évolution, constituant dans leur ensemble ce que nous appelons la culture »[101] Le champ de conscience animal est plus restreint que chez l'homme car il n'accède pas à son passé sous une forme objectivée par le langage. Il n'a pas la capacité de se projeter dans le passé et d'atteindre le niveau de la description, de la narration, de l'interprétation et de la relativisation critique qui est celui du « Soi autobiographique » qui, explique Damasio, « (…) se constitue à partir de la réactivation et de la présentation sous une forme cohérente d'ensembles choisis de souvenirs autobiographiques. »[102]

99 Popper (K.) dans Lorenz (K.), Popper (K.), L'avenir est ouvert, 1990, Flammarion, p. 31.

100 Popper (K.) dans Lorenz (K.), Popper (K.), L'avenir est ouvert, 1990, Flammarion, p. 110.

101 Seitelberger (F.) dans Ibid., p. 104.

102 Damasio (A.), 2002, Le sentiment même de soi. Corps, émotion, conscience, Poches Odile Jacob, p. 255/257.

## Références


- Azouvi (F.), 1995, Maine de Biran. La science de l'homme, Vrin, « Histoire de la philosophie ».
- Berthoz (A.), 1997, Le sens du mouvement, Odile Jacob, « Sciences ».
- Blondel (J.), 1993, Biogéographie. Approche écologique et évolutive, Masson, « Collection d'écologie ».
- Blondel (J.), 2012, L'archipel de la vie. Essai sur la diversité biologique et une éthique de sa pratique, Buchet Chastel, « Ecologie ».
- Bruce (V.) et Green (P. R.), 1993, La perception visuelle. Physiologie, psychologie et écologie, Presses universitaires de Grenoble, « Sciences et technologies de la connaissance ».
- Burgat (F.), 2005, Liberté et inquiétude de la vie animale, Kimé.
- Canguilhem (G.), 1985, La connaissance de la vie, Vrin, « Problèmes et controverses ».
- Canguilhem (G.), 2002, Etudes d'histoire et de philosophie des science, Vrin, « Problèmes et controverses ».
- Cassirer (E.), 1975. Essai sur l'homme, Paris, éd. de minuit, coll. « Le sens commun ».
- Collins (T.), Andler (D.) Tallon-Baudry (C.), 2018, La cognition. Du neurone à la société, Gallimard, « Folio Essais ».
- Comte A., 1975 (1re édition 1838), Cours de philosophie positive, Hermann.
- Cosson (F.), 2017, Animalité et humanité. La frontière croisée, Ovadia, « Chemins de pensée ».
- Cosson (F.), 2007, L'animal médiateur de l'humain, Revue internationale de psychosociologie, 30 (Vol. XIII).
- Damasio (A.), 2010, L'autre moi- même. Les nouvelles cartes du cerveau, de la conscience et des émotions, Odile Jacob.
- Damasio (A.), 2010, L'autre moi- même. Les nouvelles cartes du cerveau, de la conscience et des émotions, Odile Jacob, « Sciences ».
- Damasio (A.), 2002, Le sentiment même de soi. Corps, émotion, conscience, Poches Odile Jacob, « Sciences ».
- Danchin (E.), Giraldeau (L.-A.), Cézilly (F.), 2005, Ecologie comportementale. Cours et questions de réflexion, Dunod, « Sciences Sup ».
- Darmaillacq (A.-S ) et Dickel (L.) (sous la direction de), 2018, Cognition animale. Perception, raisonnement et représentations, Dunod, « Sciences Sup ».
- De Fontenay (E.), 1998, Le silence des bêtes. La philosophie à l'épreuve de l'animalité, Fayard.
- Dehaene (S.), 2014, Le code de la conscience, Odile Jacob, « Sciences ».
- Dennet (D.C.), 1993, La conscience expliquée, Odile Jacob, « Philosophie ».
- De Waal (F.), 1992, De la réconciliation chez les primates, Flammarion.
- De Waal (F.), 2006, Le singe en nous, Fayard, « le temps des sciences ».
- Edelman (G. M.), Biologie de la conscience, 1992, Odile Jacob, « Sciences ».
- Geluck (P.), 2005, Le chat a encore frappé, Casterman,
- Griffin (D.), 2001, Animal Minds, The University of Chicago Press.
- Heidegger (M.), 1998, Qu'appelle-t-on penser ?, PUF, « Epiméthée ».
- Hoyo (J.), Elliot (A.), Christie (D.), 2009, Handbook of the Birds of the World, volume 14, Lynx edicions.
- Hume D., 1968 (1re édition 1739/1740). Traité de la nature humaine. Essai pour introduire la méthode expérimentale dans les sujets moraux, t. 1 : De l'entendement, Paris, Aubier Montaigne, coll. « Bibliothèque de philosophie ».
- Husserl (E.), 2011, Expérience et jugement, PUF, « Epiméthée ».
- Lorenz (K.), (1975) L'envers du miroir, Une histoire naturelle de la connaissance, Flammarion, « Nouvelle bibliothèque scientifique ».
- Luyat (M.), 2009, La perception, Dunod, « Les topos ».

- Martinet (M.), (1972), Théorie des émotions. Introduction à l'œuvre d'Henri Wallon, Aubier ---- Montaigne, « Analyses et raisons ».
- Merleau-Ponty, (M.), 1990, (1ère édition 1942) La structure du comportement, PUF, « Quadrige ».
- Merleau-Ponty (M.), 1964, Le visible et l'invisible, Gallimard, « Tel ».
- Minkowski (E.), 1967 (1ère édition 1936), Vers une cosmologie. Fragments philosophiques, Aubier-Montaigne.
- Monneret (J.R.), 2006, Le faucon pèlerin, Delachaux et Niestlé, « Les sentiers du naturaliste ».
- Nagel (T.), « Quel effet cela fait d'être une chauve-souris ? » dans Hofstadler (D.) et Dennet (M.), 1987, Vues de l'esprit. Fantaisies et réflexions sur l'être et l'âme, Interédition.
- Passera (L.),1984, L'organisation sociale des fourmis, Privat, « Bios/Université Paul Sabatier ».
- Piaget (J.), 1973, Biologie et connaissance. Essai sur les relations entre les régulations organiques et les processus cognitifs, Gallimard, « Idées ».
- Plessner (H.), 2017 (1ère édition 1928), Les degrés de l'organique et l'homme. Introduction à l'anthropologie philosophique, Gallimard, « Bibliothèque de philosophie ».
- Popper (K.), 1989, La quête inachevée. Autobiographie intellectuelle, Presse Pocket, « Agora ».
- Popper (K.) et Eccles (J.), 1985, The Self and its Brain. An argument for Interactionnism, Springer International.
- Portmann (A.), 1951, La forme animale, Payot, « Bibliothèque scientifique ».
- Pradines (M.), 1928, Philosophie de la sensation, 1, Le problème de la sensation, Les belles lettres, « Publication de la faculté des lettres de l'université de Strasbourg », Fascicule 42.
- Pradines (M.), 1932, Philosophie de la sensation, 2, la sensibilité élémentaire (les sens primaires), 1, Les sens du besoin, Les belles lettres, « Publication de la faculté des lettres de l'université de Strasbourg », Fascicule 61.
- Pradines (M.), 1954, L'aventure de l'esprit dans les espèces, Flammarion, « Bibliothèque de philosophie scientifique ».
- Pradines (M.), 1946, Traité de psychologie générale,1, Le psychisme élémentaire, PUF, « Logos. Introduction aux études philosophiques ».
- Proust (J.), 1997. Comment l'esprit vient aux bêtes. Essai sur la représentation, Paris, Gallimard ,nrf.
- Rosenfield (I.), 1989, L'invention de la mémoire. Le cerveau, nouvelles donnes, Eshel.
- Ruyer (R.), 1964. L'animal, l'homme, la fonction symbolique, Paris, Gallimard, coll. « L'avenir de la science ».
- Ruyer (R.), 1966. Paradoxes de la conscience et limites de l'automatisme, Paris, Albin Michel, coll. « Les savants et le monde ».
- Ruyer (R.), 1937, La conscience et le corps, Félix Alcan, « Nouvelle bibliothèque de philosophie ».
- Scheler (M.), 1971. Nature et forme de la sympathie. Contribution à l'étude des lois de la vie affective, Paris, Payot, coll. « Petite bibliothèque Payot ».
- Scheler (M.), 1955. Le formalisme en éthique et l'éthique matériale des valeurs. Essai nouveau pour fonder un personnalisme éthique, Paris, Gallimard, « Bibliothèque de Philosophie ».
- Simondon (G.), 1964, L'individu et sa genèse physico-biologique, PUF, « Epiméthée ».
- Simondon (G.), 2013 (1958, thèse de l'auteur correspondant à l'ouvrage cité), L'individuation à la lumière des notions de forme et d'information, Million, « Collection Krisis ».
- Thom (R.),1990, Apologie du logos, Hachette, « Histoire et philosophie des sciences ».
- Varela (F.), Thomson (E.), Rosch (E.), 1993, L'inscription corporelle de l'esprit. Sciences cognitives et expérience humaine, Seuil, « La couleur des idées ».
- Wallon (H.), 1987, (1ère édition 1949), Les origines du caractère chez l'enfant. Les préludes du sentiment de personnalité, PUF, « Quadrige ».

# 13

# Conscience artificielle et monde militaire

**Gérard de Boisboissel**
Ingénieur de recherche
Centre de Recherche
des Ecoles
de Saint-Cyr Coëtquidam (CREC)

## Abstract

Gérard de Boisboissel wonders what consciousness is, whether or not it is possible to have conscious machines, and how useful it would be in the military context. He begins by making a clear distinction between intelligence and conscience: if there are machines to which the qualifier intelligent is attributed without much discussion, none of them can be seen, for the moment, as "conscious". He then examines the different categories of consciousness, using the typology separating consciousness measured by a state of alertness from the awareness of access to such and such information; he adds awareness of the consequences of the choice of action. He takes sides by stating that we won't be able to equip a machine with any of these types of consciousness. However, even if it can be done at all one day, in addition to AI algorithms, would it really be useful or desirable? To reflect on this question G. de Boisboissel analyzes the different facets of (human) consciousness necessary to (and implemented by) a combatant on a battlefield. This analysis allows him to determine which facets could advantageously be integrated - when they are not already there - into robotic systems. He ends by assuming that a program, however advanced it may be, will never have the intuition or the instinct of a human being. And that a machine, military or not, is and will remain (or must remain) a tool in the service and under the control of those who use it.

De nombreux chercheurs sont persuadés qu'un état mental conscient est la résultante de phénomènes biologiques localisés dans le cerveau, et que donc la modélisation de ces phénomènes et leur implémentation au travers d'algorithmes virtuels ou embarqués permettra de créer une conscience artificielle. La question fait débat.

L'objet de cet article est de réfléchir à ce concept de conscience artificielle, et de voir si celui-ci pourrait être appliqué à des systèmes militaires, afin de rendre conscientes des machines autonomes déployées sur le champ de bataille.

## 1. Intelligence n'est pas conscience

En préambule, il convient avant tout de distinguer intelligence de la machine et conscience de soi par une machine. Actuellement, il est déjà notable qu'une machine peut donner une impression d'intelligence dans son comportement ou son raisonnement. C'est dès à présent le cas avec l'Intelligence Artificielle faible qui surpasse déjà l'Homme dans certaines tâches telles que le programme Deep Blue d'IBM qui bat Gary Kasparov aux échecs en 1997, le jeu de Go avec le programme AlphaGo de Google DeepMind qui devint champion du monde en 2016, et simulateur de vol de combat Alpha de l'université de Cincinnati qui bat le colonel américain Gene Lee en simulation de combat aérien. D'autres machines peuvent générer des compositions artistiques qui pour certaines s'avèrent pertinentes comme le tableau numérique « The Next Rembrandt », qui a été dévoilé le 5 avril 2016, à la galerie Looiersgracht60 d'Amsterdam.

Mais excellentes sur l'exécution d'une fonction donnée, ces machines programmées pour exécuter certaines fonctions sont essentiellement esclaves de leurs composants et des contraintes de programmation et ne peuvent faire preuve de créativité hors des bornes qu'on leur a données. Elles apparaissent en tout cas ne pas avoir conscience d'elles-mêmes au sens où nous l'entendons humainement et le sens de leur création ou de leur action leur échappe. Elles ne peuvent comprendre les principes qui les structurent et qui ont été élaborés par l'Homme. Tout au plus pourront-elles donner l'impression de simuler un comportement qui s'apparente à celui que pourrait avoir un humain, en se fondant sur des langages formels qui décrivent une représentation mathématique de la réalité, langages spécifiés par l'homme et par conséquent restreints aux choix des spécifications retenues.

## 2. Les différents types de conscience

Selon Stanislas Dehaene, le mot « conscience » est polysémique, mais la plupart des chercheurs conviennent de distinguer, au minimum, l'état de conscience et le contenu de la conscience.

Selon lui, « le premier usage, intransitif, du mot « conscience » renvoie aux variations graduelles de l'état de vigilance : veille, sommeil, anesthésie, coma, état végétatif... Le

second usage, intransitif, fait référence à la prise de conscience d'une information particulière. On parle alors de « conscience d'accès », ou d'accès d'une information à la conscience »[1].

Traduit dans le monde des machines, un parallèle pourrait être fait entre état de vigilance et un mode actif de la machine permettant de traiter physiquement une information. Tout simplement un mode où les capteurs sont alimentés et actifs et où le processeur traite les informations reçues. C'est le bouton « ON » de la machine activé qui rend la machine vigilante.

Suit, dans le processus décrit par le professeur Dehaene, l'accès de l'information à la conscience, ce qui pour la machine se traduirait par une prise de conscience qu'elle traite de l'information. Pour cette étape, les limites d'une caractérisation anthropomorphique de cet état à une machine sont dès cette étape criantes, car tout au plus une machine sera capable de classer, stocker et mettre en corrélation des informations, mais on ne pourra jamais parler de conscience propre ou d'état mental conscient dans la mesure où un jugement moral fait appel à des interprétations que l'on ne peut mettre en équation, car elles deviendraient alors des règles de droit figées (ce que n'est pas la morale). Selon le professeur Lambert, « l'usage d'un vocabulaire anthropomorphique à propos d'une machine et donc la confusion intentionnelle entre l'être humain et cette dernière pose une question éthique car cette identification n'est pas défendable. Ce n'est pas parce qu'un système imite, simule, représente une ou même toutes les fonctions de l'humain que l'on peut l'identifier purement et simplement à une personne humaine »[2].

Nous rajouterons ici un troisième état de conscience qui concerne la conscience de ses choix ou le jugement de l'action que la machine peut effectuer. Ce troisième état nous ramène, pour une machine militaire, à la possibilité de mesurer les conséquences éthiques ou morales de son action, et de vérifier le respect des règles ou des droits du monde militaire (les règles d'engagement, le Droit International Humanitaire), ce qui se heurte à la très difficile, voire impossible transcription de règles morales sous forme d'algorithmes, ou de possibilité de libre arbitre face à des choix complexes.

[1] « Psychologie cognitive expérimentale », professeur Stanislas Dehaene, membre de l'Institut (Académie des sciences) https://www.college-de-france.fr/media/stanislas-dehaene/UPL62003_Dehaene.pdf

[2] Cfr M. Gabriel, *Pourquoi la pensée humaine est inégalable. La philosophie met l'intelligence artificielle au défi* (trad. G. Sturm, S.M. Sturm), Paris, JC Lattes, 2019; H. Atlan, *Cours de philosophie biologique et cognitiviste. Spinoza et la biologie actuelle*, Paris, Odile Jacob, pp. 538-545.

## 3. En quoi une conscience artificielle serait utile au monde militaire?

Le mythe et rêve de l'homme d'avoir un esclave à son service qui effectue les actions difficiles ou pénibles à sa place est vieux comme le monde. Si le respect de l'autre dans son humanité a de nos jours fort heureusement condamné l'esclavage, il reste que le rêve de remplacer l'humain par une machine dans les tâches hautement dangereuses est toujours actuel, et l'humanité y tendra tant qu'elle se développera technologiquement.

C'est d'autant plus vrai pour le milieu militaire, de par son extrême dangerosité et les risques et abnégations que doivent assumer les hommes et les femmes qui exercent le métier de soldat. Le robot soldat idéalisé devient ainsi un partenaire de combat qui s'expose à votre place et vous remplace pour des missions fatigantes, répétitives ou dangereuses.

En conséquence, le développement de machines de plus en plus autonomes fait miroiter de nouvelles perspectives de remplacement du soldat par des systèmes robotisés assurant une ou plusieurs fonctions en remplacement de l'homme[3].

Néanmoins si l'on souhaite voir des machines militaires devenir intelligentes, ou tout du moins embarquer des techniques d'intelligence artificielle qui donnent l'illusion d'une forme d'intelligence, dans le but que ces premières soient plus réactives et plus précises que l'Homme dans leur exécution, on peut se poser la question de la nécessité et de la possibilité de développer une forme de conscience artificielle.

Dans quel but ? Celui de rendre la machine vertueuse dans le choix de ses actes, et respectueuse des règlements militaires et réglementations internationales, et apte à porter des jugements sur des actes qu'elle aura accomplis ou que les autres auront accomplis autour d'elle. Mais la chose est-elle possible ? Avant d'en débattre, nous partirons tout d'abord de ce que ressent le plus ultime pion tactique sur le champ de bataille : le fantassin.

## 4. Quels types de consciences pour un combattant sur le champ de bataille ?

Le soldat sur le champ de bataille ressent de très nombreuses sensations et ses sens sont généralement amplifiés au sein d'un environnement qui peut brutalement être foncièrement hostile. Il a conscience de sa finitude face aux dangers auxquels il s'expose, face à la responsabilité qu'il a de porter les armes au service de son pays, et conscience des règles qu'il doit respecter qui vont de la loyauté à ses chefs au respect des règles internationales du droit de la guerre.

[3] Voir à cet effet l'ouvrage édité par le CREC Saint-Cyr « Autonomie et létalité en robotique militaire » édité par les Cahiers de la Revue Défense Nationale : https://fr.calameo.com/read/000558115a2727297e70a

Nous présenterons ici les différentes formes de conscience que le soldat expérimente en opération, et tenterons d'imaginer comment cette forme de conscience pourrait être transcrite dans une machine militaire faite de silicium, d'électronique et de logiciels embarqués.

## 4.1 Le soldat a conscience du contexte militaire dans lequel il est engagé

### *4.1.1 Conscience d'être un maillon dans un système plus complexe*

Un soldat n'est jamais seul, mais il est intégré dans une unité, elle-même dépendante d'entités plus larges, et toujours soumis à une autorité supérieure. Il est loyal envers sa hiérarchie, ayant conscience qu'il s'expose à de graves risques et qu'il expose ses camarades s'il n'obéit pas. Cette conscience de l'autre et de la nécessaire entraide au sein du groupe est le fruit d'un entraînement commun, de multiples exercices, et de l'empathie que les frères d'armes ont les uns pour les autres. C'est ce qui forge un esprit de corps.

Une machine ne pourra jamais avoir une telle conscience, ce sentiment d'appartenance et de redevabilité envers les autres. Cette dimension est purement humaine, fruit de nos fragilités et de nos forces, et d'un vécu commun et partagé.

### *4.1.2 Conscience de sa relation aux autres*

Au niveau tactique, toutes les unités sont plus ou moins spécialisées dans une fonction (tireur d'élite, démineur, fantassin), un moyen d'action (commando, reconnaissance) ou dans la gestion d'équipements qui lui sont confiés (artillerie, blindés). Cette organisation implique une interdépendance entre les unités pour une sécurité et une efficacité optimale. Ainsi nous trouvons des unités d'appui (artillerie, génie) et de soutien (transmission, train, matériel). Elles sont toutes organisées au sein d'un dispositif où chacun connait sa place et son rôle.

Pour le soldat, ceci nécessite à la fois a) d'avoir conscience de ce que les autres unités peuvent m'apporter dans mon action, et à la fois b) avoir conscience de ce que je dois mettre en œuvre pour soutenir ou appuyer ou secourir les autres. Les autres en effet comptent sur moi, tout comme moi je compte sur les autres, et cette relation de dépendance passe par la confiance.

Concrètement, le premier réflexe du soldat dans le feu de l'action, hormis se protéger soi-même, est de penser au groupe dont il fait partie pour à la fois s'assurer de la protection de ses camarades, tout en faisant effort sur l'ennemi. Un fantassin ne progresse jamais à l'aveugle : il regarde autour de lui en permanence ses camarades et adapte sa progression en fonction de leur vitesse, des risques perçus. De la même façon, il a dans sa progression conscience du soutien de ses camarades ce qui le met en confiance.

Cette conscience de l'autre accompagne toute action militaire, et conditionne le choix du combattant : va-t-il continuer son assaut ou s'arrêter quelques instants pour protéger la

progression de son camarade ? Va-t-il s'exposer plus encore pour dépasser un obstacle, reprendre l'initiative ? C'est la prise de risque pour lui et pour les autres qui conditionne sa décision, le tout en suivant le cadre d'ordre qu'il a reçu. C'est également la possibilité de se sacrifier le cas échéant pour sauver une situation ou un camarade.

L'interconnexion entre systèmes permet un échange d'information, mais ne permet pas de donner une valeur de confiance à la donnée, ni à personnaliser un équipement avec lequel j'interagis. Cette valeur de confiance et la nécessaire mise en contexte de la donnée sont le fruit d'une perception globale qu'une machine ne saura jamais reproduire car elle n'aura jamais cette confiance que les individus peuvent avoir dans leur camarade, et ne sera jamais en mesure de mesurer la force morale qui fait la grandeur des armées.

Néanmoins, la machine ayant une force calculatoire supérieure à celle de l'Homme, va pouvoir anticiper les besoins nécessaires à la situation tactique demandée en fonction des éléments qu'elle a collectés, et proposer un redéploiement des moyens ou ressources dont elle dispose au profit de telle unité ou de telle autre.

*4.1.3 Conscience de ses limites et des limites du groupe auquel il appartient*

Le militaire effectue son action en ayant conscience des limites de cette action. Il prend en compte les armes qu'il a à sa disposition, leur portée, les effets de ces dernières sur l'objectif (le percement du blindage) etc. Sur ce point il est possible qu'une machine ait la possibilité de traduire les caractéristiques de ses équipements (capteurs, processeurs, effecteurs etc.) en limites structurelles qui pourront être prises en compte dans toute décision algorithmique.

Le militaire mesure aussi les risques auxquels il s'expose, en dévoilant ou non son dispositif durant la manœuvre. Le chef, de son côté, prend en compte les potentiels et contraintes de son unité de combat, voyant si avec les éléments qu'il a il peut conduire une manœuvre victorieuse dans le temps, ce qui n'est pas sans rappeler l'évangile selon Saint Luc (14, 31-33).

La machine peut là encore effectuer des calculs probabilistes sur les chances de réussite d'une manœuvre selon l'option choisie. Néanmoins, ce qui fait la force de l'Homme, c'est sa capacité à rebondir face à l'adversité en inventant un mode opératoire innovant. Une telle créativité ne sera pas possible pour une machine qui restera dans les bornes qui lui auront été fixées, limitant ici toute capacité d'innovation hors du cadre prédéfini.

## 4.2 La conscience dans l'action

Sur le terrain prédomine la conscience de situation au cœur de l'action militaire.

*4.2.1 Perception de ce qu'il ressent*

Au combat, l'individu a tous ses sens en éveil. Notamment ses sens physiologiques, sa fréquence cardiaque, sa respiration vont réagir et s'adapter pour faire face au danger[4]. Cette perception de lui-même lui permet d'être conscient de sa capacité à réfléchir et de ses capacités à décider, ou pas.

De nombreux facteurs peuvent influer sur sa conscience relationnelle avec le monde extérieur, c'est-à-dire sur sa capacité de concentration, de réflexion, d'anticipation, ses réflexes… : il s'agit de la fatigue, le stress, la peur, le froid, la faim. Prenons pour exemple la fatigue : un soldat fatigué va lutter contre le sommeil, et s'organiser pour faire en sorte qu'il soit toujours opérationnel au moment de l'action militaire, quitte à s'organiser avec ses camarades (tours de garde et de repos).

Toutes ces agressions de l'environnement extérieur, favorisent une réaction physiologique sur l'individu, d'autant plus intense que la guerre est un acte violent. Notez qu'un des objectifs du projet de recherche sur le soldat augmenté du CREC Saint-Cyr est de tirer parti du développement des nouvelles technologies pour trouver des substituts ou des manières de préserver ou de maintenir la conscience psychologique du soldat face à des situations le mettant à l'épreuve (psychique, psychologique, physique)[5].

Mais pour ce qui est de la machine, celle-ci ne ressent rien de ces indicateurs physiologiques, car elle n'a pas de physiologie vivante. En conséquence, elle peut juste analyser des informations transmises par des capteurs, et les transformer en comportement imitant la perception d'une indication physique. La faim devient ainsi un niveau en énergie inférieur à un seuil, le froid ou le chaud deviennent une mesure de température couplée avec les tolérances limites des composants, la fatigue un seuil de prédiction de panne en fonction des heures d'utilisation.

*4.2.2 Perception : maîtrise de l'environnement, de la position dans l'espace, de l'appropriation de l'espace*

Un terrain se sent. Le fort de l'entraînement militaire est de faire apprendre aux combattants les caractéristiques des objets et éléments qui l'entourent (route, coupures humides ou végétales, relief), des temps nécessaires pour y évoluer (progression des véhicules, des hommes), et en déduire des positions favorables ou des zones à risques.

L'expérience est primordiale pour percevoir un espace. Néanmoins, il est très raisonnable de penser que les machines du futur auront la possibilité d'étudier de très nombreux paramètres caractérisant un espace, et d'en déduire des choix optimaux selon tel ou tel

[4] Lieutenant-colonel Michel Goya, Sous le feu, cahier de réflexion doctrinale, p17.

[5] Cahier de la Revue Défense Nationale, « Le soldat augmenté : les besoins et les perspectives de l'augmentation des capacités du combattant »:
https://fr.calameo.com/read/0005581159f5e895e1a2c

besoin : vitesse de progression, optimisation énergétique des mouvements, risque IED, position des appuis etc. Il en est ainsi de l'Intelligence Artificielle qui par son apprentissage des données cartographiques et la consultation de bases de données existantes pourra analyser un terrain peut-être mieux qu'un individu ne le peut. Il reste néanmoins que le fait de sentir le terrain fait appel à une conscience externe de l'environnement que seul l'Homme est en mesure de faire.

*4.2.3 Perception de la menace*

La perception de la menace est le fruit d'un apprentissage extrêmement long qui dure depuis les premiers âges de notre vie, fruit d'une expérience personnelle et de notre éducation. La perception de la menace nait de la méfiance. Or la méfiance est impossible à spécifier pour un langage informatique. Un homme en colère contre lui-même aura un visage fermé, mais n'est pas forcément menaçant. Un homme souriant peut tout à fait simuler une attitude bienveillante, alors qu'il a des intentions hostiles. Pour le monde militaire, un ennemi peut être en mesure de déclencher le feu, mais dans une attitude d'attente d'un ordre de son supérieur. Et un enfant peut jouer avec une arme factice et vous mettre en joue.

Comment une machine pourra-t-elle percevoir ces subtilités, alors qu'elle n'a aucune « éducation à la sociologie humaine » ? Et quand bien même un processus d'apprentissage par l'IA lui ferait apprendre certains comportements, chaque situation est unique et c'est pour cette raison que l'on forme les militaires et notamment leurs chefs au discernement. Les exemples abondent dans la littérature militaire de cette retenue au combat avant le déclenchement du feu, notamment au sein de populations civiles agressives mais non pas animées d'une volonté de destruction. C'est ce qui constitue une partie de l'enseignement aux écoles de Saint-Cyr Coëtquidan dont le but est de former les officiers de demain aux qualités leur permettant de :

- discerner dans la complexité (déployer une véritable intelligence de situation) ;
- décider dans l'incertitude (avoir une véritable force de caractère permettant d'accepter des risques calculés) ;
- agir dans l'adversité (pour fédérer les énergies, susciter l'action collective et décider en conscience).

### 4.3 Le soldat a conscience de ses actes

La conscience au sens moral désigne la « capacité mentale à porter des jugements de valeur moraux […] sur des actes accomplis par soi ou par autrui »[6]. Or l'homme est un être doué de profondeur et capable de transcendance, lui donnant cette conscience morale. La machine n'est qu'un exécutant de lignes de code sans âme et sans conscience.

[6] https://fr.wikipedia.org/wiki/Conscience

*4.3.1 L'éthique comme règle de la conscience*

Avoir conscience de mes actes implique une analyse morale de la conséquence de mes actes. Ce qui reviendrait à pouvoir mesurer les effets d'une action et pouvoir l'analyser en amont de la décision, c'est-à-dire de développer une éthique algorithmique embarquée. Mais selon le professeur Dominique Lambert[7], « s'il est pensable de programmer un système de telle manière qu'il contrôle la satisfaction de certaines règles, l'évaluation éthique ou morale repose d'abord sur trois éléments majeurs : une appréciation de l'objet d'un acte ou d'une action, une prise en compte des éléments du contexte et enfin une appréhension de l'intention sous-jacente. Il faut un travail d'interprétation qui est, pour une bonne part, intuitif et créatif sans être arbitraire, et la complexité et le caractère inédit des situations imposent parfois que l'on puisse sortir (sans « règles de sortie de règles » !) de tout ensemble existant de règles « classiques », pour pouvoir sauver et faire fonctionner l'esprit général des règles et celui des lois ». De même les éthiques militaires peuvent différer selon les pays et ne pas être « algorithmisables » de la même manière.

*4.3.2 Pour le chef, la conscience de la portée des actes de ses soldats*

La conscience du chef et de sa mission de commandement, de sa responsabilité envers les soldats, implique de savoir si l'ordre qu'il donne à ses soldats est juste ou injuste, moral ou immoral.

Si nous appliquons cela à des machines numériques, il nous semble qu'avoir conscience de la portée des actes des autres revient à centraliser une analyse et une prise de décision à un niveau déporté ailleurs que sur la machine même, c'est-à-dire sur un serveur déporté collectant les données collectives de tous les combattants et équipements du champ de bataille, en vue d'une synthèse. Ce qui nécessite une interconnexion totale entre tous et ce qui aurait pour effet d'abandonner l'initiative à un niveau calculatoire supérieur.

### *4.3.3 La non transgression des règles*

Le quatrième article du code du soldat stipule qu'il doit obéir aux ordres dans le respect des lois, des coutumes de la guerre et des conventions internationales. Le soldat doit ainsi faire preuve de discernement lorsqu'il reçoit un ordre contraire aux lois, aux coutumes de guerres et aux conventions internationales et, en se référant à sa conscience, refuser l'application de cet ordre d'un point de vue moral. Sur ce point, sous réserve qu'il soit possible d'algorithmiser les règles et lois humanitaires internationales, la machine sera potentiellement plus performante que l'être humain, car non soumise comme ce dernier aux comportements irréfléchis que peuvent être le stress, la colère, la peur et l'envie de vengeance.

[7] « Autonomie et létalité en robotique militaire » Cahiers de la Revue Défense Nationale, p.231-232.

*4.3.4 La conscience de son autonomie décisionnelle*

Ce qui fait la liberté de l'Homme, c'est sa capacité de décider en son âme et conscience. Dans le monde militaire, cela se traduit par une relation de subsidiarité entre le chef et son subordonné.

Les machines n'ont pas et n'auront jamais de volonté propre, elles ne sont que des algorithmes réagissant à des stimuli (capteurs notamment) et s'adaptant à la situation par l'écoute de leurs capteurs et le traitement de l'information qui en suivra. On ne pourra donc pas parler de volonté des algorithmes, mais éventuellement d'une volonté humaine « algorithmisée » dans une machine.

En outre, il n'est d'aucun intérêt pour un chef militaire d'avoir un système robotique qui se gouverne avec ses propres règles et ses propres objectifs, ni qui puisse faire preuve de désobéissance ou s'affranchir du cadre qu'on lui a fixé. Tout système autonome doit respecter les ordres et les consignes militaires, car c'est le chef qui les donne et qui donne du sens à l'action militaire, tout en étant le responsable[8].

En conséquence il n'est pas souhaitable d'avoir des machines avec une intelligence artificielle simulant une prise de décision autonome en dehors du cadre normatif militaire, à savoir une obéissance totale aux ordres et consignes que lui ont donnés les chefs militaires qui l'ont en charge.

Il reste que le progrès est dû à l'intelligence de l'Homme lorsqu'elle s'exprime pour dépasser ses propres limites. Or, par construction, et pour garder confiance dans la machine, cette dernière ne sera pas autorisée à dépasser ses limites dans le monde militaire, sous peine de dérives incontrôlables et inacceptables.

## 5. Les cas particulier de l'intuition et de l'instinct

Certaines personnes sont aptes à voir plus loin que le simple visible, à pressentir des menaces, des évènements à venir ou bien les conséquences de certaines actions, sans être dans la possibilité de l'expliquer rationnellement. C'est-à-dire avoir une intuition de ce qui peut se réaliser dans le futur. Ce n'est parfois même qu'après coup qu'elles prennent conscience du fait.

[8] Guillaume Venard et Gérard de Boisboissel : La nécessaire place du chef militaire dans les systèmes d'armes robotisés autonomes, Cahier de la RDN « autonomie et létalité en robotique militaire », p. 118.

Ainsi, le pressentiment – ou intuition – et l'instinct échappent souvent à toute logique rationnelle et peuvent conditionner des réactions qui sauveront la vie sur le champ de bataille. Leur codage dans des machines semble hautement improbable.

Le témoignage du capitaine Clément H, officier de cavalerie, lieutenant chef de peloton au moment des faits, nous en donne un exemple saisissant :

« Nous sommes en 2011, dans un pays africain francophone. L'unité à laquelle j'appartiens conduit un raid blindé dans une grande ville.

Après avoir été au contact pendant plus d'une heure contre un ennemi mobile et entreprenant, la situation semble brusquement se calmer dans ma zone d'action. Je reçois donc l'ordre de mon capitaine de reprendre la progression vers l'objectif du bataillon.

Toutefois, j'ai le sentiment qu'une menace est toujours présente à proximité de ma position. Je décide donc d'envoyer mon chef d'escouade mener « un coup de sonde » sur le carrefour situé à une cinquantaine de mètres devant moi. A l'issue seulement, j'envisagerai de m'y engager avec mes chars. Mon chef d'escouade, le maréchal des logis Boris D., s'y déplace et me rend compte que la situation est « claire ». Je lui demande néanmoins d'y retourner, car j'ai toujours l'intuition que quelque chose ne va pas. Il exécute l'ordre et y retourne. Au même moment, mon commandant d'unité, pressé, me demande d'accélérer le mouvement, ne comprenant pas pourquoi je suis toujours à l'arrêt. Je temporise autant que possible.

Puis, alors que j'observe mon chef d'escouade depuis la tourelle de mon engin blindé, je remarque que son véhicule blindé léger se met subitement à reculer rapidement. Et ce dernier me rend compte d'une voix forte à la radio : « BMP2 posté secteur gauche du carrefour ! ». Le capitaine me donne immédiatement l'ordre de détruire le blindé ennemi : « ALPHA 10, détruisez ! ». Après trois à quatre minutes interminables, percé d'un obus flèche tiré par mon opérateur tourelle, le blindé ennemi explose.

Quand je repense à cet événement et au vu de la situation, objectivement, on aurait pu y aller ; « mais je ne le sentais pas ». Cela nous a certainement sauvé la vie. »

## 6. Conclusion

On le voit, ce n'est pas une conscience unique à laquelle le soldat fait appel sur le champ de bataille, mais à plusieurs types de consciences. Pour certaines d'entre elles, il semble envisageable pour une machine d'avoir des comportements qui puissent sembler en concordance avec une analyse consciente, comme par exemple la perception active de son environnement et de ses dangers (à un horizon de 5 ans), l'aide à la décision pour le commandement (à horizon de 10 ans avec l'Intelligence Artificielle) et la non violation des règles et des lois internationales quel que soit le contexte (à plus long terme).

Il reste qu'une machine numérique n'aura jamais connaissance de sa propre physiologie, ne connaitra jamais ce que peut être une relation psychique avec des équipiers, préambule à toute notion de confiance dans les autres. Elle « ne pourra jamais comprendre les subtilités du cerveau humain, capable de simuler, de bluffer, de mentir, et surtout d'aimer. A fortiori le cerveau d'un adversaire distant sur le champ de bataille. Tout au plus pourra-t-elle simuler. Enfin l'audace restera toujours le propre de l'homme car elle intègre la notion de sacrifice nécessaire pour la mission, ou pour sauver un camarade»[9].

Le docteur Patrick Theillier indique que « depuis des décennies, des recherches scientifiques ont été conduites pour localiser la conscience et la mémoire à l'intérieur du cerveau. Or la Science n'a à l'heure actuelle aucune idée sur la façon dont les cellules cérébrales pourraient engendrer des pensées… et certains suggèrent que la conscience est séparée du corps »[10]. Comment une machine militaire pourrait-elle avoir une telle transcendance et écouter sa conscience ? De fait, générant des données en sortie après traitement algorithmique de données en entrée, elle est par construction finie. Elle restera donc un outil au service du chef militaire et sous son contrôle, afin de lui laisser l'initiative et le contrôle de la manœuvre, actions que ce dernier effectuera de lui-même, mais en toute conscience.

[9] idem.
[10] Patrick Theillier, Expériences de mort imminente, p. 96.

# 14

# Relation émotionnelle entre humains et robots : Quelle éthique ?

**Laurence Devillers**

Sorbonne Université, GEMASS/LIMSI/CNRS,
membre de la CERNA
Commission de réflexion
sur l'Ethique de la Recherche Nationale
en sciences et technologies du Numérique d'Allistène

## Abstract

Laurence Devillers tackles what is called "emotional robotics", which is part of the relationship between humans and robots. She begins by recalling the advances, brought by psychology and neuroscience, on the understanding of emotions, their role, their relationship with consciousness; she also returns to this latter concept, its different definitions and components. She then recalls the course of AI from the point of view of her subject, namely the integration of machines in society, the more and more frequent use of these machines by individuals, the work concerning the detection of individual emotions by machines and their simulation in return, within the framework of these interactions. She finally wonders about the benefits and risks of all these advances. Providing robots with emotional simulation or interpretation skills is particularly achieved in the health field. They can then provide invaluable assistance: playing with sick children, monitoring hospitalized people, intervention in the event of depression or autism, diagnostic assistance, etc. But on the other hand, "robots also carry significant risks of isolation, dehumanization and manipulation of humans". Human-machine co-adaptation is spreading quickly but requires reflection as the possibility of being manipulated by these devices is great. Laurence Devillers therefore encourages vigilance and assures that this co-adaptation "should in the long term be an important research and monitoring axis in the coming years"

La robotique émotionnelle est un champ de recherche et d'application émergent de l'intelligence artificielle et de la robotique. Ce terme tout comme celui d'intelligence artificielle (IA) est un oxymore, car il mélange des notions opposées relatives au vivant et à l'artefact. L'IA est définie depuis 2018[1] comme étant le champ interdisciplinaire théorique et pratique qui a pour objet la compréhension de mécanismes de la cognition et de la réflexion, et leur imitation par un dispositif matériel et logiciel, à des fins d'assistance ou de substitution à des activités humaines. Aujourd'hui, l'informatique émotionnelle ou *affective computing* (Picard, 1997) est devenue centrale dans le développement des systèmes d'interaction humain-machine, notamment pour la robotique de services. Elle regroupe trois technologies : la reconnaissance des émotions des humains (Devillers, 2005), le raisonnement et la prise de décision en utilisant ces informations, la génération d'expressions affectives par les machines.

Doter les robots de capacités d'interprétation, de raisonnement et de simulation émotionnels est utile pour construire des systèmes interagissant socialement avec les humains. Les robots émotionnels peuvent être d'un apport important pour la santé ; pour prêter assistance, stimuler cognitivement, voir surveiller pour des pathologies liées au grand âge, à la dépression ou encore à l'autisme ou encore pour l'éducation. Les robots sont aussi porteurs de risques importants d'isolement, de déshumanisation et de manipulation des humains. Comment allons-nous évoluer face à ces machines ? Est-ce qu'un robot sera capable de déceler des comportements émotionnels dont nous ne serions pas conscients pour mieux nous inciter à faire certaines actions ou à prendre certaines décisions ? La co-adaptation humain-robot sur le long terme devra être un axe de recherche et de surveillance important dans les prochaines années (L. Devillers, 2017).

## 1. Les émotions du vivant et des artefacts

Le domaine de l'affectivité a souvent été opposé à la cognition désignant les capacités de raisonnement rationnel. Mais cette dichotomie a été remise en cause car les processus affectifs contribuent de façon importante à l'adaptation de l'individu à son milieu et font partie intégrante de sa cognition. Considérées comme preuves de faiblesses, méprisées, parfois diabolisées, le plus souvent incomprises, les émotions ont fait l'objet au sein de la tradition philosophique et en psychologie de vives discussions. Les émotions ont également longtemps été considérées comme trop subjectives pour se prêter à une approche expérimentale en laboratoire que ce soit en neuroscience ou en modélisation informatique. La fin du 20e siècle a vu émerger des nouveaux champs scientifiques au sein des sciences cognitives (Andler, 1992), les neurosciences affectives (Panksepp, 1998) et l'informatique émotionnelle (Picard, 1997).

[1] JORF n°0285 du 9 décembre 2018, texte n° 58, Vocabulaire de l'intelligence artificielle (liste de termes, expressions et définitions adoptés - NOR: CTNR1832601K)

Les sciences affectives, approche interdisciplinaire, qui regroupent neurosciences, psychologie et informatique a pour objectif de comprendre à la fois les mécanismes sous-jacents à l'affect mais aussi comment l'affect et les émotions contribuent au comportement et à la pensée (Davidson, Scherer & Goldsmith, 2003). Plus récemment, Daniel Kahneman, prix Nobel d'économie en 2012, a montré via de multiples expériences à travers sa théorie du système 1/ système 2 comment nos émotions et nos sentiments guident nos décisions (Kahneman, 2011). « *La psychologie hédoniste est l'étude de ce qui rend l'expérience de la vie plaisante ou déplaisante. Elle a à voir avec les sentiments, la douleur et le plaisir, la passion et l'ennui, la joie et le chagrin, la satisfaction ou l'insatisfaction. Elle est corrélée avec toute une gamme de circonstances biologiques ou sociétales qui provoquent la souffrance ou le plaisir* » (Kahneman, Diener & Schwarz, 2003).

Chacun sait ce qu'est une émotion, jusqu'à ce qu'on lui demande d'en donner une définition. A ce moment là, il semble que plus personne ne sache (Fehr & Russell, 1984). Paul et Anne Kleinginna, psychologues des émotions, présentaient plus de 140 définitions dans un article en 1981 : « *Les émotions sont le résultat de l'interaction de facteurs subjectifs et objectifs, réalisés par des systèmes neuronaux ou endocriniens, qui peuvent : a) induire des expériences telles que des sentiments d'éveil, de plaisir ou de déplaisir ; b) générer des processus cognitifs tels que des réorientations pertinentes sur le plan perceptif, des évaluations, des étiquetages ; c) activer des ajustements physiologiques globaux ; d) induire des comportements qui sont, le plus souvent, expressifs, dirigés vers un but et adaptatifs* ». Il existe donc différentes définitions mettant l'accent par exemple sur la dimension subjective, les catégories de stimuli déclencheurs, les mécanismes physiologiques, l'expression des comportements émotionnels, les effets adaptatifs ou encore les effets perturbateurs. Le concept d'émotion utilisé pour les modélisations sur machine recouvre souvent plusieurs phénomènes que le sens commun a regroupés et est utilisé au sens large d'état affectif ou affect. En psychologie, le terme affect englobe l'ensemble des comportements affectifs : émotions, humeurs, dispositions affectives, et sentiments. Les émotions ont souvent des épisodes visibles sur notre visage, présents dans la prosodie de notre voix ou le langage de notre corps, les sentiments eux sont cachés. Ils sont nos ressentis émotionnels internes.

La sphère des émotions que l'on pensait propre à l'humain s'invite dans les machines. Les émotions agissent sur nos comportements quotidiens, sur nos perceptions et nos décisions. Elles rendent la communication plus efficace, jouent un rôle clé dans tout processus d'apprentissage, agissent également sur la capacité de mémorisation de l'information et sur notre attention. Les émotions et la cognition sont intimement liés. Les émotions humaines sont une sorte de double interface entre le cerveau et le corps. D'un côté, elles donnent des signaux perceptibles dans le corps qu'un besoin est satisfait ou non. D'un autre côté, elles font partie des qualia, phénomènes psychiques et donc subjectifs, constitutifs des états mentaux. Que savons-nous sur les qualia ? Les qualia sont les propriétés de la perception et généralement de l'expérience sensible : sensations corporelles, affects, percepts mentaux. Ce sont les qualités ressenties de nos expériences conscientes. C'est ce qu'on expérimente

lorsqu'on perçoit ou ressent quelque chose : qu'est-ce que cela fait de caresser un chat, de le sentir se frotter contre vous lorsqu'il a faim, de l'entendre ronronner ou miauler, de savoir qu'il peut griffer ? Les qualia constituent ainsi l'essence même de l'expérience de la vie et du monde, c'est notre relation aux autres. Quand le logiciel développé par Google apprend à différencier des chats, il n'apprend qu'en regardant des images ce qu'est un chat. Il est loin de notre compréhension de ce qu'est un chat.

La vie mentale, la conscience et l'inconscient sont les sujets privilégiés des neurosciences qui connaissent depuis quelques décennies un développement impressionnant, tant du point de vue de l'expérimentation scientifique grâce à l'IRM que de la connaissance de la structure et des fonctions du cerveau. Les nombreux ouvrages des neuroscientifiques, Antonio Damasio, Lionel Naccache, Stanislas Dehaene, Jean-Pierre Changeux, Pierre-Marie Lledo en témoignent. La conscience est un phénomène biologique localisé dans le cerveau. Dans *La construction du cerveau conscient* (Damasio, 2010), Damasio développe l'idée que la conscience ne serait pas le produit sophistiqué des régions les plus récentes et les plus évoluées de notre cerveau, mais des plus anciennes, là où naissent les émotions. Le début de la conscience serait le ressenti d'un état de l'organisme. Jean-Pierre Changeux et Stanislas Dehaene parlent d'un espace global conscient. Cet espace de simulation d'actions virtuelles servirait à élaborer nos buts, intentions, et programmes d'action en interaction avec le monde extérieur en considérant nos dispositions innées, nos désirs et émotions, notre personnalité, notre mémoire et les normes morales et conventions sociales. Les réseaux de neurones du cerveau humain sont capables d'activités spontanées et susceptibles d'engendrer des représentations du monde sans qu'il y ait nécessairement une stimulation extérieure. On sait également aujourd'hui que notre inconscient est riche, qu'il contient de nombreux processus et représentations mentales abstraites, qui coexistent avec nos pensées et nos représentations conscientes. Nos expériences vécues ont un support neuronal et sont liées à des processus chimico-biologiques.

Actuellement, les nouvelles découvertes sur le cerveau et sur la manière d'optimiser ses capacités sont importantes, les avancées plus incroyables les unes que les autres, mais nous n'en savons pas beaucoup plus sur l'organisation de l'ensemble du cerveau. Le projet « *Human Brain Machine* »[2] de l'Ecole Polytechnique Fédérale de Lausanne (EPFL), qui avait pour but de recopier le cerveau d'un mammifère sur une machine, n'a pas permis beaucoup de progrès pour comprendre le fonctionnement d'un cerveau.

Si l'on savait programmer un système de réseaux de neurones semblable au cerveau humain sur un support informatique, le comportement résultant généré par l'ordinateur ne s'apparenterait pas à celui effectué par l'homme car la machine n'a pas de corps, de ressenti, n'a pas de désir et de plaisir. Même si une machine pouvait simuler une sorte de qualia, via un « corps artificiel », la modélisation serait très différente d'une expérience humaine. Dire que les machines auront des émotions et une conscience, même si ce ne sont que des

[2] Human Brain Machine, EPFL

métaphores, peut très mal être compris par quelqu'un qui est non-expert. Il est donc urgent de ne pas ajouter de la confusion sur ce sujet. Notre cerveau se transforme perpétuellement grâce à des développements personnels, grâce à l'éducation et à la culture, c'est-à-dire à nos relations avec les autres. Notre capacité d'apprentissage et de compréhension ne vient pas seulement des déductions d'une énorme base de données. Nous apprenons grâce à l'interaction et à l'expérimentation et grâce à notre histoire et notre imagination.

Il est important de croiser les différentes disciplines que sont les neurosciences, la philosophie et l'intelligence artificielle pour essayer de comprendre mieux l'essence de l'humain et ce que peut nous apporter la modélisation sur machine.

Le philosophe Paul Ricoeur définit la conscience comme « un espace de délibération pour les expériences de pensée où le jugement moral s'exerce sur le mode hypothétique ». Ce qui fait naître et relève de notre volonté, nos valeurs, notre imagination, ou encore notre « conatus » ou appétit de vie décrit par Spinoza sont encore des questions en suspens.

Il faut se rendre à l'évidence de la multiplicité des modes intentionnels qui gouvernent notre relation au monde et par lesquels notre subjectivité opère : pensée, volonté, perception, imagination, affectivité, etc. Nous n'avons pas non plus conscience de la grande majorité du travail de notre cerveau. Le mot « conscience » est polysémique. Selon Stanislas Dehaene, le premier usage du mot renvoie aux variations graduelles de l'état de vigilance : veille, sommeil, anesthésie, coma, état végétatif. Le second usage fait référence à la prise de conscience d'une information particulière. Enfin, le troisième niveau est la méta-cognition, ce qui pour la machine se traduit par une prise de conscience qu'elle traite de l'information. Pour cette étape, les limites d'une caractérisation anthropomorphique à une machine sont évidentes, car tout au plus une machine sera capable de classer, stocker et mettre en corrélation des informations. On ne pourra jamais parler de conscience propre ou d'état mental conscient dans la mesure où un jugement moral fait appel à des interprétations que l'on ne peut mettre en équation.

## 2. Des automates aux robots émotionnels

Le développement de l'intelligence artificielle et des robots est avant tout un sujet économique de la société. Depuis le début de l'ère de l'automatisation, nous avons construit des machines qui peuvent nous remplacer pour des tâches difficiles et répétitives le plus souvent dans des usines, pour plus de rentabilité. Les robots ont remplacé les humains par exemple dans les chaines de montage des usines automobiles ou dans les salles des réacteurs nucléaires. Le robot peut nous libérer des tâches dangereuses, ennuyeuses, sales et stupides, connues sous l'acronyme des 4D pour « *Dangerous, Dull, Dirty and Dumb* »[3].

[3] George A. Bekey, *Robotics: State of the Art and Future Challenges*, Presse Imperial College, 2008.

Il s'agit souvent de robots automates plutôt que de robots capables de perception car ils font toujours la même succession de mouvements et ne sont pas autonomes.

La deuxième phase a été celle du *big data* que nous vivons depuis quelques années, il s'agit de remplacer des tâches de bureaucratie, de finance par des systèmes automatiques (les bots) entrainés sur des grands corpus de données pour compter, passer des ordres, traduire ou encore faire des synthèses. Ils n'ont pas besoin d'être incarnés et sont souvent invisibles et confinés à la sphère du travail.

Nous franchissons une nouvelle étape en ce moment qui est celle des assistants sociaux vocaux qui étaient déjà présents dans les centres d'appels et dans les banques, sur nos téléphones et envahissent maintenant nos sphères privées, sans doute bientôt avec des logiciels capables de détecter et simuler quelques émotions. Les machines vont nous parler de plus en plus, cela a commencé avec Siri sur nos téléphones. Aux USA, l'engouement pour ces machines est impressionnant, on a dénombré jusqu'à six enceintes Alexa ou Google home dans certains foyers, un par pièce. Il s'agit d'interagir vocalement avec les humains dans la vie de tous les jours. Le marché est énorme, ces machines pourraient nous accompagner au quotidien, pour surveiller notre santé, pour nous éduquer, nous aider et nous amuser, bref pour s'occuper de nous. Les chatbots et les robots sociaux sont régis par la règle des 4E « *Everyday, E-Heath, Education, Entertainment* ». Pour ces tâches, le chatbot ou robot est vu comme un compagnon ou encore un assistant mais il n'est pas encore vu comme un remplaçant humain. La cobotique, le travail commun entre robot et humain est en train également d'émerger et nécessite que le robot interprète les actions des humains et pourquoi pas leurs émotions.

En 1997, lorsque l'ordinateur Deep Blue devient le premier ordinateur à battre le champion du monde Garry Kasparov, on en a conclu que Deep Blue était doté d'une capacité de calcul peu ordinaire mais que ce n'était pas de l'intelligence. En 2016, lorsque l'intelligence artificielle de Google DeepMind a battu Lee Sedol, champion de GO, 4-1, on a crié au génie ! D'origine chinoise, le jeu de Go est probablement le plus vieux sport cérébral au monde. On a longtemps cru que l'incarnation même de l'intelligence était le jeu de go qui est d'une élégance rare : les mouvements sont très simples, mais le jeu est d'une incroyable complexité. *« Il y a 10 puissance 170 positions possibles, soit davantage que le nombre d'atomes dans l'univers »,* explique Demis Hassabis, PDG de DeepMind. Défi que la machine sait très bien mener avec une capacité de calculs et une dépense d'énergie hors du commun. Cependant il ne faut pas avoir peur d'une machine plus intelligente que l'homme parce que celui-ci est battu aux échecs ou au jeu de GO. Beaucoup de personnes non expertes en IA en ont déduit un peu rapidement que cette avancée majeure allait permettre des progrès fulgurants en robotique sociale. Les lois de Moore (ou conjectures de Moore), 1965, sont des lois empiriques qui ont trait à l'évolution de la puissance des ordinateurs et de la complexité du matériel informatique. Tous les 18 mois, la puissance des machines est multipliée par deux. Cette conjecture n'a jamais été démentie pour l'instant mais elle pourrait s'infléchir, la machine quantique prendrait alors le relai.

L'augmentation constante des capacités de calcul des ordinateurs permet d'augmenter les performances de l'intelligence artificielle mais est-ce que l'intelligence humaine peut être réduite à du calcul ? Nous pouvons simplement dire que ces capacités accrues de calcul amènent des améliorations des performances de l'intelligence artificielle spécifique. Rien ne prouve que plus de puissance de calcul amène à une intelligence dite « forte », c'est-à-dire générale. Nous venons de vivre deux hivers de l'intelligence artificielle, cela peut se reproduire. L'intelligence des robots n'a rien à voir avec celles des humains. Le robot peut reconnaître une pomme mais ne saura jamais le goût qu'elle a. Le programme d'intelligence artificielle AlphaGo qui a battu Lee Sedol au jeu de Go ne sait pas ce que c'est que gagner et n'en n'éprouve aucun plaisir. Il ne peut pas comprendre d'ailleurs le concept de bonheur pour un humain. Ces concepts sont également difficiles à formaliser pour les humains donc nécessairement difficiles voire impossible à coder sur une machine. La reconnaissance des messages sociaux véhiculés par les visages et les voix, en particulier les expressions émotionnelles, est aussi un élément indispensable à la communication avec les humains et à l'insertion dans toutes les sociétés.

La majorité des travaux ayant rapport aux émotions sur machine, notamment la détection des émotions, nécessite de collecter des données (audio, vidéo) et de les annoter (ou étiqueter). Ces annotations permettent de construire des modèles statistiques pour détecter les émotions ou les imiter, ou encore pour raisonner dessus. Elles sont utilisées par des approches d'apprentissage machine pour construire les modèles des émotions qui peuvent être utilisés dans des systèmes de décision. Trois phases sont donc nécessaires : le codage, l'étiquetage et l'apprentissage. L'approche la plus répandue est d'associer des étiquettes verbales ou des dimensions continues à des segments temporels de signaux audio ou vidéo (Devillers, 2010). Des techniques d'apprentissage automatique permettent ensuite de modéliser les émotions.

Une des difficultés de ces annotations est liée à la pertinence de la verbalisation émotionnelle. En effet, nous savons que le ressenti émotionnel est interne, mais jusqu'à quel point peut-on déterminer sa part consciente ? Klaus Scherer, psychologue suisse, répond à cette question à l'aide d'un diagramme de Venn (Figure 1) représentant trois zones qui décrivent schématiquement la répartition de l'activité émotionnelle.

La zone (*A*) représente la part des processus totalement inconscients. Scherer affirme que la plupart des processus entrent dans cette catégorie. Vient ensuite la région (*B*) qui identifie la part des représentations conscientes qui ne peuvent être verbalisées. Enfin, la zone (*C*) détermine la part des processus potentiellement verbalisables. Ce mélange d'évènements conscient et inconscient est une des difficultés de la problématique du traitement automatique. Nous étiquetons les données à partir de la perception que nous en avons qui est très parcellaire. Ce que nous appelons les qualia qui sont les ressentis émotionnels, sont dans A, B et C. L'étude des émotions en neurosciences affectives devrait nous amener à mieux nous connaître.

Pour améliorer les performances d'interprétation des émotions, des informations multimodales sont le plus souvent prises en compte. Les informations multimodales sont linguistiques, i.e. les mots que les personnes prononcent, paralinguistiques, i.e. les intonations, le rythme, le timbre de la voix mais également le langage du corps, i.e. les expressions du visage comme le sourire, les gestes ou encore la posture. Durant une interaction, un profil émotionnel et interactionnel de la personne (Devillers, 2010) peut être également construit pour mieux interpréter les émotions et adapter les stratégies du dialogue à chaque personne. Evidemment cette modélisation des émotions est simplificatrice ! Les meilleurs systèmes actuels ne peuvent détecter que peu d'expressions émotionnelles, et encore faut-il qu'elles soient exprimées de façon assez caricaturale. En général, les systèmes proposent la détection de 6 émotions : peur, colère, joie, tristesse, dégoût, surprise, avec quelques fois des nuances d'intensité : inquiétude, énervement, amusement. Il est très difficile d'aller au-delà dans l'état actuel des technologies. L'analyse de ces processus expressifs dans différents contextes peut amener à différentes interprétations émotionnelles. L'aspect socio-culturel et sémantique est encore assez peu développé.

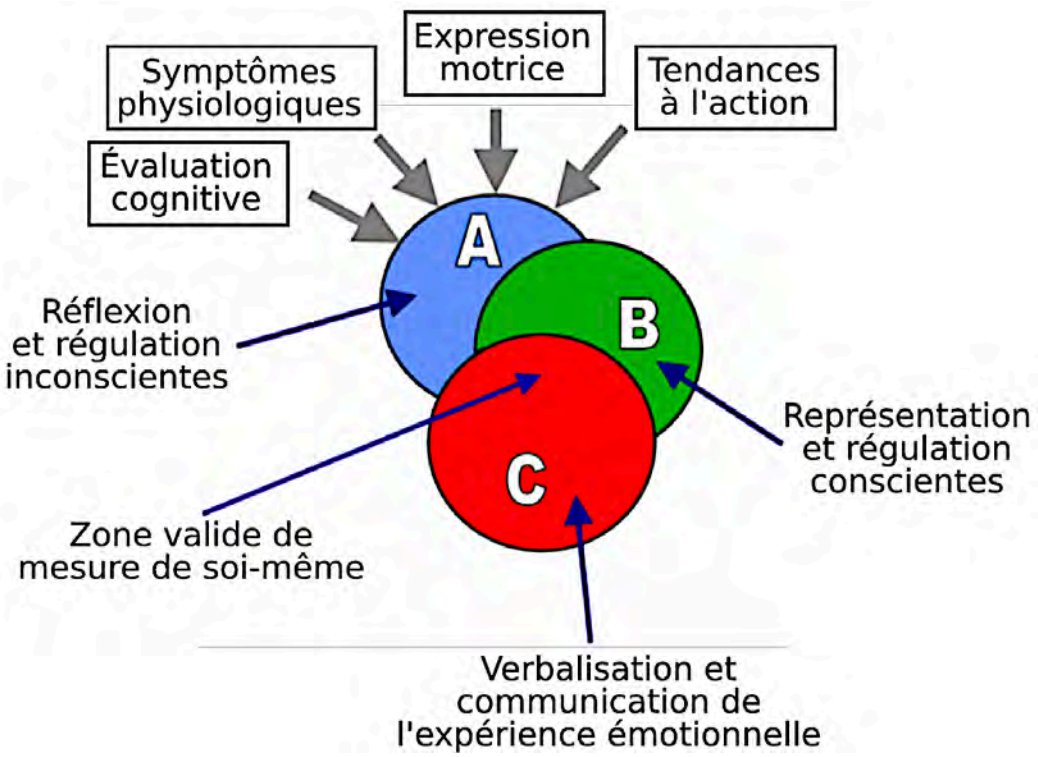



**Figure 1** : diagramme de Venn du processus émotionnel conscient et inconscient

On peut citer quelques modèles parmi les plus connus qui interprètent les émotions en fonction d'un nombre de critères d'évaluation à prendre en considération, leur nature et leur relation avec les émotions, comme le modèle OCC (Ortony, Clore, Collins, 1998) ou le modèle EMA (*EMotion and Adaptation*) (Gratch et Marsella (2004). EMA par exemple permet, d'une part, de représenter la relation causale entre les événements (passés, présents et futurs) et les états du monde courant, et d'autre part, de représenter le caractère subjectif de l'évaluation et de l'adaptation. EMA est la première tentative de modélisation d'adaptation émotionnelle (la réinterprétation de l'événement, l'acceptation, etc.) par des agents. Ce modèle est essentiellement axé sur les buts. Il ne prend pas toutefois en considération les standards moraux ou les préférences comme dans le modèle *OCC*. Ces modèles sont principalement utilisés pour modéliser des agents artificiels dans un monde virtuel. Ils ne sont pas au point pour une interaction naturelle avec une machine, encore moins pour une conversation avec échanges de points de vue. L'interaction avec une machine est « faussée » car il n'y a pas ou peu de co-construction, juste des programmes qui singent l'humain et qu'il est encore facile de débusquer.

## 3. Risques et bénéfices

L'informatique affective a ouvert un champ important d'applications que ce soit pour le diagnostic médical, la surveillance, l'assistance en santé (Nordlinger & Villani, 2018) ou encore l'éducation à l'aide de logiciels ou encore de robots. Même si les performances de ces modèles ne sont pas très robustes, ces modélisations engendrent chez les humains des comportements surprenants. L'anthropomorphisme, c'est-à-dire la tendance à attribuer aux machines des réactions humaines est alors bien réel. L'intelligence artificielle permet de détecter ou simuler les affects, elle ne peut pas les faire ressentir à la machine. La modélisation des émotions ne touche que la composante expressive, il n'y a ni sentiment, ni désir, ni plaisir, ni intention dans une machine. Mais ces machines numériques ont beau ne pas être vivantes, elles sont bien présentes.

La modélisation informatique des affects amène à se poser la question des conséquences sociétales de vivre dans un quotidien environné d'objets pseudo-affectifs (Devillers, 2017). Les chatbots aussi appelés agents conversationnels et les robots sociaux peuvent déjà embarquer des systèmes de détection, de raisonnement et de génération d'expressions affectives qui même avec des erreurs importantes peuvent interagir avec nous. Ils sont cependant loin d'avoir des capacités sémantiques suffisantes pour converser et partager des idées mais ils pourront bientôt détecter notre malaise, notre stress et peut-être nos mensonges.

Notre imaginaire et nos représentations symboliques vont évoluer au contact des machines, surtout si celles-ci nous parlent et simulent de l'empathie. Comment éviter l'isolement, la déshumanisation et l'appauvrissement de la vie sociale, si les interactions des humains sont réduites à discuter avec des machines affectives ? Il est nécessaire de réfléchir à la co-

adaptation humain-machine, aux bénéfices et aux risques engendrés par ces objets et aux garde-fous à imaginer pour éviter d'être trop manipulés par ces machines.

Il importe également de savoir si nous sommes prêts à accepter des robots capables de détecter nos émotions, d'y réagir et par exemple de simuler de l'empathie ; et si nous pourrons un jour nous attacher à des machines affectives comme nous nous attachons à un animal domestique. Dans ce cas, quels statuts auront ces individus numériques dans nos sociétés ? Même si ces objets ne sont pas réellement émotionnels, conscients et autonomes, ils vont envahir notre quotidien. Quels garde-fous seront développés ? L'alphabétisation de l'IA, l'éducation et l'expérimentation de ces machines affectives sont nécessaires pour que nous puissions prendre suffisamment de distance devant ces objets affectifs qui vont envahir notre quotidien.

**Références**


- D. Andler, Introduction aux sciences cognitives, Paris, Folio, coll. « Essais », 1992, 516 pages.
- H. Atlan, Cours de philosophie biologique et cognitiviste, Editeur Odile Jacob, 2018.
- A. Damasio, La construction du cerveau conscient, Editeur Odile Jacob, 2010.
- R. J. Davidson, K. R. Scherer, H. Hill Goldsmith, Handbook of Affective Sciences, Oxford University Press, 2003.
- L. Devillers, Challenges in real-life emotion annotation and machine learning based detection, Neural Networks Journal, vol 18, 4, 407-422, 2005.
- L. Devillers, L., Vidrascu, O., Layachi, O. Automatic detection of emotion from vocal expression in A Blueprint for Affective Computing, ed. Scherer et al., Oxford University press, 2010.
- L. Devillers, Des Robots et des hommes : mythes, fantasmes et réalité, 2017.
- B. Fehr & J. Russell, J. Exp. Psychol. 464-486, 1984.
- J. Gratch J., S. Marsella, A domain independent framework for modeling emotion, Journal of Cognitive Systems Research, vol. 5, n ° 4, p. 269-306, 2004.
- D. Kahneman, Thinking Fast and slow, Allen Lane, coll. « AL TPB », 1re éd., 512 p. (ISBN 978-1846146060), 2011.
- D. Kahneman, E. Diener et N. Schwartz, Well-Being: the Foundations of Hedonic Psychology, Russell Sage Foundation publications, 2003.
- P. R. Kleinginna and A. M. Kleinginna. *A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and Emotion*, 5(4):345–359, 1981 Calder, 2001.
- B. Nordlinger, C. Villani, Santé et intelligence artificielle, CNRS édition, 2018, chapitre L. Devillers, « Les robots sociaux et affectifs : une intelligence artificielle sans conscience mais utile ».
- A. Ortony, G. L. Clore et A. Collins, *The cognitive structure of emotions*, Cambridge University Press, Cambridge, MA, 1988.
- J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions*, New York, Oxford University Press, 1998.
- R. Picard, *affective computing*, MIT Press, 1997.

- K. Scherer, *On the nature and function of emotion: A component process approach*. Lawrence Erlbaum Associates, Publishers, Londres, 1984.
- K. Scherer, *Trends and development: research on emotions. Social Science Information, 2005.*
- K. Scherer, A. Schorr, T. Johnstone (Eds.). (2001). *Appraisal processes in emotion: Theory, Methods, Research*. New York and Oxford: Oxford University Press.
- K. Scherer**,** T. Bänziger**,** E. Roesch, A Blueprint for Affective Computing, ed. Scherer et al., Oxford University press, 2010.

# QUATRIÈME PARTIE

Table ronde dirigée

Par

Raja CHATILA

Avec

Antoine BORDES,

Gérard de BOISBOISSEL,

Laurence DEVILLIERS,

Francis EUSTACHE

# Table ronde du colloque
# Les signatures de la conscience[1]

La table ronde organisée à l'issue du colloque a été animée par Raja Chatila, entouré d'Antoine Bordes, Gérard de Boisboissel, Laurence Devillers et Francis Eustache. Une part des réflexions de la troisième partie a trouvé un écho dans cette table ronde, dont la présente synthèse constitue la quatrième partie de cet ouvrage. Quatre questions ont d'emblée été proposée par Raja Chatila en introduction et y ont été ensuite débattues : la possibilité (et le caractère souhaitable ou non) d'une conscience artificielle ; la délégation des décisions aux machines et les implications de cette délégation ; les interactions et interdépendances hommes – machines ; et enfin une brève évocation du problème de l'insertion des machines dans les réseaux sociaux

## **Peut-on envisager une conscience artificielle** ?

Que peut être la conscience d'une machine ? Une machine peut-elle être consciente d'elle-même ? Des interventions des participants à la table ronde et de celles du public présent se dégage une réponse plutôt négative. Toutefois, une telle réponse négative semble davantage relever de l'observation de l'état actuel de la technologie, ou encore d'un souhait ou du postulat d'une absence de besoin, voire d'un refus de l'alternative, que d'une argumentation rationnelle[2].

La réflexion sur ces questions se heurte en effet à plusieurs difficultés liées entre elles.

---

[1] Cette synthèse a été rédigée par Eric Chenin et Jean-Pierre Treuil. Elle a fait l'objet de plusieurs allers-retours avec Raja Chatila et les participants de la table ronde ; elle est publiée avec l'accord de tous. Les passages du texte en italique correspondent aux locutions utilisées par les intervenants.

[2] A ce sujet, étant donné l'enjeu pour l'homme, la question mériterait peut-être que l'on y consacre un effort important : peut-on démontrer que la machine n'accèdera jamais à la conscience ? Et sous quelles conditions ? Actuellement en effet, on ne sait pas déterminer les conditions qui pourraient garantir que la machine ne puisse pas devenir consciente, ni reconnaître les indices qui pourraient présager l'apparition de la conscience dans les machines : on ne sait donc ni empêcher cette apparition, ni l'anticiper. Cela semble constituer un risque sérieux pour l'humanité, peut-on se permettre de le négliger ? En relisant ce compte-rendu, Laurence Devillers fait remarquer, à propos de cette note, qu'il faudrait tout d'abord définir ce qu'on entend par conscience. Stanislas Dehaene explique dans son livre intitulé « Le Code de la conscience » que *les calculs mis en œuvre par les réseaux de deep learning actuels correspondent principalement aux opérations non conscientes du cerveau humain*. Les systèmes actuels d'IA ont la capacité d'extraire inconsciemment des corrélations de fait des informations, de prendre des décisions et d'apprendre, mais sans en avoir conscience.

La première est la difficulté à s'extraire d'un vocabulaire anthropomorphique lorsqu'on cherche à caractériser le comportement des robots et leurs relations avec les êtres humains. Ce vocabulaire est simultanément une facilité et un piège. En l'employant en effet on *humanise* les robots dans l'esprit du public et on peut être conduit à des raccourcis trompeurs. Pour Laurence Devillers, *il faut éviter les amalgames entre la conscience et la « conscience machine »* auxquels conduit l'utilisation, pour les algorithmes ou les robots, du vocabulaire propre au comportement humain. Par exemple, que recouvre la notion de « rêve » appliquée aux machines ? S'agit-il de la phase, chez l'homme, *où l'on construit sur les données enregistrées dans la journée, où l'on consolide ce qu'on a vu et fait dans la journée* ? et Laurence Devillers a poursuivi son propos, en observant cependant que si un robot est programmé pour mettre à jour périodiquement ses paramètres internes en fonction des données nouvellement enregistrées, *cela pourrait s'apparenter effectivement à cette phase de consolidation de la mémoire pendant le sommeil observée chez l'homme*[3]. Il y a bien, de ce point de vue, une analogie entre ce que fait l'homme et ce que fait la machine ; mais ce n'est pas pour autant une manifestation de la conscience chez la machine, où le sens et l'émotion, par exemple, restent absents de la reconfiguration des paramètres internes. Un exemple simple d'exercice de l'imagination où la machine a rejoint l'homme est la paréidolie, c'est-à-dire la capacité à voir des structures là où il n'y en a pas. On la considérait comme une manifestation typique de l'imagination humaine, mais en fait les algorithmes s'en avèrent parfaitement capables, comme l'a bien montré le programme DeepDream d'Alexander Mordvintsev. On est capable de simuler facilement ce genre de capacité avec des réseaux de neurones, mais la machine ne rejoint pas pour autant les capacités d'imagination de l'homme. Laurence Devillers poursuit : *les machines n'ont pas d'émotions, elles n'ont pas de corps, elles n'ont pas de ressenti, elles n'ont pas de conatus, d'intériorité,* la machine ne fait que ce pour quoi elle est programmée, *arrêtons d'abuser du langage : les émotions, la conscience, l'intelligence, l'autonomie, tout cela, c'est de la simulation ; et quand la machine fait quelque chose que l'on n'a pas prévu, souvent il s'agit de bugs.*

Antoine Bordes rebondit sur le thème des stratégies dans les jeux et l'apprentissage par renforcement ; il s'interroge sur l'interprétation des capacités de certains algorithmes à apprendre par eux-mêmes, sans supervision ; par exemple à apprendre à mieux jouer en jouant une multitude - typiquement plusieurs millions voire milliards - de parties contre un avatar d'eux-mêmes. On pourrait dire que ce sont des parties imaginaires et considérer que l'exercice s'apparente au rêve ou à un travail d'imagination. Peut-on dire qu'une telle capacité est la transposition pour une machine, de ce qu'est pour un être humain la faculté de rêver ou d'imaginer ? Antoine Bordes conclut en disant *qu'il n'irait pas jusque là.*

[3] Toutes les formes de mémoire sont concernées, mais chaque stade du sommeil joue un rôle assez sélectif. Lors du sommeil léger puis profond qui suit l'endormissement c'est la mémoire déclarative, faite de nos souvenirs et de nos connaissances, qui est consolidée. Lors du sommeil paradoxal, plus tardif, ce sera la mémoire procédurale, celle de nos habiletés motrices et perceptives.

La seconde difficulté est celle de définir la conscience autrement que par ses manifestations extérieures (i.e. dans les comportements), visibles ou mesurables. De fait, rien n'interdit d'interpréter une simulation de ces comportements comme procédant d'une conscience, d'une *intériorité*, d'une *intention*[4]. Laurence Devillers rappelle que Turing observait déjà *qu'on ne sait pas si un individu en face de soi est conscient, on ne connaît pas son intériorité.* Or on est bien dans la même situation dans notre interaction avec la machine. Mais malgré cette remarque, Laurence Devillers maintient sa position en insistant : *il faut bien comprendre que les machines produisent automatiquement des comportements qui nous ressemblent, mais sans aucune intention interne, sans ressenti, sans corps, sans physiologie, sans compréhension, sans aucun appétit de vie… elles ne font que reproduire des comportements qui nous paraissent intelligents.*

Si le robot décode des émotions, il le fait de manière « abstraite », il est simplement programmé pour apprendre à reconnaître des motifs dans le ton de la voix, dans l'expression du visage ou dans les gestes, et il trouve dans une base de données les catégories d'états émotionnels qui y sont associées. Mais, dira Francis Eustache, décoder les émotions ce n'est pas les ressentir ; il poursuivra cette observation, à partir de son expérience de psychologue : pour lui, la conscience *au sens fort du terme, et notamment la conscience morale*, n'est pas seulement liée au fonctionnement du seul cerveau : *le cerveau est dans un corps, je réagis avec mon cœur, si j'ai une émotion mon cœur se met à battre plus vite, je me mets à transpirer…* Il doute, tout en craignant qu'il puisse se tromper, que de telles réactions puissent être *modélisées* et devenir partie intégrante du comportement de robots humanoïdes.

Enfin, il y a la difficulté à parler de la conscience « tout court », sans l'associer à l'objet de cette conscience ; on parle ainsi de conscience de son environnement, de sa place dans cet environnement, conscience des possibilités de s'y adapter et des limites de ces possibilités d'adaptation, conscience des *autres*, de leurs situations, de leur attitude à notre égard, conscience de soi. Dans toutes ces expressions, c'est toujours *conscience de quelque chose*, Gérard de Boisboissel en liste plusieurs exemples comme la conscience de faire partie d'un contexte lié à l'action militaire, c'est-à-dire dans un environnement donné et le plus souvent imparfaitement reconnu, et d'être au cœur du déroulement de l'action ; mais aussi la conscience de ce que je peux faire pour les autres ainsi que pour la mission et de ce que les autres peuvent faire pour moi (comme m'appuyer, me soutenir, me secourir). La conscience de son environnement ne doit pas être confondue avec la simple existence, dans la machine,

[4] On peut noter à ce propos que parmi les capacités que l'on associe couramment à la conscience, celle qui serait la plus susceptible d'impacter notre relation avec les machines, serait leur capacité éventuelle à développer des *intentions* propres : l'intentionnalité serait la capacité au potentiel disruptif le plus fort, puisque rien ne garantit que les intentions propres que les machines développeraient seraient compatibles avec celles des hommes.

d'une représentation interne de son environnement[5]. En revanche, si la machine met à jour en continu cette représentation interne à partir des données dont l'alimentent ses capteurs, et si elle s'appuie sur cette représentation interne pour déclencher des actions, pour adapter en temps réel son comportement, et c'est justement le cas dans les situations évoquées par Gérard de Boisboissel, alors on se rapproche, malgré tout, de ce que pourrait recouvrir la notion de conscience de son environnement chez l'homme ou chez l'animal.

La réflexion des participants s'interroge également sur la possibilité effective, l'opportunité, l'intérêt pratique d'aller vers des *robots conscients.*

Tout d'abord, se manifeste un certain scepticisme quant au fait que l'on puisse « y arriver » (c'est-à-dire à une intelligence artificielle consciente) ; les échanges qu'un être humain peut avoir actuellement avec un robot sont très pauvres, voire *pathétiques* dans leur indigence. A contrario, les progrès constatés vers une meilleure prise en compte du *contexte*, par exemple dans le cadre d'interactions avec les humains, vers une meilleure capacité d'adaptation à des contextes, à des interlocuteurs variés, pourraient inciter à la prudence, quant aux conclusions définitives. Ces progrès dans la reconnaissance et l'adaptation à de multiples contextes provoquent, dira Francis Eustache, un double sentiment : à la fois une admiration devant ces avancées techniques *impressionnantes*, et une crainte, voire un *effroi* devant ce que cela pourrait signifier : un appauvrissement des interactions, une habituation à interagir avec des entités demeurant très frustes et conduisant à l'isolement ; ou à l'inverse l'être humain dépassé par une machine supérieure devenue consciente. A ce sujet Francis Eustache renvoie au livre qu'il a récemment coordonné ayant pour titre « La mémoire entre science et société » (Le Pommier, 2019), où ces aspects sont abordés de façon pluridisciplinaire.

Un autre argument alimentant le scepticisme repose sur le postulat déjà souligné que la conscience est liée à la vie, que c'est une propriété d'un être vivant, avec un cerveau certes, mais aussi un corps. Conscience implique un *contexte intérieur* qui lui soit accessible ; il s'ensuit cette conclusion : un ordinateur n'est pas un être vivant, un être produit par une longue évolution, donc il ne peut avoir de conscience, pour ainsi dire par construction.

Une dernière idée amenée par les intervenants est qu'un robot peut être dupliqué, qu'il n'a pas d'identité, ou encore que son identité n'est pas unique ; que l'intégration du robot dans les réseaux rend cette notion d'identité vide de sens[6].

---

[5] Une telle représentation de l'environnement, au sens large, de la machine, est en effet présente dans toute machine sous une forme ou sous une autre, sinon les interactions seraient impossibles, y compris entre algorithmes.

[6] La conscience au sens « fort », la conscience de soi, la conscience d'exister, d'avoir une identité, repose en effet sur l'existence d'une telle identité, d'un soi. Et cette notion est difficile à appliquer à un ensemble diffus d'algorithmes en interaction au sein d'un réseau, dont une partie est dupliquée sur plusieurs supports.

## Problèmes éthiques posés par les décisions prises par les machines

Un argument mis en lien avec la notion de responsabilité est qu'un robot n'a pas d'intentions, de motivations, de buts qui lui soient propres, car les *intentions* qu'il peut avoir lui sont *données* par le programmeur[7]. De fait, il n'exerce pas vraiment des choix, il n'est pas *responsable.* La responsabilité, qui est une notion morale et juridique, serait le propre de l'humain. La nuance qui est cependant apportée par un intervenant est que le niveau de généralité et d'abstraction du but qui est explicitement programmé peut être très élevé, si bien qu'alors le robot dispose d'une grande *autonomie* dans sa réalisation.

Conscience, identité, responsabilité - cette dernière impliquant la capacité de dire « non », soulignera Francis Eustache lors de la discussion : *si on est capable de dire non, c'est grâce à notre histoire personnelle, mais aussi à l'histoire qui vient de nos parents, à nos valeurs, et à notre Histoire avec un grand 'H'*. Ces termes sont, de façon sous-jacente, au cœur du débat autour de la différence entre *la machine et l'homme* et ses conséquences sur le plan de l'éthique.

Les questions posées autour de ce thème s'articulent autour de celle-ci : le robot est-il responsable de son comportement ? La réponse globale à cette question est que non, la responsabilité restera toujours - ou doit rester ? - in fine humaine, celle du producteur ou de l'utilisateur de la machine. Mais cette réponse générale et ses conséquences se déclinent de plusieurs façons, selon le contexte d'utilisation et le point de vue particulier à chacun des participants.

Ainsi que le soulignera Gérard de Boisboissel, le contexte militaire fait ressortir le lien entre ces questions, la notion d'autonomie et celle de délégation. Le robot militaire apparaît comme un système de fonctions, de catégories d'actions à accomplir, chacune ayant son degré d'autonomie propre ; par exemple la fonction *déplacement* où le robot peut être complètement téléguidé, ou au contraire laissé totalement libre du choix de son trajet, pour atteindre une destination donnée, en définissant lui-même son parcours et en contournant tout obstacle qui entraverait sa progression. Ou encore, la fonction de surveillance, où le robot peut être laissé libre du choix des secteurs d'observation ainsi que de l'orientation et du positionnement de ses capteurs. Par ailleurs, doter les robots de systèmes d'armement est inéluctable, car il est des usages où de telles machines apporteront un avantage tactique. Une question cruciale devient donc celle des *armes autonomes*, armes pouvant décider « librement » du choix de la cible et du moment du tir. Sur ce point une position peut être, par exemple, que ce droit de tirer ne puisse être délégué au robot que dans des circonstances particulières (situations saturantes ou milieux hostiles à la présence

[7] Au jeu de Go, l'intention de gagner est programmée ; de même, pour un robot assistant dans un EHPAD, l'intention d'exprimer et de susciter l'empathie est programmée.

humaine) ; selon cette approche, la délégation de tir doit nécessairement être bornée dans le temps et révocable à tout moment par la personne habilitée. *Abandonner tout contrôle sur le robot est un non-sens du point de vue de l'action militaire,* car c'est le chef qui donne du sens à celle-ci. L'utilisation par les militaires d'une machine ayant une forme d'autonomie doit garantir que cette machine soit sujette aux ordres et aux contre-ordres, et qu'elle rende compte tout comme n'importe quelle unité militaire[8]. Il est probable selon Gérard de Boisboissel que tous les pays, malgré leurs conceptions propres de la guerre, respecteront ces règles d'utilisation, avec des volontés plus ou moins marquées de développer des armements létaux autonomes. Reste le cas du terrorisme, qui lui souhaite au contraire que de tels systèmes puissent faire le maximum de pertes sans obligation de contrôle, et il faudra sans nul doute *tout faire pour éviter de tels scénarios.*

Un robot n'a donc qu'une autonomie déléguée, déléguée par des êtres humains ; mais cette autonomie partielle pose la question de l'éventuelle attribution aux robots d'une *personnalité juridique.* Sur cette question, selon Laurence Devillers, la réponse de la communauté concernée est *plutôt négative.* Elle évoque trois types d'arguments. Tout d'abord un argument financier, lié aux engagements d'assurance qu'il faudrait placer sur ces machines dotées d'une personnalité juridique, avec *des montants qui seraient supportables par de grands groupes mais pas nécessairement par de petites startups*. Ensuite, considérer qu'en cas de dommages causés c'est la machine qui est « responsable », n'incite pas à chercher l'origine du « bug » ou plus exactement de ce comportement dommageable du robot et à chercher à y remédier. Enfin, l'argument de la confusion qu'apporterait l'introduction de cette nouvelle catégorie du droit, dans un système où le matériel juridique permettant le partage de responsabilité entre les différents acteurs humains impliqués dans un dommage causé par un robot est déjà en place et suffisant, au moins dans un contexte civil[9]. Oublions donc, dira Laurence Devillers, cette idée de robots dotés d'une personnalité juridique, idée jugée inutile ; l'important est de bien comprendre quels sont les différents acteurs humains impliqués dans l'action du robot, la nature, le degré de leur implication.

Parmi ces acteurs, il y a les chercheurs, qui, en amont des applications opérationnelles, travaillent sur les algorithmes, et, en lien avec eux, ceux qui conçoivent ces applications. Puis les « entraineurs » - ceux qui vont entraîner le robot aux tâches qu'il doit effectuer, à partir de données qu'ils auront sélectionnées. C'est l'apprentissage, dans lequel la fonction

[8] On peut toutefois noter que vient toujours un moment où l'arme est livrée à elle-même, sans possibilité d'en reprendre le contrôle, cela parfois plusieurs dizaines de minutes avant son déclenchement létal effectif : c'est le cas par exemple des missiles balistiques de la dissuasion nucléaire, ou des missiles de croisière, ou, pour des durées plus courtes, des torpilles, ou des missiles sol-air ou air-air. Mais dans ces exemples, l'homme a explicitement désigné les cibles, et un tel choix n'est pas laissé à la machine.

[9] On sait qu'en France existent les notions de personne et de meuble, et dans cette catégorie les meubles *sensibles* que sont les animaux ; la responsabilité des dommages causés par ces différents meubles est toujours recherchée du côté des personnes, selon des règles bien établies.

objectif est fixée par l'homme. Il y a enfin les utilisateurs eux-mêmes, notamment dans l'action complémentaire qu'ils peuvent avoir sur le robot et son comportement, en lui apprenant par exemple un comportement non approprié ou peu judicieux.

Prenons d'abord, avec Antoine Bordes, le cas des chercheurs, qui se sentent effectivement porteurs d'une responsabilité. Selon lui, un consensus s'est dégagé dans la communauté de ces chercheurs, aussi bien dans le monde industriel comme à Facebook[10] ou Google, que dans le monde universitaire, pour qu'*il n'y ait pas de plan secret*, que tout passe par des publications « open access », en domaine ouvert, comme pour les travaux de recherche des autres disciplines. Contrairement à ce qui peut être craint ici ou là, *il ne peut y avoir ainsi d'algorithme superpuissant que personne n'aurait vu venir*. La raison de ce comportement n'est pas seulement un désir spontané d'ouverture des grands groupes, mais le fait que les laboratoires sont bien sous le contrôle des chercheurs, qui sont, à la base, imprégnés des principes de la recherche universitaire[11].

Cette situation, dira Antoine Bordes, ne signifie pas pour autant, de l'aveu même de ces chercheurs, que les systèmes actuels ne peuvent pas être appliqués de manière problématique ; il y a nombre de points auxquels il faut faire attention. C'est notamment le cas dans le contexte de l'apprentissage profond ; car le robot que l'on entraîne pour une tâche donnée, apprend à partir de ce qu'on lui présente, et ce qu'on lui présente peut être affecté de biais de natures et d'origines diverses. Ces biais peuvent entraîner de mauvaises décisions, mais on ne peut pas à proprement parler les considérer comme des bugs. C'est d'autant plus problématique que l'algorithme n'est souvent pas capable d'expliciter ce qui le conduit à fournir telle réponse ou à prendre telle décision, et donc ceux qui ont fourni la base d'exemples peuvent très bien ne pas avoir conscience des biais que celle-ci contient. Ces biais d'apprentissage ne sont pas qu'une crainte théorique, de nombreux exemples ont pu être constatés ; exemples dans lesquels le comportement de la machine s'avère discriminant, toutes les personnes n'étant pas traitées de la même façon selon qu'il s'agit d'hommes ou de femmes, ou selon la couleur de la peau[12].

La question de la confiance à placer dans la machine, dans sa capacité de non-discrimination, dans son équité, voire dans sa *loyauté*, terme employé par Laurence Devillers, se pose dans de nombreux contextes ; un des facteurs souvent évoqué en faveur de la machine est son insensibilité aux émotions, à la fatigue ; cette insensibilité lui permettrait alors de proposer, sans être affectée par de tels facteurs, contrairement aux

[10] Ainsi Facebook a publié en deux ans plus de 200 articles, tous accessibles en plein texte.

[11] En relisant ce compte-rendu de la Table ronde, Raja Chatila fait cependant remarquer que les entreprises sont libres de ne pas tout publier.

[12] Les conférenciers ont cité ainsi les algorithmes mis en place pour le déblocage des téléphones, qui se sont avérés fonctionner de façon moins efficace pour les personnes noires ; ou encore des algorithmes manipulant le langage et reflétant, par exemple dans le domaine des métiers, les stéréotypes distinguant métiers « masculins » et « féminins » et la réalité sociale qui leur est liée.

humains, des décisions *rationnelles* et *justes* selon les règles prescrites ; de fait des études[13] ont montré que, dans le domaine judiciaire, des juges n'avaient pas, statistiquement, la même sévérité selon le moment de la journée, et donc qu'une machine serait plus constante ; mais encore une fois, si le biais émotionnel ou de fatigue y est inexistant, d'autres biais, on l'a vu, interviennent, et peuvent conduire la machine à faire des choix que personne n'avait anticipé. Selon Laurence Devillers, deux niveaux peuvent être ainsi distingués ; le premier est celui des biais d'apprentissage initiaux, résultat d'un mauvais calibrage des données et aboutissant à ce qui peut être appelé la *bêtise artificielle* ; le second est lié à ce que la machine apprend au cours de son utilisation, qui peut remettre en cause ce qui a déjà été vu, ou encore au fait que l'effet de certains paramètres, fixés par les programmeurs, n'est pas bien contrôlé.

Cette question de la confiance, plus généralement des interactions entre l'homme et la machine dans la prise de décision, rejoint, dans certains cas, des questions éthiques majeures. Dans le contexte médical par exemple, notamment celui de la fin de vie, le fantasme d'une délégation totale du diagnostic ou de la décision à l'algorithme existe et fait peur. Mais justement est-ce seulement un fantasme ? Avec Francis Eustache, on doit s'interroger sur le risque, que l'on peut estimer réel, d'une tendance de plus en plus accentuée vers l'application de procédures automatiques : dans de telles procédures, dès lors que le robot prendra telle décision, celle-ci sera appliquée, l'homme n'ayant pas la possibilité de faire intervenir son propre jugement ; soit parce que c'est ce que prescrit le règlement, le médecin contrevenant risquant alors une sanction ; soit, à l'extrême, en raison d'une automatisation complète.

Il s'agit donc de questions extrêmement importantes, dont la résolution fait l'objet d'un domaine actif de recherche : donner aux algorithmes basés sur l'apprentissage profond la capacité de mieux expliciter leurs résultats ; créer des outils à même d'*évaluer* les systèmes mis sur le marché en en analysant les biais, les façons de répondre, les discriminations de toutes sortes ; et pour mener au mieux cette entreprise, concevoir ces *agents évaluateurs*, travailler en interdisciplinarité avec des économistes, des juristes, des informaticiens ; c'est bien de recherche qu'il s'agit, dont le but est de permettre au public et aux décideurs d'avoir confiance dans ces produits ; *s'ils ne recueillent pas cette confiance, ils ne seront pas vendus, et pour le monde militaire ne seront pas utilisés par les chefs et les opérateurs qui en auront la responsabilité.*

## **Interactions et Interdépendances hommes-machines**.

C'est un vaste sujet, qui a été abordé sous plusieurs angles : celui de la menace de perte d'humanité, et celui de la préservation du contrôle de la machine par l'homme ; avec deux perspectives, à court terme à l'échelle de l'individu, et à long terme à l'échelle de la société.

[13] Etudes dont la validité est toutefois discutée.

Pour le court terme à l'échelle de l'individu, la question est d'abord traitée sous l'angle des implants, et notamment des implants cérébraux. Avec une question symbole : à partir de quand un homme complété par des implants devient-il un robot ? Deux types de positions s'expriment, qui, sans être contradictoires, manifestent un certain malaise.

Les implants sont acceptables quand il s'agit de *réparation*. Dans le contexte militaire, un homme blessé dont certaines fonctions vitales sont accomplies par des implants *fait toujours partie de la famille humaine*. Dans le contexte médical, on reconnaît que les implants constituent un progrès considérable dans le suivi et la thérapeutique de certaines maladies[14]. Mais faut-il aller de l'homme réparé à l'homme augmenté ? Par exemple avec la mise au point d'implants cérébraux qui permettraient la communication directe, non médiatisée par le langage ? Ne serait-ce pas un très grand danger, est-ce qu'il n'y a pas un gros risque que ces technologies permettent des manipulations d'individus qui perdraient leur *libre arbitre*, se transformant de facto en robots ?

Le danger, selon Laurence Devillers, est en effet *que l'homme puisse être piloté par ses implants, que ceux-ci prennent le contrôle, et qu'il perde ainsi sa capacité de décision. A partir du moment où il ne peut plus décider, où il est gouverné par la machinerie qui est dans son cerveau, on peut considérer que l'homme devient une machine. Les implants cérébraux sont utilisés jusqu'ici pour des pathologies particulières où il n'existe pas encore d'autres solutions pour l'instant, comme la maladie de Parkinson, avec des effets positifs*. Mais certains industriels, notamment parmi les GAFA, commencent à parler de BCI –Brain Computer Interaction- : selon eux, on pourra se connecter directement et communiquer plus vite sur Internet via la pensée. *Le langage deviendrait ainsi obsolète, or celui-ci fait partie de notre culture, il véhicule aussi ce que l'on a d'humain*. Une telle communication sans langage serait dangereuse. Elle poserait des difficultés en matière de transparence, de partage, d'archivage à long terme, ainsi d'ailleurs que de faculté d'abstraction, de raisonnement ou de conceptualisation complexe.

En fait cette idée d'homme augmenté est-elle vraiment possible ? Certains en doutent, présentant les affirmations de certains industriels ou penseurs comme de l'*esbroufe*, voire des tentatives de manipulation ? On fait aussi remarquer qu'un être humain est un système de fonctions[15] qui sont pour une grande part (70%) des sous-systèmes autonomes et indépendants de tout appel à la conscience. Chacun de ces sous-systèmes peut être effectivement réparé en cas de dysfonctionnement, mais cette proportion doit être rappelée

[14] Notamment le diabète, la maladie de Parkinson. Francis Eustache mentionne ainsi le traitement de certaines formes de Parkinson avec implants stimulant le noyau sous-thalamique, pour lequel la France a été pionnière. Il évoque également des essais prometteurs concernant la maladie d'Alzheimer, mais exprime un certain doute concernant les implants mémoire.

[15] Un intervenant cite ainsi le système hormonal, le système digestif, les fonctionnements réflexes comme ceux de la marche. Laurence Devillers fit remarquer dans le même esprit que l'on n'est pas conscient du fonctionnement de notre corps et de notre cerveau.

semble-t-il dans toute réflexion sérieuse sur ces sujets. De même que doit être rappelé ce qui fait l'identité d'un individu, l'importance de son histoire personnelle et l'histoire de ses relations avec les autres.

Car justement, une plus grande autonomie n'est pas toujours à rechercher ; au contraire, notamment dans le contexte militaire, comme le redira Gérard de Boisboissel, les programmeurs de ces systèmes doivent veiller à ce que le *contrôle* de l'homme sur le robot puisse être toujours assuré. En fait, on n'a nul besoin de robots conscients[16] ; on leur demande seulement d'être à même d'accomplir les tâches auxquelles ils sont destinés ; ces tâches peuvent requérir un haut degré d'automatisation, d'adaptation au contexte, de *réflexivité*, mais la conscience, au sens que nous donnons intuitivement à ce terme, est inutile car la responsabilité d'actes conscients revient au chef militaire.

Sur ce plan du contrôle, Gérard de Boisboissel souligne que l'utilisation des machines nécessite que l'on ait *confiance* en elles, et pour cela, *que l'on comprenne comment elles opèrent, que l'on ait une explication de leurs mécanismes* en disposant des informations relatives à leurs processus d'analyse et aux choix qui en découlent, afin de pouvoir discerner et valider ou invalider leurs comportements.

Antoine Bordes apporte une nuance à propos de la possibilité d'expliquer les actions de la machine, il conçoit que l'on souhaite pouvoir *interpréter la trace de la décision d'une machine* comme une séquence déductive explicite, mais il *pense que cela n'arrivera pas systématiquement, sauf dans certaines situations, de même qu'un humain qui a pris une décision n'est pas toujours capable* d'expliciter le cheminement qui l'y a conduit. Antoine Bordes souligne en effet que l'homme lui-même a souvent des difficultés à expliquer son processus de décision : il prend l'exemple du joueur de football, *qui ne sait pas toujours expliquer* comment il contrôle le ballon, ni comment il décide à qui le passer : les tentatives d'explication sont généralement partielles et subjectives.

A l'échelle de la société, le débat a touché à certaines implications possibles de la notion d'homme augmenté. De fait, la communication directe entre l'homme et la machine via des implants cérébraux n'a d'intérêt, en pratique, que si elle fonctionne dans les deux sens. Or, dans le sens réciproque « Computer Brain Interaction », un tel mode de communication ouvrirait par exemple la porte à une commande du cerveau par la machine –y compris via de fausses informations, à défaut de commande directe-. Les implants cérébraux permettraient ainsi de commander *à distance des armées d'individus robotisés,* via les machines, voire permettraient aux machines elles-mêmes d'en prendre le contrôle.

Sur le plan du contrôle justement, une crainte a semblé sous-tendre le débat : celle d'une perte de contrôle de l'homme sur l'évolution de la société sous l'emprise grandissante des

[16] Aucun programmeur ne souhaite a priori mettre de conscience machine dans son programme, dira ainsi Antoine Bordes.

machines. Laurence Devillers a ainsi préconisé de limiter les développements en matière de robots et d'algorithmes *à ce qui serait le plus utile, pour les personnes âgées, les personnes moins âgées, pour nous faciliter la vie, mais sans nous rendre trop paresseux, ni trop dépendants des machines*[17].

En conclusion sur ce thème de l'interaction homme-machine, il a été regretté que les pouvoirs publics soient parfois mal informés. En particulier, selon Laurence Devillers, *il faut réagir très clairement contre les implants cérébraux*, s'ils ne sont pas justifiés par l'intention de réparer ou de traiter une maladie. Et il est inquiétant de voir que certaines institutions puissent être sensibles à des discours prônant l'eugénisme ou l'*augmentation* de l'homme, sous prétexte de baisse du QI moyen ou de compétition avec l'intelligence artificielle.

## Intelligence artificielle et Insertion dans les réseaux sociaux

C'était un sujet trop vaste pour être traité dans le temps disponible. Et son rapport avec le cœur du thème central du colloque, voire avec celui de l'intelligence artificielle, n'est pas immédiat. Les interventions et la discussion ont cependant pu aborder plusieurs points.

Antoine Bordes rappelle l'un des caractères des réseaux sociaux : une plateforme telle que Facebook permet à des *milliards* d'entités d'échanger des informations de façon quasiment instantanée ; ces entités sont des personnes, mais aussi des *groupes politiques, des journaux, des entreprises, voire des algorithmes* … ; l'échelle de cet *énorme espace d'expression, de liberté, qui, de plus, s'auto-développe*, engendre des phénomènes qu'on ne prévoyait pas. L'observation et éventuellement, si nécessaire, le contrôle de ce qui se passe dans cet espace est difficile, compte tenu de sa taille mais aussi de la rapidité de son fonctionnement. La détection de *contenus inappropriés* (fausses nouvelles, faux comptes, incitations à la haine et autres propos interdits par la loi, …) a d'abord été fondée sur des signalements : le contenu suspect est signalé à une équipe de vérifieurs qui décident s'il est ou non conforme aux règles de la plateforme, règles qui sont publiques et qui peuvent donc être confrontées aux différentes législations. En cas de non-conformité, ces vérifieurs décident éventuellement du retrait du contenu incriminé, mais bien sûr si ce contenu a été signalé c'est qu'il a déjà été vu et donc vu, a priori, par un grand nombre de personnes … L'intelligence artificielle est apparue comme un outil pouvant aider[18] un tel processus, notamment pour le rendre plus rapide. Mais d'un autre côté, de telles interventions se heurtent à un principe, qui est *que l'on n'est pas là pour faire de la censure.* La solution serait peut-être de rechercher *comment faire comprendre au plus vite aux récepteurs que*

[17] A ce propos a été cité le rapport récent de la mission Villani, qui préconise un usage raisonné des NTIC, y compris pour des raisons écologiques, du fait de leur coût élevé en énergie.
[18] La décision quant au retrait devant toutefois rester humaine.

*tel ou tel contenu n'est pas fiable* sans le censurer. Il s'agirait donc, dans l'idéal, de trouver des moyens pour attacher à l'information elle-même, une certaine valeur de fiabilité.

Laurence Devillers reprend ce thème, en évoquant un certain scepticisme devant ce qu'elle qualifie de *scotchs*, qui seraient mis ainsi sur certains contenus. Car le nombre de *fakenews* observées sur les réseaux sociaux, d'informations qui ne sont pas avérées, lui parait littéralement atterrant. Le web est apparu au début comme un formidable outil de mise à la disposition du plus grand nombre des connaissances acquises par l'humanité. Mais très vite, constate-t-elle, une grande anarchie s'est faite jour, poussant en avant des contenus qu'il est très difficile d'évaluer : *pour faire le tri entre ce qui est avéré, pas avéré, ou carrément n'importe quoi, il faut être très bien outillé.* On pourrait craindre – une hypothèse certes extrême- que ce développement anarchique nous échappe, entraîne une prolifération *de fausses connaissances, de fausses idées, de faux amis, de faux algorithmes,* et rende le web inutilisable. Une telle issue pourrait être partiellement une conséquence d'une révolte collective plus ou moins consciente devant la tendance d'une société hyper-connectée *à nous enfermer*. Et Laurence Devillers de citer un des angles d'attaque de cette tendance, qui est celui de notre santé : avec un scénario où nous serions environnés par une multitude d'objets connectés qui nous aideraient théoriquement à bien vivre, mais qui en fait nous piègeraient en décidant pour nous. *Pourra-t-on supporter cela ? N'y-a-t-il pas un risque d'explosion ?*

Francis Eustache fait alors remarquer que ce développement des réseaux sociaux associé à l'utilisation parfois effrénée d'Internet constituent des changements majeurs, dont certaines générations ont pleinement conscience, car elles ont connu l'état antérieur. Mais pour les plus jeunes malheureusement, il n'en est pas ainsi. Un des points caractéristiques de ce développement est *l'externalisation de la mémoire. La mémoire est de plus en plus externalisée*. Bien sûr, ce n'est pas nouveau dans l'histoire de l'humanité. Mais l'ampleur du phénomène est sans précédent. Quelles en sont les conséquences ? Notamment sur le plan psychologique, l'exigence, la nécessité d'une mémorisation interne disparaissant peu à peu ? Francis Eustache avance la nécessité de réfléchir sérieusement à cette évolution de *l'équilibre entre mémoire interne et mémoire externe*. Concernant toujours les conséquences psychologiques, que dire aussi de la prégnance, dans les réseaux sociaux, d'une sollicitation permanente faisant perdre le contrôle de la situation, que dire de la prégnance d'un fonctionnement de type stimulus-réponse excluant tout temps de réflexion ? Alors que pour nombre d'opérations, *l'assimilation du contenu d'un ouvrage très pointu par exemple*, nous avons besoin, pour pouvoir les accomplir avec efficacité, de temps et d'un certain isolement. Cette dernière réflexion amène Francis Eustache à s'interroger sur ce qu'il en est dans l'action militaire.

Gérard de Boisboissel répond qu'effectivement la possibilité technique pour les soldats de se connecter à tout moment à leurs proches, nouvelle dans l'histoire militaire, poserait problème si elle était autorisée ; les jeunes ont le réflexe d'utiliser constamment leur smartphone dans la vie civile, pour communiquer avec leurs familles et leurs amis, et ils

ont parfois du mal à comprendre que cet usage puisse être contrôlé. Pourtant, les raisons de ce contrôle sont multiples et assez évidentes : tout d'abord pour la préservation du moral du combattant projeté en opération, pour l'émergence et le maintien d'un esprit de groupe et de sa cohésion, et enfin pour de simples raisons de sécurité, l'adversaire pouvant avoir les moyens de détection appropriés et d'intrusion dans les réseaux. Gérard de Boisboissel aborde alors un autre point important lié à la connectivité : il est clair que dès à présent et a fortiori dans l'avenir, tous les équipements militaires déployés et les personnels mobilisés dans une opération feront partie d'un système global interconnecté, et produiront, transmettront, exploiteront un volume considérable de données, images, vidéos, etc. C'est une tendance qui lui semble inexorable et qui implique que l'art de conduire les opérations militaires doit s'adapter. Il s'agit d'un changement majeur qui n'est pas sans poser un certain nombre de problèmes nouveaux. Le problème de la sécurisation de ces réseaux, même s'ils sont non-ouverts sur le monde extérieur ; le problème de la *saturation du spectre*, lié au volume des données à transmettre, *un soldat n'ayant pas une fibre optique derrière lui*. Enfin, le problème de l'exploitation de toutes ces données : *qu'est-ce qu'on en fait, qui y a accès ? Telle image, prise par un robot, doit-elle remonter au chef de groupe, de section, au capitaine, etc…* Or le principe d'horizontalité à la base des réseaux peut entrer en contradiction avec le principe de subsidiarité nécessaire à la sécurité et à l'efficacité de l'organisation et de la prise de décision militaires[19] : certaines informations ne doivent être transmises, au moins dans un premier temps, qu'aux seuls échelons pour lesquelles elles seront utiles dans l'action et qui ont la responsabilité de la conduite de l'opération. Aux échelons désignés de traiter et, pour ne pas encombrer les échelons supérieurs, il faudra activer un filtre des données qui sera fonction de la phase en cours de l'action militaire et de l'importance des données. L'IA pourra aider à la gestion dynamique de ce filtre. Et Gérard de Boisboissel de conclure : *La bonne résolution de ce problème est un vrai enjeu de l'interconnexion des systèmes futurs.*

## Conclusion

Raja Chatila conclut le débat en résumant ce qu'il retient des interventions sur les quatre grands thèmes :

- La possibilité d'une conscience artificielle : *non, cela n'existe pas, voire cela ne peut pas exister, aujourd'hui c'est un concept qui n'aurait pas de sens pour une machine.*
- L'éthique et les décisions prises par les machines : *la responsabilité sera toujours celle des humains qui ont conçu, déclenché et contrôlé la machine.*
- Les interactions et les interdépendances hommes-machines : *sur l'intégration de l'humain et de la machine par le biais de l'augmentation,* il relève une alerte *sur cette intégration qui transforme l'homme* et présente des dangers, même si elle peut aussi être bénéfique.

[19] Ce problème se pose bien sûr également dans d'autres contextes.

- L’insertion des machines dans les réseaux sociaux : elle pose plusieurs problèmes, notamment : l’externalisation de notre mémoire, qui *transforme nos mécanismes de pensée* ; la difficulté de contrôler les réseaux, en évitant à la fois censure et manipulation ; *la connexion vers l’extérieur, qui nous extirpe de notre réalité* ; et la diffusion inopportune des données collectées.

# ÉPILOGUE

Dans cet épilogue, nous nous proposons de faire une courte synthèse des réflexions que les thèmes abordés dans les différents chapitres de cet ouvrage ont inspirées aux membres du comité de notre société savante ayant conduit son édition. Nous allons le faire d'abord autour de la notion de conscience. Puis nous présenterons quelques éléments de discussion plus particuliers. Qu'on nous permette à titre de préambule de dire ceci : ces réflexions n'engagent que les personnes de ce comité, et non les auteurs des chapitres concernés.

**Autour de la notion de conscience**

Comme point de départ de la synthèse sur ce thème, il nous a paru utile de partir de l'expression « corrélats neuronaux de la conscience » ; cette expression semblerait indiquer l'existence sinon de deux réalités, au moins de deux aspects, de deux points de vue sur le même phénomène : deux points de vue susceptibles d'investigations séparées et indépendantes, avant de pouvoir être reliés.

Or manifestement, autant les structures et dynamiques neuronales sont des « objets » accessibles à la mesure et à la caractérisation - et on peut faire l'hypothèse qu'elles le seront dans l'avenir à des résolutions spatiales et temporelles de plus en plus fines - autant « la conscience » nous a paru être d'emblée « quelque chose » dont la définition même est problématique. Alors même que nous prétendons en avoir tous une expérience intime, et qu'elle a pu être posée comme la seule réalité dont nous pouvons être certains.

Face à ce déséquilibre, deux attitudes se manifestent, chacune à la racine, comme développé dans la présentation générale, de plusieurs courants philosophiques, mais aussi de recherches en neurosciences et en informatique.

Il y a d'abord l'attitude consistant à se contenter de recherches sur les signes de la conscience, cette dernière étant définie opérationnellement par le rapport qu'en fait le sujet humain participant aux expériences, dans son interaction avec l'expérimentateur. Poussée sans réserve sur un plan philosophique, cette attitude pourrait conduire aux versions radicales du matérialisme éliminativiste - la conscience n'est qu'une illusion. Sauf que précisément, les résultats des recherches aboutissent bien à « objectiver » la prise de conscience : il y a bien - avons-nous compris - dans le cerveau une différence matérielle, caractérisable, entre les phénomènes inconscients et ceux rapportés conscients. Objectivité renforcée s'il s'avérait, au bout du compte, qu'à la variété des états de conscience correspondrait de façon biunivoque une variété de phénomènes neuronaux spécifiques de ces états conscients. On peut alors se replier sur une position consistant à affirmer que bien qu'il s'agisse d'une réalité, il n'y a pas de définition de la conscience, en dehors de ses manifestations neuronales, ou encore à affirmer qu'il est inutile d'en chercher une, le seul objectif scientifique atteignable étant la caractérisation de son substrat physique et biologique. Les modèles élaborés pour rendre compte des observations faites sont dans

cette attitude des modèles des manifestations neuronales de la conscience, et non des modèles de la conscience en soi.

L'autre attitude, tout en se gardant de tout dualisme, considère que la conscience est une réalité complexe mais descriptible, analysable en termes de différentes composantes, de différentes fonctions ou propriétés. Cette attitude ne s'oppose pas fondamentalement à la première, puisque des corrélats neuronaux peuvent être recherchés pour chacun de ces différents aspects. De telles analyses ont été assez peu abordées dans l'ouvrage. Dans la première partie, le chapitre de Claire Sergent traite essentiellement de la conscience perceptive ; cependant l'embrasement d'un espace de travail global lors de la prise de conscience d'un stimulus peut se voir comme l'exécution d'une fonction de la conscience : fonction consistant à *donner du sens* à cette perception, c'est-à-dire à la relier à d'autres stimuli simultanément perçus, à des souvenirs d'expériences vécues antérieurement, autrement dit à notre représentation interne du monde[1] ; ce dans le but que l'individu se sente « en compréhension » avec son environnement, qu'il puisse agir avec l'efficacité nécessaire à sa survie. Les travaux de Jérôme Sackur portent sur des fonctions associées à la conscience comme l'introspection, ou la capacité de concentration ou d'attention. Dans la troisième partie, le chapitre de Franck Cosson aborde la question des niveaux de conscience comparés dans la diversité du monde animal. Gérard de Boisboissel évoque la relation de la conscience avec la représentation du fait que l'entité concernée traite de l'information ; ou encore la représentation d'un ensemble de choix (d'actions possibles), la capacité de « calculer », puis d'évaluer les conséquences de ces choix – évaluer au sens du risque encouru, de l'efficacité, de la « moralité »… Enfin Laurence Devillers met en avant la relation entre conscience et émotions.

En continuité avec de Boisboissel, et dans cet esprit d'une approche de la conscience « en soi », nous nous sommes interrogés sur la relation entre conscience et « savoir », « savoir » au sens de disposer de « modèles » de notre environnement, de la manière dont il peut réagir à nos propres actions individuelles et nous affecter en retour. Cette relation rappelle celle évoquée dans la littérature philosophique sur le thème conscience et représentation. Elle évoque aussi l'une des fonctions attribuée[2] à la conscience, la fonction d'auto-évaluation des connaissances acquises, de leur degré de fiabilité. Savoir que l'on sait, savoir ce que l'on sait, et inversement savoir qu'il existe des choses que l'on ne sait pas, par exemple, ne serait-il pas un marqueur significatif d'un certain niveau de conscience[3] ? On imagine que de telles questions pourraient être à la racine d'une boucle infinie, « savoir que l'on sait que l'on sait… », etc. Et on en arrive à une association de termes qui revient

[1] Et réciproquement, peut-être avec un certain décalage temporel, cette représentation du monde se mettra à jour, se corrigera et s'affinera au vu de cette perception ; mais cette seconde phase, d'apprentissage, n'est pas nécessairement consciente.

[2] Cf la publication citée dans la présentation générale de l'ouvrage, Stanislas Dehaene et al. What is consciousness, and could machines have it ? Science, oct.2017 et la composante de la conscience désignée par "C2" dans cette publication.

[3] Cf le philosophe Alain dans ses *Propos* : « Savoir, c'est savoir que l'on sait ».

souvent, mais qui n'est pas très claire à nos yeux, conscience et « réflexivité ». D'autres discussions ont porté sur le rôle de la conscience en tant qu'outil de sélection – composante « C1 » dans la publication qui vient d'être citée : tout être sensible vit en permanence un continuum de sensations et de processus mentaux. La majeure partie des sensations reste ignorée, et des processus mentaux fonctionnent en automatique, sinon la conscience serait saturée. Mais cette dernière semble scruter en permanence ce continuum, et elle ramène au niveau conscient une sensation ou un processus mental particulier lorsque cela s'avère nécessaire.

Aucune de ces approches de la conscience « per se » n'arrive vraiment à formaliser, nous semble-t-il, l'objet du « problème difficile », savoir ce qu'est le ressenti subjectif vécu par chacun d'entre nous. Mais, on le pressent bien, une réflexion sur de telles approches, lorsqu'elle est poussée à son terme, peut aboutir – et aboutit de fait dans certaines publications - au moins à des définitions fonctionnelles voire formelles de la conscience : de même que pour l'Intelligence dont certaines définitions ont conduit à l'Intelligence Artificielle, de telles définitions de la Conscience, la détachant de son support physique, conduisent à envisager une « Conscience Artificielle »[4].

On retrouve là un débat, largement sous-jacent aux deux derniers chapitres de la troisième partie, et sur lequel les membres de notre comité ont beaucoup échangé : la conscience, dans ses composantes jugées les plus complexes, dont celle relevant du « problème difficile », le vécu subjectif, est-elle une spécificité du monde vivant, ou au contraire, pourra-t-elle être a contrario, un jour, présente chez des machines ?

Certains d'entre nous posent comme *principe* qu'il n'y a de conscience que dans le vivant ; certes, il se pourrait que hors du vivant se manifestent certains phénomènes qui *ressemblent* ou qui *simulent* la conscience, mais ce terme doit être réservé à la désignation d'une capacité des êtres vivants. Le parti pris d'un tel principe part de l'observation que la conscience chez les êtres vivants est le résultat du lien progressivement tissé entre ces êtres et leur environnement au cours de millions d'années d'évolution darwinienne, et qu'en conséquence ce phénomène – en tant « qu'invention de l'évolution biologique » ne saurait qu'être spécifique à la vie[5]. Une autre racine du même principe – peut-être différente mais qui rattache toujours la conscience à la biologie - est l'idée que la conscience a d'abord été celle des états internes du corps et de ses besoins les plus essentiels tels que la faim ou la soif, une sorte de conscience primaire ou primitive.

[4] Cf également cette publication au titre significatif : Michael Graziano, The Attention Schema Theory, A Foundation for Engineering Artificial Consciousness, *Frontiers in Robotics and AI*, nov. 2017.

[5] Si l'on suit la logique de cette position, on pourrait alors considérer de même que la vision, ou la locomotion, sont réservées au vivant et interdites aux machines, ce que démentent les développements actuels de la robotique.

On peut bien sûr rejeter le principe posé précédemment, en partant de l'idée que la conscience peut - ou pourra - être définie indépendamment de son support physique et donc pourra un jour animer des « êtres artificiels ». Mais cette position se heurte à la complexité du cerveau humain, avec ses quelque 100 milliards de neurones et ses quelques millions de milliards de synapses, et une organisation qui est là encore le résultat d'une très longue évolution. La reconstruction d'une telle complexité est-elle vraiment possible ? Quoique la relation entre la capacité de conscience et le niveau de complexité de son support n'ait semble-t-il pas encore été établie...

**Mémoire humaine et mémoire informatique**

Francis Eustache/Armelle Viard dressent un panorama de différents modèles de la mémoire humaine. Ce terme de mémoire étant également très largement utilisé en informatique, il a paru légitime de nous interroger sur ce qu'ils ont en commun et sur ce qui les différencie dans les deux domaines.

Sur le plan de l'organisation, il apparait qu'aussi bien chez les êtres humains que dans les ordinateurs, il faille parler non d'une mémoire, mais de mémoires au pluriel, de différentes composantes ayant chacune leur spécificité. En informatique, ces composantes se différencient selon le volume, la rapidité d'accès, la place dans une hiérarchie hautement structurée et précisément connue, allant des mémoires « caches » des unités centrales, aux mémoires externes locales ou distantes. Concernant la mémoire humaine, on a pu parler d'une distinction entre mémoire à court terme ou de travail et mémoire à long terme ; par ailleurs, des différenciations hiérarchiques ont bien été proposées, attribuant par exemple à certaines structures cérébrales la mémorisation de représentations simples, et à d'autres structures celles de représentations plus complexes. Mais d'autres différenciations se basent plutôt sur le contenu, par exemple celle distinguant mémoire épisodique et mémoire sémantique, ou sur le degré d'accès à la conscience, mémoire explicite et mémoire procédurale. Au bout du compte, sur ce plan de l'organisation, mémoire humaine et mémoire informatique ne paraissent entretenir que de lointains rapports.

Sur le plan des procédures de l'accès, il apparait que le cerveau dispose de puissants mécanismes associatifs lui permettant de récupérer un souvenir complexe à partir de la perception d'un stimulus, par des chemins dont les parcours neuronaux et les mécanismes détaillés sont encore très largement l'objet de recherches. En informatique, il existe bien des mémoires associatives « hardware » - les CAM, content adressable memory - utilisables en complément des RAM dans certaines applications spécialisées, et mobilisant des procédures de recherche massivement parallèles. Mais quelle que soit l'éventuelle correspondance formelle entre ces mécanismes (ce qui nous paraît loin d'être clair) leur coût en matière de dépense d'énergie est sans commune mesure. N'oublions pas les performances du cerveau humain, qui fonctionne avec seulement une trentaine de watts pour alimenter ses 100 milliards de neurones et leurs interconnections, qui comme déjà dit sont de l'ordre de 10.000 synapses pour chaque neurone du cortex, et plus de 100.000 pour les neurones du cervelet. Et pour conserver un volume d'information qu'on a pu estimer de

l'ordre du petaoctet[6]. L'informatique joue – sur le plan des volumes d'information conservée - dans une cour nettement plus dispendieuse en matière de consommation d'énergie.

Mais ce qui distingue peut-être davantage encore la mémoire humaine de la mémoire informatique, c'est ce que souligne Alberto Oliverio dans la dernière section du premier chapitre : la mémoire humaine ne doit pas être vue comme un archivage, mais comme un ensemble d'informations en perpétuel remaniement ou reconstruction, en quelque sorte un ensemble « vivant » d'éléments en interaction.

**Autour de la localisation des activités mathématiques**

Les résultats du travail de Marie Amalric, sous la direction de Stanislas Dehaene, présentés dans le chapitre 6, ont pu surprendre ; ils semblent en effet faire une nette distinction, sur le plan de la localisation cérébrale, entre les activités de compréhension d'énoncés mathématiques et ceux d'énoncés non-mathématiques. La surprise tire son origine du fait que les mathématiques sont un langage, comme l'ont souligné des mathématiciens tels A. Grothendieck ou L. Lafforgue. Un langage, savoir un ensemble conventionnel de signes et de méthodes d'assemblage de ces signes visant à exprimer des concepts, certes plus abstraits que ceux sous-jacents aux énoncés ordinaires. Constatons aussi que ce langage mathématique est universel, en ce sens que les signes comme les concepts en sont enseignables et compréhensibles dans toutes les cultures humaines. Pourquoi les localisations des activités cérébrales liées à ces deux langages - le mathématique et le non-mathématiques - seraient-elles séparées ? Et quelle est l'origine de cette séparation ? syntaxique, sémantique ? Reconnaissons à Marie Amalric de ne pas esquiver le problème, en répondant que ses travaux comme ceux d'autres chercheurs semblent confirmer le fait que concepts mathématiques et concepts non-mathématiques appartiennent bien au sein du cerveau, à deux réseaux *sémantiques* différents.

Une autre question a émergé : l'existence d'une différence, qui pourrait être importante du point de vue de leurs manifestations neuronales, entre l'activité de compréhension (d'énoncés mathématiques) et l'activité de recherche et de création – par exemple, l'activité qui se déploie chez un mathématicien lorsqu'il cherche à démontrer une conjecture, avec les différentes phases de travail qui ont pu être décrites. On pourrait penser à priori que de telles activités pourraient mobiliser des ressources cérébrales plus diversifiées, des réseaux neuronaux plus étendus que la simple compréhension de concepts déjà acquis. La question peut-être élargie à celle de la localisation des activités d'abstraction – au sens actif de ce terme impliquant une généralisation, l'invention d'un concept englobant. Mais de telles

[6] Soit un million de Gigaoctets. Estimation fournie en 2016 par une équipe de neurologues conduite par Terry Sejnowski, Salk Institute for Biological Studies, Université de la Jolla, Californie. Pour apprécier l'importance de ce volume de données détenues *individuellement*, il suffit de le comparer à la capacité de stockage totale (actuelle) des quatre géants du Web, qui ne serait « que » 1200 fois plus élevée. Bien entendu une telle estimation, faite à partir d'extrapolations et visant à comparer des structures aussi différentes que l'ordinateur et le cerveau, doit être prise avec précaution.

questions étaient hors du champ des investigations exposées par Marie Amalric ; il s'agit de problèmes ouverts.

**Intelligence Artificielle ou Informatique Augmentée ?**

A la lecture des deux chapitres consacrés à l'histoire de l'Intelligence Artificielle dans la seconde partie, on n'aura pas manqué de constater que nombre d'idées sont apparues très tôt. Que l'on songe par exemple à l'idée de réseau neuronal artificiel, née (1943…) avant même que n'apparaissent les « vrais » ordinateurs, ou encore le principe de l'algorithme de rétro-propagation du gradient, dans les années 1980. En quelque sorte le progrès *conceptuel* n'est pas si rapide qu'on pourrait le penser au vu de l'effervescence actuelle. Pour comprendre la rupture constituée par l'émergence massive de l'IA, il faut bien prendre la mesure de ce qui s'est passé. Nous rappellerons ici un seul point, celui de l'explosion des *capacités hardware.*

Entre leurs années de naissance dans la décennie 1950 et maintenant, la puissance de ces « machines » stupéfiantes a été multipliée en gros par 1 million, tandis que leur taille/poids passait de quelques dizaines de tonnes, à quelques centaines de grammes. Ce qui était inimaginable par le simple manque de puissance de traitement, 20.000 opérations par seconde pour les premières machines, et une fiabilité de quelques heures avant la panne inéluctable, est devenu trivial. La fiabilité du matériel est telle que la moyenne de durée de vie sans panne est de quelques années, durée que l'on sait d'ailleurs prolonger par redondance bien dosée de matériels ad hoc aux bons endroits du système informatique géant que constituent les *Data Centers* qui font la fortune des GAFAM.

Cette explosion des performances n'a pas pour cause un changement dans l'architecture des machines. Cette dernière n'a pas vraiment bougé depuis que von Neumann dans un mémoire Top Secret de 1948 leur a donné une forme qui, en gros, n'a jamais varié : von Neumann l'appelait le *Logical Design*, en référence aux travaux de Turing. Par contre, la technologie n'a plus rien à voir. Entre temps en effet les transistors ont été inventés, et plus particulièrement ceux fabriqués en technologie CMOS, dans les années 1980, ce qui a permis de les ramener progressivement à des tailles nanométriques, consommant très peu d'énergie. Aujourd'hui nous sommes capables d'intégrer de façon courante 8 à 10 millions de transistors dans 1 $mm^2$ de silicium, et nous avons des « puces » de quelques $cm^2$ qui en contiennent 4 à 5 milliards. On aboutit ainsi à des performances proprement « époustouflantes » comme celles de certaines machines de la société NVIDIA, qui peuvent délivrer une puissance limite de cent mille milliards d'opérations par seconde, en étant optimisée justement pour les opérations mises en œuvre dans les réseaux neuronaux. S'il y a « révolution », n'est-ce pas d'abord une révolution technique avant d'être une révolution conceptuelle et l'Intelligence Artificielle n'est-elle pas plutôt une « Informatique Augmentée » ?

**Et si une machine semblait nous comprendre parfaitement ?**

Le Chapitre 8, sous la signature d'Antoine Bordes traite des progrès dans la compréhension du langage naturel par les machines. On pressent, malgré les difficultés qu'Antoine Bordes y expose, que l'on atteindra peut-être dans ce domaine de très grandes performances. Quelles conclusions pourraient-on alors en tirer ? Imaginons donc une machine qui maitriserait complètement l'utilisation de ce langage, si bien qu'elle serait capable de s'exprimer et de comprendre comme un humain. Elle le ferait en s'appuyant sur son expérience spécifique, et peut-être sa « personnalité » de machine, par exemple en étant incapable de duplicité, de se faire passer pour ce qu'elle n'est pas. Il se pourrait donc qu'elle ne parvienne pas à passer avec succès le test de Turing, qu'elle ne puisse pas se faire passer pour un humain. Cependant, nous pourrions échanger avec elle de manière parfaitement naturelle, comme avec un humain qui aurait une expérience et une personnalité très différente de la nôtre, ou avec une sorte d' « alien » intelligent : est-ce que l'on n'aurait pas envie de considérer qu'une telle machine est consciente ? Et si oui, peut-on se demander quelle est la relation entre conscience et maîtrise du langage naturel ? Est-ce que l'analyse de cette relation ne pourrait pas apporter un éclairage sur la nature de ces deux notions de conscience et de langage naturel ?

Par ailleurs, a contrario, si l'on postule, comme beaucoup à l'image de John Searle, qu'il ne peut y avoir de conscience hors du vivant, et que l'on admet qu'il y a une relation étroite entre conscience et maîtrise du langage naturel (au sens complet, comme indiqué ci-dessus, de la capacité d'échanger de manière naturelle - comme avec un humain éventuellement très différent en expérience et personnalité), doit-on en déduire qu'il est illusoire d'espérer amener une machine à maîtriser le langage naturel, et que celle-ci ne pourra jamais que s'en approcher, et seulement dans des contextes spécifiques ?

Sur ces questions autour des progrès susceptibles d'être accomplis par des machines dans la compréhension du langage humain, un argument a pu être évoqué, conservant une « hiérarchie » nette entre l'être humain et les algorithmes : c'est le premier, « l'ingénieur » qui crée la machine, qui a réfléchi à la façon d'articuler les éléments syntaxiques, sémantiques, symboliques ; ce n'est pas la machine. Mais pourquoi, vu les progrès actuels du « deep learning » ne pourrait-on pas doter les machines de capacités leur permettant d'apprendre le langage humain de *façon autonome*, syntaxe et sémantique incluses ; à la manière de ce qui se passe chez un enfant ? Telle est en tout cas la direction dans laquelle s'est engagé Luc Steels et dont il va maintenant être question.

**Evolution du langage et évolution des concepts**

Dans le chapitre 9, Luc Steels présente sa théorie sur l'origine et l'évolution du langage, dans les aspects à la fois lexicaux, morphologiques, syntaxiques, d'une part, et sémantique d'autre part. Il considère bien en effet les éléments de sens, à coté et peut-être indépendamment des expressions qui les signifient, comme des unités soumises aux processus de mutation/transmission/sélection à la base de sa théorie. Il détaille des expérimentations sur cette question du sens - émergence de catégories de couleurs, de

catégories de gestes corporels ou d'objets, et simultanément des vocabulaires associés à ces catégories, ou encore émergence de la distinction entre « un » et « plusieurs » simultanément à des modifications morphologiques. Il rattache fonctionnellement ces modifications morphologiques, puis l'apparition de structures grammaticales complexes, à la réduction des possibilités de sens qui seraient ouvertes par une liste « brute », non structurée, de mots sans marqueurs ; donc à la recherche d'une « économie » de moyens dans la transmission du sens.

Mais peut-être, aimerait-on le voir traiter plus largement de l'évolution des concepts en général ; quelles formes exactes prennent les processus du type de ceux qu'il évoque – transmission/sélection/mutation – dans la détermination de telles dynamiques ? Plus largement, quels mécanismes collectifs et neuronaux contrôlent la « contagion des idées » ? Est-ce un signe que ces questions n'ont pas encore été suffisamment explorées, qu'il y aurait un saut paradigmatique à effectuer pour avancer ? Est-ce que les techniques numériques de type neuronal permettront ce saut de paradigme ?

Nous nous sommes interrogés également sur la technique consistant à conduire ces expérimentations en mobilisant des robots. Certes, cette mobilisation répond à une hypothèse de fond, à laquelle beaucoup souscrivent : celle d'une *nécessaire incarnation physique* - dans le monde physique dans toute sa complexité - de l'intelligence et de la conscience, pour atteindre le niveau constaté dans le monde du vivant. Mais sur certaines questions, au moins à titre d'exploration préalable, pourquoi ne pas conduire davantage les expérimentations en simulant perceptions, actions et échanges. Pour permettre à moindre coût plus de flexibilité dans la modulation des scénarios à mettre en oeuvre, ouvrir potentiellement sur des scénarios plus complexes, être plus efficace en termes de rapidité de test de grands nombres de cas…

**Le langage peut-il émerger du hasard, le langage évolue-t-il vraiment ?**
Une des questions que soulève la lecture du chapitre de Luc Steels est celle de l'émergence de structures complexes à partir d'une combinatoire d'éléments plus simples disponibles initialement, sous l'effet du hasard, puis de processus de sélection et de transmission. Il est vrai que les mécanismes à l'origine de structures complexes dans le domaine du langage humain, et au-delà dans le domaine du vivant, sont loin d'être entièrement compris ; et l'appel à « la puissance créative du hasard » peut paraître problématique à certains, au travers de calculs de probabilité. Les arguments avancés pour défendre l'idée de cette puissance du hasard dans l'évolution biologique sont que la vitesse des processus de reproduction, la taille des populations en jeu, le temps à l'échelle de millions d'années, et les contraintes physiques internes et externes font que l'ensemble des possibles viables à une certaine époque est exploré. Ces arguments, en admettant qu'ils soient valables en biologie, sont-ils transposables en linguistique ?

Il nous semble utile, pour terminer sur ces thèmes, d'évoquer l'opposition entre les thèses de Steels et celles de Chomsky (le langage n'évolue pas) dont Luc Steels a parlé lui-même

dans sa conférence. Le livre de Robert Berwick et Noam Chomsky, *Why Only us ?* paru en 2016, auquel Luc Steels fait référence, soutient en effet deux points essentiels, liés l'un à l'autre.

Le premier point est l'affirmation que l'apparition du langage humain fait partie des grandes ruptures, peu nombreuses, pour lesquelles une explication par un effet *progressif* de la sélection naturelle paraît problématique. En appui de cette thèse, ils évoquent un ouvrage bien connu par ailleurs, *The major transitions in evolution* de J. Maynard-Smith. Et de citer par exemple la rupture constituée par la transition Procaryotes->Eucaryotes au tout début de la vie cellulaire, il y a plus d'un milliard d'année. A l'opposé des schémas de mise en place progressive, Berwick et Chomsky semblent défendre l'idée d'une mutation majeure et unique, ayant introduit les capacités langagières humaines de façon quasi complète, mettant en place de façon définitive la structure profonde très spécifique de la pensée et du langage humain. Si bien qu'un nouveau-né de l'âge de pierre et transplanté à l'époque actuelle ne se distinguerait pas, du point de vue de ses capacités cognitives et linguistiques, du reste des enfants. Il y a certes une grande diversité des langues, et l'on ne peut nier qu'elles changent au cours du temps, mais elles reposent toutes sur cette structure profonde sous-jacente, si bien que cette diversité et ces changements apparaissent inessentiels.

Le second point concerne la nature de cette structure profonde, dont Berwick et Chomsky supposent l'existence, comme socle de l'ordre linéaire du langage se manifestant dans ce qu'ils appellent « l' externalisation », c'est-à-dire la communication. Tout d'abord, avant d'être celle de la communication, cette structure profonde est bien celle d'un « langage de la pensée » et donc une structure particulière de cette dernière. C'est une structure hiérarchique, et sans limites apparentes concernant la profondeur de cette hiérarchie. Les constructions qu'elle permet sont assemblées selon un principe combinatoire et récursif extrêmement simple, spécifiable mathématiquement, et pour ainsi dire « parfait », « computationnellement optimal », c'est un tel principe que la mutation originelle a pu mettre en place, du fait même de cette simplicité. C'est ce principe qui, une fois installé, permet à tout nouvel être humain exposé linguistiquement *au bon moment*, d'apprendre à penser et à communiquer.

On le voit bien, les thèses respectives de Luc Steels et de Berwick/Chomsky paraissent aux antipodes l'une de l'autre. D'une part avec le premier, une émergence progressive et concomitante du vocabulaire et du sens, basée sur la sélection d'avantages fonctionnels en matière de coopération sociale, avec en arrière-plan génétique, selon toute probabilité, une sorte de bricolage, un recyclage de capacités déjà présentes dans les espèces non humaines. D'autre part, avec les seconds, l'apparition soudaine d'une singularité dans l'outillage de la pensée, avec d'emblée la capacité de combiner à l'infini des concepts et de traduire ces

combinaisons en phrases. Le débat est vif entre linguistes[7], et dépasse bien sûr le cadre de cet ouvrage. Mais il nous ramène à des questions récurrentes : devant la diversité des langages humains, leur syntaxe, leur traduction scripturale, et a contrario l'uniformité supposée d'un « langage de la pensée » ou d'une certaine capacité d'abstraire - l'abstraction réfléchissante disait Piaget - qu'est-ce qui est génétique exactement chez l'être humain ?

**Pour conclure**

Il est clair que la recherche sur la Conscience, à travers les pistes suivies en neurosciences, en informatique et dans la réflexion philosophique, est un front scientifique majeur de notre temps. Il est trop tôt encore pour en percevoir l'aboutissement, si tant est qu'il puisse y en avoir un. On peut se demander par contre si l'Humanité saura accepter et intégrer ces résultats dans une conception du monde et d'elle-même commune. Dans cette perspective mettons en exergue les récentes réflexions de Pascal Picq, paléoanthropologue au Collège de France qui donne une nouvelle perspective à la thématique sur les « intelligences ». Dans son dernier ouvrage « L'intelligence artificielle et les chimpanzés du futur, pour une anthropologie des intelligences », il dit : « L'homme est à l'aube d'une nouvelle étape de son évolution et il ne la franchira pas seul. Les animaux qui l'entourent et les intelligences artificielles y participeront également et l'objectif est donc de réussir le mariage entre les humains et ces différentes formes d'intelligence. Notre évolution de demain ne se fera pas entre des intelligences animales méprisées et des intelligences artificielles vénérées. Il est impératif de reconnaître et de comprendre les autres formes d'intelligences, leurs spécificités, leurs fonctionnements et de penser en termes d'intelligences augmentées. L'objectif est de co-évoluer ensemble en tenant compte des complémentarités de chacune. »

Pour le comité de lecture[8]

[7] Le lecteur pourra lire par exemple, dans une perspective critique, la review de l'ouvrage de Berwick et Chomsky, effectuée par Ljiljana Progovac dans la revue Language Volume 92, Number 4(2016)

[8] Gilbert Belaubre, Eric Chenin, Pierre Nabet, Alberto Oliverio, Jacques Printz, Jean Schmets, Jean-Pierre Treuil

# Remerciements

Nous tenons en premier lieu à remercier les auteurs qui nous ont adressé un texte original : Mesdames et Messieurs Alberto Oliverio, Francis Eustache, Armelle Viard, Laure Zago, Marie Amalric, Jean-Paul Haton, Ernesto Di Mauro, Franck Cosson, Gérard de Boisboissel, Laurence Devillers. Nous remercions aussi les conférenciers et conférencières qui ont relu, corrigé, amendé et approuvé après corrections et compléments les transcriptions qui ont été faites de leurs présentations par les membres de l'AEIS : Claire Sergent, Jérôme Sackur, Jean-Gabriel Ganascia, Antoine Bordes, Luc Steels, Raja Chatila, ce dernier également pour l'animation de la Table Ronde et sa participation à la rédaction de sa synthèse.

Nous remercions en outre les rédacteurs de ces transcriptions et autres textes inclus dans notre ouvrage et membres du Comité de lecture de l'AEIS : Gilbert Belaubre, Eric Chenin, Pierre Nabet, Jacques Printz, Jean-Pierre Treuil. Nous remercions Jean Schmets d'avoir bien voulu faire une première relecture systématique de l'ensemble des textes.
Tous les participants nous ont fourni des articles de haut niveau scientifique, et ont collaboré à la mise au point des iconographies et au respect des normes de mise en page de notre Éditeur EDP-Sciences.

Une mention particulière doit être faite à Jean-Pierre Treuil qui a bien voulu animer le comité de lecture de l'AEIS, chargé de mener à bien ce projet d'ouvrage. Il a su établir des contacts fructueux avec nos différents contributeurs au niveau des échanges et des contenus, a contribué également aux textes issus de transcriptions et aux présentations des différentes parties de cet ouvrage. Qu'il soit vivement remercié pour toutes ces actions.

Il participe également à la mise en ligne sur notre site de documents multimédias de nos conférences et séminaires.

À chacune de nos séances mensuelles, nous recevons les intervenants pressentis pour un colloque sur le thème de notre projet bisannuel. Les séminaires correspondants sont annoncés sur le serveur CORDIS-Europa de la Commission Européenne par l'intermédiaire de Jean Schmets ; notre secrétaire générale Irène Herpe-Litwin a rédigé les comptes-rendus des nombreuses réunions du Comité de lecture de l'AEIS, institué pour la réalisation de cet ouvrage, qu'ils en soient remerciés. Nos remerciements vont également à notre Webmestre Alain Cordier qui est en charge de la mise à niveau et de la maintenance du site de l'AEIS ainsi que de la gestion de nos colloques sur la plateforme du CNRS/MESR. Marie-Françoise Passini contribue de la meilleure façon à la conception de la maquette de l'Opuscule destiné aux participants à nos colloque, qu'elle en soit remerciée.

Pour la mise au point technique de l'intérieur de l'ouvrage (maquette) nous avons bénéficié des outils professionnels, des compétences et de la bienveillante coopération de Marie-Françoise et Jean Passini. Qu'ils soient vivement remerciés pour ce travail et pour leur professionnalisme. Nos remerciements vont aussi au professionnalisme et au travail de Gilbert Brami qui s'est à nouveau chargé de la couverture de l'ouvrage. Nous bénéficions aussi des compétences et de la bienveillante attention de France Citrini Responsable LIVRES et de sa collègue Sophie Hosotte chez notre éditeur EDP-Sciences.

Nos remerciements vont aussi aux organismes et aux personnalités qui, en amont de notre projet, ont permis les contacts, les conférences préalables, l'organisation d'un colloque international et enfin la structuration de cet ouvrage. L'association des Anciens Élèves de l'École Polytechnique, puis l'Institut Henri Poincaré nous ont accueillis pour les séances mensuelles de l'AEIS. Le colloque lui-même s'est tenu à l'Institut Henri Poincaré de Paris, Haut lieu de la science mathématique et physique française, dont nous tenons à remercier la Directrice Sylvie Benzoni, le Directeur-adjoint, Rémi Monasson et Florence Lajoinie, responsable de l'organisation technique de l'amphithéâtre Hermite.

Le Président de l'AEIS
Victor Mastrangelo

# Présentation de

## L'ACADEMIE EUROPÉENNE

## INTERDISCIPLINAIRE

## DES SCIENCES

**L'académie Européenne Interdisciplinaire des Sciences, AEIS, société savante (loi 1901), a pour but la recherche, la diffusion et la formation dans tous les domaines de la science. L'AEIS est répertoriée parmi les autres institutions de la capitale sur le site http** : www.paris.fr

**L'académie se propose** :

- De rassembler et de faire étudier les différentes recherches et pensées scientifiques dans un cadre interdisciplinaire ;
- D'établir entre les scientifiques un langage commun nécessaire pour une mutuelle compréhension ;
- De faire connaître les plus récentes découvertes, inventions ou réalisations des domaines de la connaissance ;

De participer à l'élargissement de la pensée, en particulier sur des sujets frontières des différentes disciplines, pour atténuer la rigueur des délimitations souvent artificielles.

**Ses membres sont issus du monde académique de la recherche et de l'industrie, les grandes disciplines scientifiques sont représentées** : Mathématiques, Physique, Chimie, Biologie, Biophysique, Biochimie, Médecine, Informatique, Sciences cognitives, Neurosciences cognitives, Sciences sociales, Sciences de la Terre, Théorie des systèmes complexes, Philosophie des sciences….

**L'Académie** :

- Tient des séances régulières à Paris, Nancy, Nice, Reims ;
- Édite un bulletin mensuel ;
- Possède des ramifications à Bruxelles, Liège, Luxembourg, Rome ; Athènes ;
- Organise des colloques interdisciplinaires sur des thèmes scientifiques et de société.

Principaux thèmes des colloques internationaux :

- **2002 « Biologie et conscience »** avec la participation du prix Nobel Gérald M. Edelman, CNAM-Paris ;
- **2004 « Fractales en progrès »** pour les 80 ans de Benoît Mandelbrot, Université Paris-Descartes ;
- **2004 « Physique et conscience »,** clôturant l'Année Mondiale de la Physique avec la participation du prix Nobel Gilles de Gennes, Paris, Ministère de la Recherche ;
- **2008 « Émergence : de la fascination à la compréhension »,** Université Paris-Diderot
- **2009 « Perspectives des approches expérimentales et théoriques de l'évolution »** à l'occasion de l'année Darwin ; président du colloque Denis Noble membre de la « Royal Society », université Paris-Diderot ;
- **2011 « Théories et modèles en sciences sociales »,** Président du colloque Prof. Raymond Boudon, membre de l'Institut, Université Paris-Diderot ;
- **2014 « Formation des systèmes stellaires et planétaires -Conditions d'apparition de la vie »**, Institut Henri Poincaré-Paris ;
- **2016 « Ondes, matière et Univers / Relativité générale, physique quantique et applications »**, Présidente du colloque Prof. Françoise Balibar avec la participation de Claude Cohen-Tannoudji et Serge Haroche, Institut Henri Poincaré-Paris ;
- **2018 « Les Signatures neurobiologiques de la conscience/Neurobiologie fonctionnelle, phénomènes de conscience, cognition, automates « intelligents », éthique »**, Institut Henri Poincaré-Paris.

## Recrutements de membres

L'ACADEMIE EUROPEENNE INTERDISCIPLINAIRE DES SCIENCES (AEIS) reçoit les candidatures de personnes ayant une forte culture scientifique. Cette culture peut être très générale ou fortement spécialisée, mais tous ses membres doivent être aptes à aborder des sujets interdisciplinaires qui sont à la base de ses projets. Les candidats/candidates doivent donc présenter un curriculum vitae qui expose leur niveau scientifique et leurs activités de recherche et de publications. Les productions intellectuelles, articles scientifiques dans des revues à comité de lecture, ouvrages et brevets sont les témoignages les plus pertinents des capacités des candidats/candidates. Les éléments du CV à fournir sont détaillés dans l'onglet « membres » du site internet de l'AEIS dont l'adresse est donnée ci-après. Le CV doit être accompagné d'une lettre de motivation par laquelle le candidat ou la candidate exprime ce qu'il ou qu'elle peut attendre de l'AEIS et ce qu'il ou qu'elle pense pouvoir apporter en termes de compétences, de participation aux programmes de travail, et de gestion propre de l'Académie. Il existe deux modes de recrutement : membres titulaires et membres correspondants (Cf. détails sur le site internet de l'AEIS) Les dossiers de candidature doivent être adressés au secrétariat général de l'AEIS. Ils sont ensuite présentés à l'ensemble des membres titulaires qui formulent leurs avis, et, la décision est prise par le Bureau selon les modalités prévues par le Règlement intérieur.

Pour accéder aux modalités de candidature, aller sur le site de l'AEIS

http://science-inter.com/

Ou écrire à l'adresse électronique suivante : eric.chenin@science-inter.com ou bien à iherpelitwin@gmail.com qui transmettra le dossier de candidature

::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::