Académie des technologies

Vers une technologie de la conscience ?

Communication présentée à l'Académie

en juin 2012

Imprimé en France
ISBN : 978-2-7598-1009-3

UNE ACADÉMIE DES TECHNOLOGIES POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

L'Académie des technologies, créée en 2000, est, depuis 2006, un Établissement public administratif auprès du ministère de l'Enseignement supérieur et de la recherche.

L'Académie des technologies a pour vocation d'être l'institution de référence et l'intermédiaire privilégié entre le monde de la recherche, les pouvoirs publics et les acteurs socio-économiques sur les questions technologiques. Les technologies sont abordées dans une approche transversale et prospective, prenant en compte les risques, l'impact sur l'environnement et la santé, les aspects économiques et sociétaux.

L'Académie des technologies rassemble à ce jour 275 académiciens élus, couvrant de très larges domaines de compétences : chercheurs, entrepreneurs et industriels, économistes et avocats, philosophes des sciences et des techniques, urbanistes et architectes, physiciens, chimistes, ingénieurs, astronautes, médecins et chirurgiens, spécialistes des technologies de l'information et de la communication, agronomes et spécialistes du génie agroalimentaire, etc. Cette grande variété de compétences en fait un lieu d'expertise et de réflexion interdisciplinaire sur les bouleversements technologiques et les grands défis auxquels la société se trouve confrontée.

Elle contribue ainsi à bâtir une Europe de la connaissance et de l'innovation compétitive, condition de la création de richesse et d'emplois.

Une expertise collective, ancrée au sein même des choix politiques en matière de technologies innovantes

L'Académie des technologies mène ses travaux en toute indépendance, en associant à ses réflexions le secteur de la production, les milieux de la recherche scientifique, le monde politique et social et les acteurs socio-économiques. Elle développe de nombreux partenariats avec d'autres académies en France et à l'étranger.

Elle participe au développement des réflexions menées au niveau international ou européen. Elle assure le Secrétariat général d'EURO-CASE, qui regroupe les académies technologiques européennes de 21 pays.

Outre les nombreuses publications depuis sa création (une soixantaine d'ouvrages), les académiciens vont à la rencontre des décideurs, des acteurs et du public, à Paris et dans les régions de France à travers des colloques, des actions thématiques régionales, *etc.* Des rencontres-débats sont organisées régulièrement avec des personnalités du monde économique, politique, industriel… Les publications font l'objet de conférences-débats où est conviée la presse. Un grand débat public est organisé une fois par an sur un sujet d'actualité (changements climatiques, économies d'énergie, principe de précaution …).

PUBLICATIONS DE L'ACADÉMIE

Les travaux de l'Académie des technologies sont l'objet de publications réparties en quatre collections[1] :

- Les rapports de l'Académie : ce sont des textes rédigés par un groupe de l'Académie dans le cadre du programme décidé par l'Académie et suivi par le Comité des travaux. Ces textes sont soumis au Comité de la qualité, votés par l'Assemblée, puis rendus publics. On trouve dans la même collection les avis de l'Académie, également votés en Assemblée, et dont le conseil académique a décidé de la publication sous forme d'ouvrage papier. Cette collection est sous couverture bleue.

[1] - Les ouvrages de l'Académie des technologies publiés entre 2008 et 2012 peuvent être commandés aux Éditions Le Manuscrit (http://www.manuscrit.com). La plupart existent tant sous forme matérielle que sous forme électronique.
- Les titres publiés à partir de janvier 2013 sont disponibles en librairie et sous forme de ebook payant sur le site de EDP sciences (http://www.edition-sciences.com). À échéance de six mois ils sont téléchargeables directement et gratuitement sur le site de l'Académie.
- Les publications plus anciennes n'ont pas fait l'objet d'une diffusion commerciale, elles sont consultables et téléchargeables sur le site public de l'Académie www.academie-technologies.fr, dans la rubrique « Publications ». De plus, l'Académie dispose encore pour certaines d'entre elles d'exemplaires imprimés.

- Les communications à l'Académie, rédigées par un ou plusieurs Académiciens. Elles sont soumises au Comité de la qualité et débattues en Assemblée. Non soumises à son vote elles n'engagent pas l'Académie. Elles sont rendues publiques comme telles, sur décision du Conseil académique. Cette collection est publiée sous couverture rouge.
- Les « Dix questions à… et dix questions sur… » : un auteur spécialiste d'un sujet est sélectionné par le Comité des travaux et propose dix à quinze pages au maximum, sous forme de réponses à dix questions qu'il a élaborées lui-même ou après discussion avec un journaliste de ses connaissances ou des collègues (Dix questions à…). Ce type de document peut aussi être rédigé sur un thème défini par l'Académie par un académicien ou un groupe d'académiciens (Dix questions sur…). Dans les deux cas ces textes sont écrits de manière à être accessibles à un public non-spécialisé. Cette collection est publiée sous une couverture verte.
- Les grandes aventures technologiques françaises : témoignages d'un membre de l'Académie ayant contribué à l'histoire industrielle. Cette collection est publiée sous couverture jaune.
- Par ailleurs, concernant les Avis, l'Académie des technologies est amenée, comme cela est spécifié dans ses missions, à remettre des Avis suite à la saisine d'une collectivité publique ou par auto saisine en réaction à l'actualité. Lorsqu'un avis ne fait pas l'objet d'une publication matérielle, il est, après accord de l'organisme demandeur, mis en ligne sur le site public de l'Académie.
- Enfin, l'Académie participe aussi à des co-études avec ses partenaires, notamment les Académies des sciences, de médecine, d'agriculture, de pharmacie…

Tous les documents émis par l'Académie des technologies depuis sa création sont répertoriés sur le site www.academie-technologies.fr. La plupart sont peuvent être consultés sur ce site et ils sont pour beaucoup téléchargeables.

Dans la liste ci-dessous, les documents édités sous forme d'ouvrage imprimé commercialisé sont signalés par une astérisque. Les publications les plus récentes sont signalées sur le site des éditions. Toutes les publications existent aussi sous forme électronique au format pdf et pour les plus récentes au format ebook.

AVIS DE L'ACADÉMIE

1. Brevetabilité des inventions mises en œuvre par ordinateurs : avis au Premier ministre – juin 2001
2. Note complémentaire au premier avis transmis au Premier ministre – juin 2003
3. Quelles méthodologies doit-on mettre en œuvre pour définir les grandes orientations de la recherche française et comment, à partir de cette approche, donner plus de lisibilité à la politique engagée ? – décembre 2003
4. Les indicateurs pertinents permettant le suivi des flux de jeunes scientifiques et ingénieurs français vers d'autres pays, notamment les États-Unis – décembre 2003
5. Recenser les paramètres susceptibles de constituer une grille d'analyse commune à toutes les questions concernant l'énergie – décembre 2003
6. Commentaires sur le Livre Blanc sur les énergies – janvier 2004
7. Premières remarques à propos de la réflexion et de la concertation sur l'avenir de la recherche lancée par le ministère de la Recherche – mars 2004
8. Le système français de recherche et d'innovation (SFRI). Vue d'ensemble du système français de recherche et d'innovation – juin 2004
   - Annexe 1 – La gouvernance du système de recherche
   - Annexe 2 – Causes structurelles du déficit d'innovation technologique. Constat, analyse et proposition.
9. L'enseignement des technologies de l'école primaire aux lycées – septembre 2004
10. L'évaluation de la recherche – mars 2007
11. L'enseignement supérieur – juillet 2007
12. La structuration du CNRS – novembre 2008
13. La réforme du recrutement et de la formation des enseignants des lycées professionnels – Recommandation de l'Académie des technologies – avril 2009
14. La stratégie nationale de recherche et l'innovation (SNRI) – octobre 2009
15. Les crédits carbone – novembre 2009
16. Réduire l'exposition aux ondes des antennes-relais n'est pas justifié scientifiquement : mise au point de l'Académie nationale de médecine, de l'Académie des sciences et de l'Académie des technologies – décembre 2009
17. Les biotechnologies demain – juillet 2010
18. Les bons usages du Principe de précaution – octobre 2010
19. La validation de l'Acquis de l'expérience (VAE) – janvier 2012
20. Mise en œuvre de la directive des quotas pour la période 2013–2020 – mars 2011

21. Le devenir des IUT – mai 2011
22. Le financement des start-up de biotechnologies pharmaceutiques – septembre 2011
23. Recherche et innovation : Quelles politiques pour les régions ? – juillet 2012
24. La biologie de synthèse et les biotechnologies industrielles (blanches) – octobre 2012
25. Les produits chimiques dans notre environnement quotidien – octobre 2012
26. L'introduction de la technologie au lycée dans les filières d'enseignement général – décembre 2012
27. Évaluation de la recherche technologique publique – février 2013
28. L'usage de la langue anglaise dans l'enseignement supérieur – mai 2013

RAPPORTS DE L'ACADÉMIE

1. Analyse des cycles de vie – octobre 2002
2. Le gaz naturel – octobre 2002
3. Les nanotechnologies : enjeux et conditions de réussite d'un projet national de recherche – décembre 2002
4. Les progrès technologiques au sein des industries alimentaires – Impact sur la qualité des aliments / La filière lait – mai 2003
5. *Métrologie du futur – mai 2004
6. *Interaction Homme-Machine – octobre 2004
7. *Enquête sur les frontières de la simulation numérique – juin 2005
8. Progrès technologiques au sein des industries alimentaires – la filière laitière, rapport en commun avec l'Académie d'agriculture de France – 2006
9. *Le patient, les technologies et la médecine ambulatoire – avril 2008
10. *Le transport de marchandises – janvier 2009 (version anglaise au numéro 15)
11. *Efficacité énergétique dans l'habitat et les bâtiments – avril 2009 (version anglaise au numéro 17)
12. *L'enseignement professionnel – décembre 2010
13. *Vecteurs d'énergie – décembre 2011 (version anglaise au numéro 16)
14. *Le véhicule du futur – septembre 2012 (publication juin 2013)
15. *Freight systems (version anglaise du rapport 10 le transport de marchandises) – novembre 2012
16. *Energy vectors – novembre 2012 (vesion anglaise du numéro 13)

17. *Energy Efficiency in Buildings and Housing – novembre 2012 (version anglaise du numéro 11)
18. * Première contribution de l'Académie des technologies au débat national sur l'énergie / First contribution oF the national academy oF technologies oF France to the national debate on the Future oF energies supply – ouvrage bilingue, juillet 2013
19. *Les grands systèmes socio-techniques / Large Socio-TechnicaL Systems – ouvrage bilingue, juillet 2013

COMMUNICATIONS DE L'ACADÉMIE

1. *Prospective sur l'énergie au XXIe siècle, synthèse de la Commission énergie et environnement – avril 2004, MàJ décembre 2004
2. Rapports sectoriels dans le cadre de la Commission énergie et environnement et changement climatique :
3. Les émissions humaines – août 2003
   - Économies d'énergie dans l'habitat – août 2003
   - Le changement climatique et la lutte contre l'effet de serre – août 2003
   - Le cycle du carbone – août 2003
   - Charbon, quel avenir ? – décembre 2003
   - Gaz naturel – décembre 2003
   - Facteur 4 sur les émissions de CO_2 – mars 2005
   - Les filières nucléaires aujourd'hui et demain – mars 2005
   - Énergie hydraulique et énergie éolienne – novembre 2005
   - La séquestration du CO_2 – décembre 2005
   - Que penser de l'épuisement des réserves pétrolières et de l'évolution du prix du brut ? – mars 2007
4. Pour une politique audacieuse de recherche, développement et d'innovation de la France – juillet 2004
5. *Les TIC : un enjeu économique et sociétal pour la France – juillet 2005
6. *Perspectives de l'énergie solaire en France – juillet 2008
7. *Des relations entre entreprise et recherche extérieure – octobre 2008
8. *Prospective sur l'énergie au XXIe siècle, synthèse de la Commission énergie et environnement, version française et anglaise, réactualisation – octobre 2008
9. *L'énergie hydro-électrique et l'énergie éolienne – janvier 2009
10. *Les Biocarburants – février 2010
11. *PME, technologies et développement – mars 2010.

12. *Biotechnologies et environnement – avril 2010
13. *Des bons usages du Principe de précaution – février 2011
14. L'exploration des réserves françaises d'hydrocarbures de roche mère (gaz et huile de schiste) – mai 2011
15. *Les ruptures technologiques et l'innovation – février 2012
16. *Risques liés aux nanoparticules manufacturées – février 2012
17. *Alimentation, innovation et consommateurs – juin 2012
18. Vers une technologie de la conscience – juin 2012
19. Profiter des ruptures technologiques pour gagner en compétitivité et en capacité d'innovation – juin 2012
20. Les produits chimiques au quotidien – novembre 2012
21. Profiter des ruptures technologiques pour gagner en compétitivité et en capacité d'innovation – novembre 2012
22. Dynamiser l'innovation par la recherche et la technologie – novembre 2012
23. La technologie, école d'intelligence innovante. Pour une introduction au lycée dans les filières de l'enseignement général – octobre 2012

DIX QUESTIONS POSÉES À…

1. *Les déchets nucléaires – 10 questions posées à Robert Guillaumont – décembre 2004
2. *L'avenir du charbon – 10 questions posées à Gilbert Ruelle – janvier 2005
3. *L'hydrogène – 10 questions posées à Jean Dhers – janvier 2005
4. *Relations entre la technologie, la croissance et l'emploi – 10 questions à Jacques Lesourne – mars 2007
5. *Stockage de l'énergie électrique – 10 questions posées à Jean Dhers – décembre 2007
6. *L'éolien, une énergie du XXIe siècle – 10 questions posées à Gilbert Ruelle – octobre 2008
7. *La robotique – 10 questions posées à Philippe Coiffet, version franco-anglaise – septembre 2009
8. *L'intelligence artificielle – 10 questions posées à Gérard Sabah – septembre 2009
9. *La validation des acquis de l'expérience – 10 questions posées à Bernard Decomps – juillet 2012

GRANDES AVENTURES TECHNOLOGIQUES

1. *Le Rilsan – par Pierre Castillon – octobre 2006
2. *Un siècle d'énergie nucléaire – par Michel Hug – novembre 2009

HORS COLLECTION

1. Libérer Prométhée – mai 2011

CO-ÉTUDES

1. Progrès technologiques au sein des industries alimentaires – La filière laitière. Rapport en commun avec l'Académie d'agriculture de France – mai 2004
2. Influence de l'évolution des technologies de production et de transformation des grains et des graines sur la qualité des aliments. Rapport commun avec l'Académie d'agriculture de France – février 2006
3. *Longévité de l'information numérique – Jean-Charles Hourcade, Franck Laloë et Erich Spitz. Rapport commun avec l'Académie des sciences – mars 2010, EDP Sciences
4. *Créativité et Innovation dans les territoires – Michel Godet, Jean-Michel Charpin, Yves Farge et François Guinot. Rapport commun du Conseil d'analyse économique, de la Datar et de l'Académie des technologies – août 2010 à la Documentation française
5. *Libérer l'innovation dans les territoires. Synthèse du Rapport commun du Conseil d'analyse économique, de la Datar et de l'Académie des technologies. Créativité et Innovation dans les territoires Édition de poche – septembre 2010 – réédition novembre 2010 à la Documentation française
6. *La Métallurgie, science et ingénierie – André Pineau et Yves Quéré. Rapport commun avec l'Académie des sciences (RST) – décembre 2010, EDP Sciences.
7. Les cahiers de la ville décarbonée en liaison avec le pôle de compétitivité Advancity
8. Le brevet, outil de l'innovation et de la valorisation – Son devenir dans une économie mondialisée – Actes du colloque organisé conjointement avec l'Académie des sciences le 5 juillet 2012 éditions Tec & doc – Lavoisier

AVERTISSEMENT

Le fait que les publications de l'Académie des technologies soient regroupées en quatre collections distinctes découle d'un classement interne des textes par les instances académiques.

Les avis et rapports de l'Académie engagent celle-ci, dès lors que les textes, préalablement visés par le comité de la qualité, ont été soumis à débat et ont donné lieu à vote favorable par l'Assemblée. Les avis constituent des réponses de l'Académie à des saisines d'autorités, notamment gouvernementales ; ils ne sont publiés qu'avec l'accord du destinataire.

Les communications à l'Académie, d'une part, font l'objet de présentations à l'Assemblée et de débats, d'autre part, sont visées par le comité de la qualité ; elles ne sont pas soumises à un vote et il revient au Conseil académique de décider de l'opportunité de leur publication. Ces textes engagent la seule responsabilité de leurs auteurs.

Les annexes des rapports et des communications, visées également par le Comité de la qualité, sont signées et engagent la seule responsabilité de leurs auteurs (souvent des experts non membres de l'Académie) qui peuvent en disposer. Elles sont le plus souvent réunies avec les corps de texte votés afin de constituer des publications complètes et à jour au moment d'être mises sous presse.

Le lecteur est invité à visiter le site Internet de l'Académie :

www.academie-technologies.fr où apparaissent non seulement les textes votés, les « communications », les « dix questions posées à… » et la série « Grandes aventures technologiques françaises » dans leur version intégrale pour ceux qui sont libres de droits mais aussi des textes qui ne font pas (ou pas encore) l'objet d'une publication dans l'une ou l'autre des quatre collections.

Les travaux de l'Académie se poursuivant sur certaines thématiques, des versions plus récentes de textes et/ou d'annexes sont régulièrement mises en ligne .

SOMMAIRE

CONSTITUTION DU GROUPE DE TRAVAIL

Membres de l'Académie :

Laurent Alexandre,
Sigrid Avrillier,
Alain Berthoz,
Danièle Blondel,
Yves Caseau,
Philippe Coiffet, président,
Jean-Pierre Dupuy,
Michel Frybourg,
Jean-Charles Hourcade,
Jean-Pierre Marec,
Jean-Claude Millet,
Paul Parnière,
Dominique Peccoud,
Marc Pélegrin,
Pierre Perrier,
Gérard Sabah, président,
Joseph Sifakis,
Erich Spitz.

Participants extérieurs à l'Académie :

Claude Mangeot (médecin thérapeute cognitif)

Experts auditionnés :

Henri Condé,
Peter F. Dominey,
Nayla Farouki,
Marc Jeannerod,
Dominique Laplane.

*La conscience est un **être***
*pour lequel il est dans son **être***
*question de son **être***
*en tant que cet **être***
*implique un **être** autre que lui.*

Jean-Paul Sartre

La seule façon d'exister pour la conscience
c'est d'avoir conscience qu'elle existe.

Jean-Paul Sartre

RÉSUMÉ

Réussir l'inclusion des processus de traitement d'information présents dans les organismes vivants sur des substrats artificiels, quelle qu'en soit la nature, constitue sûrement l'un des défis technologiques les plus en vue actuellement. L'Académie des technologies se devait de mettre ce sujet à l'ordre du jour de ses travaux.

Le groupe de travail qui a élaboré cette communication s'est ainsi posé la question générale suivante : peut-on doter des machines d'une certaine forme de conscience à l'instar de ce que manifestent l'homme ou les animaux supérieurs ?

La définition et la nature de la conscience sont de vieux problèmes qui ont produit une immense littérature, d'abord dans les domaines de la philosophie, de la métaphysique et de la théologie, puis de la psychologie et de la sociologie, enfin, avec la science et la technologie modernes, de la biologie, de la génétique, de la théorie de l'évolution, de l'automatique et de l'informatique. Malgré cet énorme réservoir de connaissances, nul ne semble avoir pu répondre de façon satisfaisante aux questions entourant la genèse et le fonctionnement de la conscience.

Le groupe de travail, après avoir tenté de revenir sur ce débat général, s'est vu contraint, par évidence, de se limiter à une problématique réduite et réaliste, mais lourde de conséquences pratiques au cas où une réponse positive pourrait être donnée à la question : *peut-on établir un modèle d'entité, implémentable sur un ordinateur, qui ferait apparaître, dans son interaction avec le monde, des propriétés qui sont attribuées à la conscience humaine par les spécialistes, alors que l'on n'a pas encore réussi à synthétiser ces propriétés sur les machines passées et présentes ?*

Bien qu'un ensemble relativement important de chercheurs ait proposé des modèles de conscience, peu sont décrits d'une manière suffisamment précise et détaillée pour permettre d'en envisager des implémentations sur machine. Toutefois, même si ces modèles ont chacun leurs originalités, ils présentent aussi des similitudes et des complémentarités qu'il convient d'exploiter.

La conscience n'est pas un objet physique bien que sa manifestation suggère la nécessité de la présence, au départ, d'un conteneur physique qui la supporte (cerveau fondé sur le carbone ou éventuel ordinateur fondé sur le silicium). La conscience est un processus dynamique évolutionnaire. Autrement dit, quand les conditions de l'apparition de son étincelle sont réunies, elle se manifeste dans un phénomène d'auto-amélioration issu d'une capacité d'auto-apprentissage.

Le groupe de travail a ainsi sélectionné pour examen trois modèles (parmi d'autres) offrant cette caractéristique. Celui d'Alain Cardon qui part de l'affirmation que « toute pensée est calculable » et propose un modèle fondé sur un système multi-agent massif dont la stabilisation correspond à un « état de pensée ». Puis, celui de Sabah avec un « carnet d'esquisses » qui traite les perceptions, accompagné d'un niveau supérieur profitant de la réflexivité et des techniques de l'intelligence artificielle[1] pour engendrer des programmes capables de représenter leurs propres actions ; les deux niveaux sont mis en relation par l'intermédiaire de mémoires et le fonctionnement de l'ensemble doit produire des « effets » de conscience. Enfin, celui de Pitrat qui développe un système de résolution de problèmes fondé sur l'amorçage et sur la déclarativité des connaissances et des méta-connaissances, et dont les aspects réflexifs lui permettent de s'améliorer continuellement.

Ainsi, chacune de ces trois approches permet de faire apparaître des fonctionnalités que l'on attribue à la conscience.

Ces trois modèles présentent un intérêt nouveau car ils sont compatibles, complémentaires, et qu'on peut les fusionner en un seul modèle plus générique, comme cela est expliqué dans le corps de cette communication. La gamme des fonctionnalités offerte par le modèle global s'élargit quantitativement aussi bien que qualitativement.

[1] Abrégée en IA dans la suite du texte.

Si ce modèle global devenait mis en œuvre opérationnellement, de nombreuses propriétés non encore atteintes avec l'IA deviendraient effectives et permettraient d'avancer qu'outre une intelligence d'un bon niveau, la machine manifeste une certaine conscience de sa propre existence, ce qui changerait beaucoup ses comportements en tant que génératrice d'actions plus ou moins libres et plus ou moins volontaires.

Seules des parties des trois modèles examinés ont donné lieu à des tests expérimentaux (le modèle de Cardon est resté « sur le papier », divers processus de celui de Sabah ont été mis en œuvre, alors que le programme de Pitrat fonctionne depuis plus de 20 ans). On reste donc dans un état de conjecture quant au résultat final présenté par la concrétisation complète et les tests du modèle global (qui exigeront un gros travail).

Le rapport montre qu'il n'y a pas d'obstacle ou de barrière technologique s'opposant à une telle concrétisation du modèle global. En conséquence, si la conjecture se révèle correcte, il faut s'attendre à une émergence plutôt rapide de machines pouvant exhiber une certaine autonomie de « pensée » et d'actions non incluses – ni prévisibles – dans la programmation initiale. Le concepteur et le programmeur fournissent les bases d'un potentiel qui va s'épanouir du fait même de la machine en dépassant son supposé ou évalué déterminisme d'action.

Si cette conjecture devient affirmation, il y a évidemment lieu de réfléchir aux conséquences de l'introduction de telles machines dans notre monde. Il s'agit d'un saut technologique considérable. Il implique précautions, prévisions et contraintes à observer, tant au niveau conceptuel qu'à celui des usages. Le groupe de travail va maintenant aborder ces questions, et le prochain rapport traitera de ces sujets, en partant *a priori* de la conjecture qui affirme que des machines exhibant diverses fonctionnalités qu'on attribue à la conscience humaine sont dès maintenant faisables.

ABSTRACT

Implementing data processing functions – as they exist in living bodies – in man-made electronic substrates, whatever the matter used, is one of the most challenging technological issues today. Hence the choice made by the National Academy of Technologies of France (NATF) to include this topic among its current case-studies.

The members of the academic Working Party (WP) who authored this paper framed a general question: is it now possible to endow a machine with a conscience comparable to that shown by Man and so-called higher animals.

Defining and characterising conscience are age-old problems that have generated an abundant range of publications, initially in philosophy, metaphysics and theology, then in psychology and sociology and more recently using science and modern technologies, in biology, genetics, in the theory of evolution, control theory and computer sciences. Yet, despite the huge size of this knowledge base, nobody today seems able to provide a satisfactory answer to the questions that surround the genesis and the functioning of conscience and consciousness.

The NATF Working Party (WP) voluntarily limited itself – after a tentative approach to the general thematic above – to the analysis of a more limited and realistic problematic which, nonetheless, would lead to produce considerable consequences if a positive answer were to be found for the following question: *is it possible to establish a computer implemented model that would exhibit properties, in its relationship with the outside world, similar to those specialists attribute to human conscience – properties that we have not hitherto been able to synthesise on machines?*

A relatively large number of research scientists have proposed models for the conscience. Few of them, however, did so with sufficient precision to enable the establishment of machine implementable models. Notwithstanding, it will be noted that those that exist present original features, with similarities and complementarities.

Conscience is not a physical entity even though its manifestation suggests the existence *ab initio* of a physical "container" as its support (brains are made of carbon, computer chips of silicon). Consciousness is an evolutionary, dynamic process. In other words, when the primitive conditions for the initial spark were fulfilled, conscience appeared as a self-improving phenomenon resulting from an inherent self-learning function.

The WP selected three models that present this characteristic (among others) : 1) the Cardon Model which asserts that "any thought is computable". He proposes a model based upon a massive multi-agent system, stabilization of which corresponds to a "state of thought" ; 2) the Sabah model, with a "Sketchboard" that addresses and embodies the perception functionalities, to which is associated a higher level to benefit from reflexivity and artificial intelligence[2] techniques to generate programmes capable of representing themselves and taking self-decided actions. Both levels are interconnected through memory functions. It is the functionality of the whole that produces "signs" of conscience ; 3) the Pitrat Model; which develops a general problem-solver based on a bootstrapping mechanism and the use of declarative knowledge and meta-knowledge.

In short, each of the three approaches should enable us to reveal functionalities that we attribute to conscience.

The three models are interesting inasmuch as they are compatible, complementary and could be merged into a more generic, single model, as explained in this Report. The range of functionalities made feasible by the global model is enlarged from both from qualitative and qualitative points of view.

If this global model becomes operational, numerous as yet unattained properties of AI would become effective and would allow us to assert formally that the machine manifests a certain consciousness of its existence and this alone would change to a greater or lesser degree its behaviour (more or less free and more or less voluntary).

[2] Abbreviated "AI" in the following English text and "IA" in the fully developped French text.

The three models examined by the authors have only been subjected to partial experimentations (the Cardon Model stayed in "paper format"; various processes of the Sabah's model were implemented; the Pitrat program has been running for more than twenty years). We are therefore restricted to making a conjecture as to the final outcome that would result from complete implementation and testing of the global model (requiring a colossal amount of work).

This paper shows there are no known obstacles or technological barriers that would prevent implementation of the global model. Consequently, if this conjecture is proven, we should expect a fairly rapid emergence of machines endowed with a certain form of autonomous "thought" and able to undertake actions not initially included in the initial programming stages. The designer's and the programmer's roles will be to implant the basic potentialities that the machine itself would reveal in crossing the frontiers of its supposedly closed determinism.

If the conjecture becomes an assertion, we shall necessarily be led to think about the consequences of producing and using such machines in our Society. This in itself would represent a considerable technological leap. It implies that precaution, forecast and constraints be invoked at each step, both from a conceptual and from an end-users' point of view. The next Report will address these very topics, starting with the *a priori* conjecture that machines that exhibit functions hitherto attributed to human conscience can be assembled with today's technology.

AVANT-PROPOS

Lors des tout débuts de l'informatique, l'ordinateur fut considéré comme une super-machine à calculer, mais on s'aperçut très vite qu'il s'agissait de la machine la plus générale permettant de traiter des symboles. En effet, les éléments qu'elle manipule peuvent être interprétés comme des codes et permettent alors de représenter n'importe quel type d'information. Depuis, l'informatique a fait des avancées spectaculaires en matière de puissance de calcul, de rapidité, de miniaturisation, de convivialité, de variétés d'utilisations et d'applications.

Les systèmes informatiques ne sont plus aujourd'hui limités à l'exécution de tâches ponctuelles, en réponse à des commandes humaines. Leur autonomie et leurs capacités d'initiative les rendent capables d'assister l'usager dans des activités de raisonnement variées et complexes : en le guidant et en lui fournissant les connaissances qui lui font défaut ; en prenant en charge la résolution de sous-problèmes précis ; ou, enfin, en proposant des outils que l'utilisateur peut mettre en œuvre et combiner entre eux à son gré, dans le cadre de ses propres stratégies cognitives. La gestion de l'information et l'interaction homme-machine sont devenues des tâches essentielles de l'informatique et de l'intelligence artificielle (IA). Ainsi, le principe du contrôle asservi de l'automatique classique ne peut plus, à lui seul, répondre aux exigences du concepteur et de l'utilisateur.

On est passé des robots industriels et télécommandés (robots asservis) qui restent sous le contrôle de l'Homme aux robots autonomes d'aujourd'hui

qui disposent de capteurs, d'effecteurs et de moteurs de mouvement ; ils sont gérés par techniques d'IA et disposent d'un mode de contrôle dynamique et hiérarchique capable d'adapter les tactiques et les stratégies à tous les niveaux, en fonction de buts immédiats et à plus long terme. Après les petits robots ménagers, pensons à Aldebaran Robotics, qui, malgré son nom, est une entreprise française, leader mondial de la robotique humanoïde, et qui commercialise le petit robot NAO ; elle a le projet de développer un robot d'assistance aux personnes en perte d'autonomie (Romeo) ainsi qu'un robot d'intervention industrielle pour pallier l'emploi d'êtres vivants en situation périlleuse, à la suite des accidents de Fukushima, ce qui évoque également le robot Asimo de Honda, dont l'autonomie lui permet de s'adapter à des éléments non programmés. On sait aussi que les voitures totalement automatisées sont maintenant autorisées au Nevada…

On va aller maintenant vers de vrais robots autonomes qui seront gérés par l'IA évolutive ; celle-ci se réfère aux programmes de la vie artificielle (qui s'inspire du vivant) qui ont la capacité de se modifier eux-mêmes et de s'adapter aux modifications de leur environnement. Peut-on alors parler de conscience ?

Chez l'Homme, il est clair que la conscience naît de l'interaction d'un cerveau, d'un corps physique et d'un environnement (et que ces trois éléments sont tous simultanément nécessaires). L'hypothèse de base ici est donc que ces robots modernes réuniront les conditions nécessaires à l'apparition de la conscience, qui pourra se développer sur n'importe quel support physique. Mais, comme nous ne connaissons pas exactement ce qui fonde notre conscience, nous ne pourrons pas les programmer directement, mais nous pouvons construire des systèmes qui acquerront des fonctionnalités similaires à celles de notre conscience, sans toutefois leur être obligatoirement identiques.

À l'origine de l'IA, l'hypothèse fondamentale sous-jacente est que les processus de pensée sont automatisables et peuvent être simulés sur ordinateur. Cette hypothèse revient à considérer que l'intelligence se manifeste en transformant de l'information et en produisant, à partir de données, des résultats appropriés. L'IA postule que l'intelligence est une propriété générale de systèmes matériels symboliques et essaye de traiter sur ordinateur des problèmes complexes qui sont résolus par l'homme de façon sémantique : alors que l'informatique classique traite les problèmes résolus classiquement par des algorithmes connus, l'IA s'intéresse aux problèmes d'une complexité élevée (appelés NP-difficiles ou exponentiels

par les spécialistes[3]) ou pour lesquels aucun algorithme satisfaisant (c'est-à-dire n'entraînant pas une explosion combinatoire ingérable) n'est connu.

D'autres courants de pensée sont venus enrichir la discipline : la métaphore des réseaux (l'esprit est ramené au fonctionnement du cerveau et l'intelligence est conçue comme la diffusion d'activations, non symboliques, dans des réseaux) et l'IA distribuée (avec les systèmes multi-agents, la pensée est conçue comme un phénomène collectif émergent produit par de nombreux événements élémentaires).

Toutefois, aujourd'hui, le terme « intelligence artificielle » correspond essentiellement à l'utilisation astucieuse de méthodes d'optimisation, auto-adaptatives, très efficaces quand le problème se pose en termes quantitatifs, mais beaucoup moins dès qu'on aborde une approche plus qualitative. Que les bases de données sur lesquelles se fondent leurs raisonnements soient fournies par les programmeurs ou apprises automatiquement, ces programmes restent incapables de prendre des décisions pertinentes dans un monde ouvert et face à des situations totalement imprévues.

En vue d'augmenter les performances des systèmes d'IA, certains chercheurs croient à l'IA forte, qui fait référence à une machine capable non seulement de produire un comportement intelligent, mais d'éprouver une compréhension du monde, une identification de sa propre place dans ce monde et de représenter ses propres raisonnements, bref, de donner l'impression d'une conscience de soi (en limitant cette notion à ce qu'on appelle conscience réflexive). Cette approche est fondée sur l'hypothèse pour des machines d'avoir des expériences sensorielles et émotionnelles, et souligne la nécessité de l'incarnation qui permet des bouclages féconds entre le réel et le subjectif (les émotions relatives à l'état et à la posture du corps physique sont nécessaires à l'intelligence). Ces tenants de l'IA forte posent qu'il n'y a aucune limite fonctionnelle à produire une telle intelligence artificielle

[3] Une mesure de la complexité d'un problème est le temps que prend un algorithme pour le résoudre sur machine, en fonction de la taille des données à traiter (T). Les problèmes les plus simples sont de complexité constante ou linéaire (aT+b), puis polynomiaux déterministes (P(T)). Les plus complexes sont dits NP-difficiles (non-déterministes polynomiaux) ou exponentiels : ils peuvent être traités en énumérant l'ensemble des solutions possibles et en les testant à l'aide d'un algorithme polynomial ou exponentiel (le jeu d'échecs, par exemple, appartient à cette dernière catégorie), ce qui, dans la pratique, prend un temps généralement inacceptable.

consciente et divers modèles correspondant à des niveaux divers de conscience ont été proposés, bien qu'aucune implémentation n'ait encore été menée à son terme. Par ailleurs, si la notion de survie est la première des valeurs essentielles à l'évolution chez l'homme, ce n'est pas le cas chez les machines qui peuvent ainsi prendre des risques beaucoup plus élevés.

Allant encore plus loin, Ray Kurzweil[4] suppose que, dans quelques décennies, l'accélération des capacités des ordinateurs permettra à l'intelligence des machines de rattraper, puis de dépasser l'intelligence humaine. Des nanorobots exploreront nos cerveaux et nous libéreront de nos contraintes physiques. Il pense qu'à terme, l'humanité fusionnera avec la technologie informatique. Si l'émergence d'une conscience n'est que fonction de la complexité d'un système, vu la croissance prévisible et exorbitante des puissances de calcul, ce scénario n'est pas si aberrant qu'il y paraît, bien qu'il fasse abstraction de la capacité « spirituelle » de l'homme, mal cernée, et de la puissance électrique considérable nécessaire pour alimenter des machines complexes à puissance de calcul très élevée.

Le but de ce groupe de travail est d'analyser la pertinence de ces positions, d'en analyser les enjeux (extrêmement importants !) et d'évaluer les chances que de tels projets aboutissent positivement, partiellement ou totalement. Selon les résultats de cette première étape, le groupe de travail envisagera par la suite d'examiner ce qu'une telle nouvelle fonctionnalité implantée sur les machines aura comme impact sur notre société. Des recommandations allant dans ce sens seront explicitées.

[4] Directeur technique de Google et transhumaniste.

LES POINTS QUE NOUS N'ABORDONS PAS

Nous analysons la capacité que nous avons d'observer notre propre fonctionnement, et nous n'étudions pas les possibilités de doter un système artificiel de la capacité à distinguer le bien du mal (même si, à terme, cela sera certainement nécessaire). Nous n'approfondirons pas non plus les éventuelles relations entre les différents sens du mot conscience, comme par exemple, tenter de déterminer si l'un est premier et si les autres en découlent. Nous nous intéressons donc à la fonction *introspective* de la conscience (dite *conscience réflexive)* qui se distingue aussi bien du fait d'être éveillé que de la conscience morale qui juge du bien et du mal.

Certains termes qui interviennent dans cette étude possèdent plusieurs significations. Nous avons tenté de les définir (fichier terminologie à consulter en fin d'ouvrage) tout en précisant le sens dans lequel nous les utilisons ici selon notre interprétation propre. Nous ne considérons pas forcément toutes les ambiguïtés de ces termes.

Nous nous limitons au domaine des TIC (*technologies de l'information et de la communication*) et nous n'abordons pas le sujet du point de vue de la convergence NBIC (nanotechnologies, biologie, informatique et cognition) ; en particulier, nous ne traitons pas la constitution biologique de la conscience, ni les questions liées aux pathologies de la conscience.

Nous n'approfondissons pas les technologies informatiques les plus pertinentes, nous n'en proposons pas de nouvelles, mais nous utilisons les modèles

disponibles (programmation classique, IA distribuée, systèmes multi-agents, réseaux neuronaux artificiels…). Nous ne traitons pas le problème d'une modélisation informatique de la conscience humaine, ni sur le plan matériel (recherche d'une structure informatique analogue au cerveau), ni sur le plan logiciel (constitution d'un programme ayant les fonctionnalités de la conscience), même si certains des modèles que nous évoquons peuvent avoir cette ambition.

Nous visons simplement à identifier certaines fonctionnalités de la conscience et à déterminer celles qui seraient simulables sur machine, en analysant les modèles existants.

Nous ne considérons aucun aspect philosophique, métaphysique ou religieux lié à la notion de conscience, hors du champ épistémologique abordé dans ce rapport. Nous n'avons conservé que les modèles suggérant des pistes d'implémentation de fonctionnalités de la conscience.

DOCUMENTS DISPONIBLES

L'ensemble des documents produits par le groupe de travail est accessible à l'adresse : http://gscns.free.fr/

Ce sont :

- les présentations effectuées lors des réunions par les membres du groupe ou par les experts invités (essentiellement des fichiers Word ou pdf, et des PowerPoint) ;
- les comptes rendus de toutes les réunions ;
- un glossaire (78 termes) (joint en annexe au présent document) ;
- une bibliographie (253 références) (joint en annexe au présent document) ;
- divers autres documents pertinents.

En outre, sont disponibles les fichiers suivants :

- « TowardConsciousMachines.doc » : une synthèse (en anglais) des travaux du groupe de travail ;
- « Vers une conscience artificielle » : la version française de « TowardConsciousMachines », qui constitue en fait l'ossature du présent rapport ;
- « ModelsValidation.doc » : une réflexion sur la notion de modèles et leurs validations, en particulier dans le domaine des sciences humaines.

PROLÉGOMÈNES

Cette partie « prolégomènes » nous semble importante pour bien saisir l'esprit de cette communication ; cependant, après celle-ci, le lecteur pressé de découvrir nos apports effectifs pourra se reporter directement à la page 52 (le lecteur très pressé pourra même aller directement à la page 57).

Introduction

Le groupe de travail « vers une technologie de la conscience ? » s'inscrit dans la suite du groupe de travail « Interaction homme-machine » qui a clos ses travaux en 2004 et a donné lieu à un rapport de l'Académie.

Depuis lors, la robotique et l'IA ont beaucoup évolué et il semble qu'aujourd'hui le défi visant à doter des objets ou des systèmes informatiques d'une « *conscience réflexive* » (capacité à se représenter soi-même et à raisonner sur ces représentations ainsi que sur ses propres actions) ayant certaines caractéristiques de celle de l'homme soit à la portée des chercheurs. Si cela se confirme, on mesure l'importance des conséquences potentielles et des précautions à prendre.

Il paraît donc nécessaire de mener une réflexion sur ce sujet. C'est l'objet de ce groupe de travail qui vise à identifier, à inventorier les éléments qui fondent la

conscience et déterminer ceux qui pourraient raisonnablement être intégrés à des machines ou des systèmes.

Le premier point à souligner est la multiplicité de concepts que recouvre la notion de conscience (comme d'ailleurs les notions liées d'intention, d'attention, de mémoire…). Différents sens de « conscience » apparaissent selon la langue considérée, aussi bien à partir d'exemples issus de la Bible que de la langue anglaise qui propose déjà quatre mots pour exprimer des états très différents. Quand on parle de la conscience pour qualifier l'état d'éveil par rapport au sommeil ou au coma, on emploiera le mot *awakeness*, de la conscience de *telle ou telle information particulière*, on emploiera le mot *awareness*, de la conscience d'être soi-même, le mot *self-consciousness*, de la conscience morale pour qualifier tel ou tel acte, le mot *conscience*.

Il existe différentes catégorisations plutôt convergentes de nos différentes formes de conscience. Par exemple, pour le philosophe Ned Block [1], 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 prémisse pour le raisonnement et peut jouer un rôle dans le contrôle rationnel de l'action et de la parole. Ce concept rappelle celui d'espace de travail neuronal ;
- *la conscience phénoménale*, qui correspond aux aspects qualitatifs de notre vie mentale (ou *qualia*). En d'autres termes, « l'effet que cela fait » de ressentir une douleur, de percevoir une couleur, etc.
- *la conscience réflexive* (ou conscience de « *monitoring* ») 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.

Bien entendu, concevoir un système possédant toutes les facultés correspondantes est totalement inenvisageable à l'heure actuelle. Toutefois, certaines de ces significations sont utiles pour les machines ; ainsi, la conscience réflexive (*consciousness*) est nécessaire pour apprendre et s'adapter à de nouvelles situations : elle donne des informations sur nos actions et sur le pourquoi on les

accomplit ; la conscience morale (conscience) est nécessaire pour être autonome : elle donne des informations sur le caractère bon ou mauvais, adéquat ou inadapté (choix des buts, des actions, conséquences de celles-ci) des décisions prises. Par ailleurs, considérant les activités qui sont possibles pour les machines, mais pas pour les êtres humains, on pourrait aussi envisager l'émergence de nouveaux sens spécifiques de cette notion pour elles.

Quand le groupe de travail a commencé à réfléchir, son ambition était très modeste. Beaucoup de membres du groupe ayant travaillé en informatique ou en robotique, la question était de savoir s'il était possible de fabriquer des artefacts munis de capteurs et d'actionneurs, susceptibles d'interagir avec d'autres artefacts ou avec des humains et ayant des réactions qui ne soient pas issues d'un déterminisme programmatique évident. Allant plus loin, le groupe s'est demandé si ces artefacts pourraient avoir des réactions inattendues, raisonnées et décidées, par une capacité perceptive de la complexité de leur environnement qu'ils sauraient confronter à des bases de connaissances d'environnements comparables, à des bases de connaissances d'un ordre supérieur, contenant des scénarios de réactions possibles dans ces environnements, voire à des bases de connaissances, d'un ordre supérieur encore. Seraient-ils également capables d'apprécier ces scénarios au regard de valeurs susceptibles de régir leurs interactions avec d'autres artefacts ou avec des humains ; toutes ces bases de connaissances, de surcroît, pouvant s'enrichir grâce à des processus d'apprentissage constants au cours de l'histoire de ces artefacts.

En décrivant ainsi des artefacts, on ne décrit pas l'Homme comme ensemble de composants matériels entrelacés dans une extrême complexité, et capables des meilleurs ou des pires relations avec ses semblables. La procréation humaine reproduit cette complexité lui permettant d'évoluer à chaque génération, et souhaitant que l'être procréé le soit avec de multiples degrés de liberté qui constitueront une altérité individuelle radicale entre les procréateurs et le procréé. La question que nous pouvons nous poser est de savoir si nous serons un jour capables de profabrication d'entités à qui nous pourrions reconnaître la double caractéristique de nous être apparemment totalement semblables dans une altérité individuelle radicale inaliénable qu'ils pourraient revendiquer. Est-ce le rêve prométhéen ou celui du Golem qui nous anime en cherchant à construire de telles entités ? Peut-être … Ne serait-ce pas plus modestement le désir de mieux comprendre l'Homme dans ses propres interactions avec ses semblables et son environnement ?

Aussi le groupe a-t-il auditionné non seulement des spécialistes des systèmes intelligents de traitement de l'information (comme le système mis en œuvre par Google) ou des roboticiens pouvant nous entretenir des plus récentes avancées dans leur domaine, mais encore des neurologistes spécialistes des sciences cognitives, qui nous ont passionnés par les modèles du cerveau qu'ils élaborent, et des philosophes qui nous ont présenté bien des points de vue sur la conscience humaine.

Ce processus d'exploration, culturellement passionnant, mais ouvert à de si nombreuses directions, nous conduisait à une explosion cognitive qui ne pourrait jamais donner lieu à un rapport de synthèse. Il a alors été décidé de revenir à la problématique de systèmes informatiques ou de robots et à leur capacité de simuler certaines fonctions aux contours bien précis que l'on peut voir à l'œuvre dans la conscience humaine quand on en fait une analyse fonctionnelle.

Poser la bonne question

Depuis les années 1990, où l'étude de la conscience s'est développée et a été reconnue comme une recherche scientifique, on trouve de nombreuses études en neurobiologie, sociologie, psychologie, philosophie et même en théologie. Pour s'en convaincre, il suffit de regarder les 28 490 entrées de la bibliographie proposée par Chalmers et Bourget [2], toutes concernées par les sciences cognitives et l'IA et dont plus de 12 000 mentionnent effectivement la conscience. De la même façon, le web propose plus de dix millions de références associées au mot « robot » dont un pourcentage important a trait à l'intelligence de ces machines, précurseur de la conscience. On se trouve donc submergé par un flux informationnel dans lequel il est difficile de sélectionner ce qui est important pour notre sujet en posant les questions pertinentes.

On remarque toutefois, qu'à propos de la conscience, très peu des recherches vont vers des modèles programmables : l'impossibilité d'étudier concrètement ce qui se passe dans le cerveau explique pourquoi il y a tant de théories souvent contradictoires sur son fonctionnement.

La question que nous posons ici (et que la plupart de ces études ne posent pas !) **n'est pas** de savoir si un robot ou un système est (ou pourra être) conscient (nous avons vu ci-dessus qu'on ne sait pas définir de façon univoque ce qu'est

« la conscience »), **mais quelles sont les fonctionnalités que l'homme attribue à sa conscience et qui pourront être mises en œuvre dans les futurs robots ?**

Pour ce faire, nous identifions ci-dessous diverses fonctionnalités liées à la conscience humaine. Puis, au paragraphe *Modélisations scientifiques et concrétisables*, nous présentons quelques propositions pratiques donnant de telles capacités à des machines, robots ou systèmes en général, et nous en proposons une synthèse. Nous sommes loin d'être exhaustifs, mais, notre réflexion est fondée sur de nombreux travaux tels que : Harth et ses boucles créatives [3], Edelman et sa TSGN [4-6],Baars et son « *Global Workspace* » [7, 8], Dennett et son *pandémonium des esquisses multiples* [9, 10], Varela et sa théorie de l'autopoïèse [11] ou Damasio et le rôle important donné aux émotions [12]. Nous analysons alors les capacités potentielles de cette synthèse par rapport aux fonctionnalités présentées au paragraphe *Fonctionnalités attachées à la notion de conscience*, avant de conclure.

Une machine consciente, pour quoi faire ?

Dès ses débuts, l'IA avait deux buts distincts : simuler les raisonnements intelligents pour comprendre certains aspects du fonctionnement de l'homme, ou bien résoudre des problèmes le plus efficacement possible, sans vouloir à tout prix faire comme l'homme. Les recherches sur la conscience en découlent et présentent les mêmes aspects. L'intérêt de la notion de machine consciente est montré ci-dessous dans deux domaines : la résolution de problèmes et le traitement automatique des langues.

Les recherches en résolution de problèmes s'inspirent généralement du fonctionnement de l'homme mais ne se contraignent pas à le simuler ; on pourra donc éventuellement aboutir à de nouvelles méthodes de mise en œuvre de consciences ayant peut-être des possibilités différentes (supérieures ou inférieures) aux nôtres.

La conscience réflexive permet d'analyser les différentes étapes d'un raisonnement pendant que celui-ci a lieu et donc, grâce à un méta-raisonnement, d'évaluer constamment la direction à choisir ; elle donne ainsi des informations sur les actions effectuées et sur les raisons pour lesquelles elles ont été accomplies, permettant, d'une part, de justifier les plans choisis et, d'autre part, d'acquérir de nouvelles connaissances en analysant les raisons d'un succès ou d'un échec.

Dans CAIA (chercheur artificiel en intelligence artificielle, développé par Pitrat [13]) par exemple, l'auteur analyse les mécanismes informatiques utilisés dans les cas de succès et constate que ceux qui relèvent d'une analogie avec la conscience, lorsqu'ils sont présents, sont nécessaires à la résolution du problème.

La conscience morale (au sens anglais de *conscience*) est nécessaire pour être autonome : elle donne des informations sur le caractère bon ou mauvais, adéquat ou inadapté des décisions prises. Comme elle se fonde sur des principes généraux, elle est nécessaire pour s'adapter efficacement à des situations nouvelles en donnant des heuristiques pour choisir les buts à atteindre, les actions à effectuer et les conséquences de celles-ci. Bien entendu, une telle conscience n'exerce son discernement que par rapport à des règles de conduites qui peuvent être appliquées ou non, et non par rapport à des lois auxquelles on ne peut se soustraire. Par ailleurs, les machines peuvent faire fi de certaines de nos limitations : leur conscience réflexive peut être ajustable (elles peuvent observer tout de leur fonctionnement), elles peuvent être clonées à faible coût (ce qui est intéressant pour tester divers modes d'apprentissage, en fonction de différences clairement identifiables) ; enfin, en ce qui concerne la conscience morale, les machines peuvent être considérées comme esclaves et, comme nous l'avons déjà souligné, la survie n'étant pas leur valeur essentielle, elles peuvent prendre de grands risques.

Ces raisons sont également valables pour le traitement automatique des langues. Du point de vue de l'efficacité informatique, [14, 15] ont mis en évidence l'inadéquation des architectures classiques d'exécution sérielle de tâches, et la nécessité de mettre en œuvre des systèmes coopératifs, dits « *multi-agents* », notamment pour ne pas introduire d'ambiguïté artificielle dans la compréhension du langage. Ces systèmes permettent un contrôle dynamique utile pour l'adaptation à des situations diverses mais peuvent se révéler insuffisants face à la résolution des problèmes où une « conscience réflexive » est nécessaire. Donnons-en deux exemples où le texte lui-même (partie soulignée ci-dessous) décrit la procédure à utiliser pour le comprendre. Le premier est inspiré de « ixage d'un paragrab », de Poe [16]

« Jean écrit **en remplaçant les « o » par des « x »**. Il écrit à Paul : "nxus viendrxns demain" ».

Ici, il faut modifier la procédure classique d'analyse morphologique : remplacer d'abord les x par des o et voir si le nouveau mot existe, quand on a détecté que c'est Jean qui écrit.

Autre exemple, tiré de La maison d'âpre-vent, de Dickens [17] :

« Et il m'a dit, ajouta-t-il, **en jouant de petits accords aux endroits où je mettrai des points**, que Chécoavins avait laissé. Trois enfants. Sans mère. Et que la profession de Chécoavins. Étant impopulaire. La génération montante des Chécoavins. Était dans une situation très difficile[5] »[6].

Là, il faut modifier l'analyse syntaxique en ne considérant pas les points, et modifier la procédure d'interprétation de la situation en mémorisant que des accords ont été joués là où figuraient ces points.

Il est clair que ces exemples restent hors du champ des systèmes actuels de traitement automatique des langues ; seuls des systèmes multi-agents *réflexifs* et conscients pourront les traiter efficacement.

[5] "And he told me" he said, playing little chords where I shall put full stops, "The Coavinses had left. Three children. No mother. And that Coavinses profession. Being impopular. The rising Coavinses. Where at considerable disadvantage."

[6] Cité dans *Dialogues et sciences cognitives* – Gérard Sabah – 1994 et dans *La linguistique cognitive* – Catherine Fuchs – 2004).

GÉNÉRALITÉS SUR LA CONSCIENCE

Le cerveau est l'interface entre le réel perçu par les sens et ses représentations mémorisées ; ces représentations, dites mimèmes, expliquent la possibilité d'apprentissage par mimétisme (rôle des « neurones miroirs »), influencé par le contexte et les émotions ; en liaison avec l'attention, un langage élaboré permet ensuite une coopération efficace entre individus. Le monde extérieur impose ses structures à différents niveaux : matériel bien sûr, informationnel ensuite (relations entre les éléments matériels et la (les) signification (s) qu'ils transmettent), puis relations entre le vivant et l'information (mémoire et outils), et enfin les différents points de vue selon lesquels considérer les êtres vivants (du niveau biochimique à la conscience).

Coste, le traducteur de Locke, introduisit vers 1700 l'usage moderne du mot conscience [18] (en français, mais le sens du mot *consciousness* était tout aussi nouveau), associé à l'idée d'un soi-même dont la conscience exprime l'identité. Freud localise la conscience sur la couche externe du cerveau [19] ; la couche interne, selon lui, correspond à l'inconscient. Plus récemment Sperry [20], ayant sectionné le corps calleux reliant les deux hémisphères du cerveau pour traiter l'épilepsie, constata un certain dédoublement de la conscience. Une des deux consciences était verbale et analytique et semblait correspondre à l'hémisphère gauche du cerveau. L'autre conscience, plus subjective, semblait correspondre à l'hémisphère droit du cerveau. Sperry en déduisit que la conscience n'est pas localisée à un endroit particulier dans le cerveau : toutes les parties du cerveau y contribuent simultanément.

Ainsi, en relation avec la notion d'apprentissage par mimétisme et interaction, diverses formes de conscience peuvent être évoquées (homme, animal, robot). Mais, la conscience dynamique d'exister et la possibilité d'un dialogue conscient avec un autre individu susceptible de compréhension seraient caractéristiques de l'humain.

Recensement des questions générales à propos de la conscience

Parmi les questions générales abordées dans la communauté « étude scientifique de la conscience », Chalmers [21] pose sa « question difficile » (on ne cherche pas seulement une corrélation entre états du cerveau et états de conscience, mais une explication ou « pourquoi l'exercice des fonctions cognitives s'accompagne-t-il d'une expérience subjective ? »). Puis, Dennett [22] exprime un avis contradictoire en proposant l'analogie des arguments de Chalmers avec 1) la vie (avoir toutes ses fonctionnalités mais ne pas être vivant ?) et 2) avec la perception (on ne peut simuler une vraie vision qu'avec de l'interprétation) ; il conclut que la question posée par Chalmers est absurde. Il prétend que la résolution de l'ensemble des problèmes « faciles » conduira nécessairement à la résolution du problème difficile (l'ensemble des fonctionnalités considérées débouche naturellement sur les qualia et la conscience).

Quelles sont les questions que la communauté pose usuellement à propos de la conscience ? Voici les principales :

1. Qu'est-ce que la conscience ?
2. Les expériences subjectives peuvent-elles être expliquées en termes physiques ?
3. Quelles sont les méthodes les plus appropriées et potentiellement fructueuses pour l'étude de la conscience ?
4. Quels sont les corrélats neuronaux de la conscience ?
5. Quelles nouvelles méthodes d'imagerie cérébrale pourraient aider à clarifier la nature et les mécanismes de la conscience ?
6. Quelle est la relation entre les processus conscients et inconscients dans la perception, la mémoire, l'apprentissage, et d'autres domaines ?

7. Quelles sont les propriétés de l'expérience consciente dans des domaines spécifiques tels que la vision, l'émotion et la métacognition ?
8. Comment pouvons-nous comprendre les troubles et les formes inhabituelles de la conscience, que l'on trouve dans la « vision aveugle », la synesthésie, et d'autres syndromes ?
9. Est-ce que la conscience joue un rôle fonctionnel dans la vie, et si oui quel est ce rôle ?
10. Pouvons-nous développer des méthodes rigoureuses d'étude et de formalisation des données sur l'expérience consciente (point de vue en première personne) ?
11. Quel rôle joue l'expérience subjective dans les théories existantes de la science moderne ?
12. Quelles seraient les implications d'une science de la conscience sur l'éthique et la société ?

Parmi celles-ci, le GT s'est essentiellement intéressé aux questions : 1-5-6-8-10-11. Il pose également d'autres questions fondamentales :

- une conscience artificielle est-elle possible ? (en principe ; techniquement ; financièrement…)
- y a-t-il des contre-arguments ?[7]
- si elle est possible, quelles sont les chances qu'elle soit réalisée un jour ? Quand ?
- si elle est réalisée un jour, quelle sera son autonomie ?
- quels dangers pourra-t-elle impliquer pour l'humanité ?
- la réalisation d'une conscience artificielle est-elle souhaitable ?

[7] On peut par exemple, se référer à Dennett, qui conteste les contre-arguments suivants, et les considère comme non valides (http://users.ecs.soton.ac.uk/harnad/Papers/Py104/dennett.rob.html) :

(1) Les robots sont purement matériels, et la conscience exige un esprit immatériel. (*Old-fashioned dualism*)

(2) Les robots sont inorganiques, et la conscience ne peut exister que dans un cerveau biologique

(3) Les robots sont des artefacts, ce que la conscience abhorre ; seul quelque chose de naturel, non fabriqué, pourrait présenter une conscience authentique.

(4) Les robots seront toujours beaucoup trop simples pour être conscients.

Fonctionnalités attachées à la notion de conscience

Sans vouloir prétendre à l'exhaustivité, nous précisons ici diverses fonctionnalités, usuellement reconnues comme des propriétés de la conscience humaine (en surligné ce qu'on sait faire ou modéliser, souligné ce qui est prévisible, en **gras** ce qui n'est pas encore accessible).

Capacités de représentation et d'interprétation

Interprétation et unification des données des sens

Les images mentales ne sont pas des répliques d'objets du monde : elles sont combinées avec les connaissances antérieures grâce à des chemins descendants qui modifient les messages des sens et injectent des informations supplémentaires, à la commande des niveaux supérieurs. La conscience fait émerger les interprétations les plus pertinentes pour nos connaissances et le contexte. Exemple : voyant dans un texte écrit : /-\ on l'interprète comme un A s'il est entre C et R, ou comme un H entre C et AT.

Représentation et interprétation de l'environnement par rapport à soi

À partir d'informations sensorielles parcellaires et complexes, la conscience construit une cohérence de l'ensemble et donne ainsi une image simplifiée, unifiée ou enrichie du monde. Exemple : dans une expérience où un point vert s'allume quelques dixièmes de seconde après un point rouge situé quelques centimètres à sa gauche, on construit systématiquement un point intermédiaire (qui n'existe en fait pas) allant de l'un à l'autre et changeant de couleur au milieu.

Représentation d'éléments imaginés

La conscience permet de construire des représentations d'éléments qui ne figurent pas dans l'environnement courant, et même qui pourraient ne pas exister. *Exemple :* souvenirs modifiés de scènes vécues antérieurement ou imaginaires : licornes, Pégase et autres phénomènes.

Fonction de constitution et de perception de soi

L'impression que la conscience nous donne d'avoir un esprit détaché du corps, même si elle est fausse, augmente la valeur que nous donnons à nos existences ; cela revient à construire une « théorie de l'esprit » (de soi, des autres). Exemple : on se représente ses propres caractéristiques et on affecte aux autres des aptitudes analogues.

Articulation avec l'inconscient ; capacités de sélection

Des activités diverses se déroulent en parallèle dans le cerveau ; tous les niveaux de traitement ne sont pas conscients (que ce soit au niveau des perceptions, des sensations ou des sentiments). Or, on ne peut être conscient que d'une chose à la fois. La conscience produit donc un effet de séquentialisation et de sélection d'éléments pertinents dans les représentations. Exemple : cela justifie la raison pour laquelle on interdit l'utilisation du téléphone portable en voiture.

Le **contrôle de soi et de l'environnement**, de nos pensées et de nos comportements procède d'un système s'informant constamment de l'activité d'une multitude de sous-systèmes inconscients fonctionnant en parallèle afin de coordonner toute cette activité et de la contrôler. *Exemple :* cette faculté nous permet d'intérioriser les règles sociales, d'avoir une personnalité constante, et d'être capables de faire des choix sensés ou conformes à nos besoins.

On peut *choisir* rapidement des *comportements* appropriés à l'environnement en fonction d'intentions et de buts à atteindre sans se soucier de la complexité des commandes motrices à donner. Exemple : l'adaptation automatique du niveau de planification selon ce qu'on connaît de nos capacités, sans entrer dans des détails inutiles.

Gestion de l'attention

L'attention joue le rôle de contrôle métacognitif de la conscience : l'attention volontaire correspond à un contrôle conscient de l'accès à la conscience et s'oppose à l'attention automatique, liée à un contrôle inconscient de l'accès à la conscience. Possibilité de lutte entre attention volontaire et attention involontaire. Exemple : cas où quelqu'un, regardant un film, voit son attention attirée par l'émission de son nom dans la pièce où il se trouve.

Capacités de décision ; actions

Elles concernent *l'ordonnancement des buts, la gestion d'hypothèses, la planification, le déclenchement de processus planifiés, le contrôle de l'exécution des buts* et la résolution de problèmes explicites. L'IA traite ces questions depuis ses débuts. À partir d'une description de l'état initial d'un monde, d'une description d'un but à atteindre et d'un ensemble d'actions possibles (décrites de telle façon que chaque action spécifie des préconditions qui doivent être présentes dans l'état actuel pour qu'elle puisse être appliquée, et des postconditions qui représentent les effets sur l'état actuel du monde), on sait calculer la (ou les) séquence(s) d'actions pertinentes pour atteindre les buts visés et contrôler leur exécution. Quand les moyens élémentaires nécessaires à la résolution des problèmes qui apparaissent sont disponibles, les processus de planification peuvent résoudre la question aisément, mais s'il est nécessaire d'être créatif et de produire des processus nouveaux, la situation devient beaucoup plus délicate.

La plupart de nos comportements se font automatiquement sans faire appel à la conscience : ce n'est que lorsque les choses ne se passent pas comme prévu que nous faisons appel à des raisonnements conscients. Là aussi, cette question fut abordée par l'IA, dans un premier temps en essayant de recenser tous les « imprévus prévisibles » (! ?) et en proposant des réactions adéquates dans chacun de ces cas. Mais, calculer automatiquement des comportements adéquats dans des situations totalement nouvelles et réellement imprévisibles reste actuellement hors de portée.

Capacités de mémoire

Elles ont trait essentiellement à *la gestion de la mémoire à court terme*, aux *mécanismes* d'apprentissage, *à rendre accessible de l'information pertinente* et à la décision d'apprentissage volontaire.

La mémoire est un processus intellectuel dont le but est d'acquérir des informations pour les stocker et les réutiliser. Les relations entre mémoire et apprentissage sont si étroites qu'on confond souvent les deux. Toutefois, la distinction essentielle porte sur le fait que la mémoire désigne la capacité à retrouver des expériences passées, et l'apprentissage tout processus susceptible de modifier un

comportement ultérieur. Il convient de distinguer plusieurs sens au mot mémoire, et en particulier la mémoire informatique (servant à stocker des informations plutôt figées), et la mémoire humaine (qui est essentiellement une capacité à répéter une performance).

On distingue plusieurs systèmes de mémoire. Le premier système mémorise les informations sensorielles issues du monde extérieur. La première étape du travail de ce système (mémoire à court terme ou mémoire immédiate) est transitoire et correspond à une mémoire attentionnelle, c'est-à-dire aux capacités de concentration. Cette mémoire est très sensible à certaines interférences et permet de retenir simultanément quelques informations (comme les chiffres d'un numéro de téléphone) sans qu'ils soient conservés longtemps. Toute perturbation des ressources attentionnelles rendra difficile l'étape ultérieure de la mémoire à long terme, dont le but est de stocker et de rappeler les souvenirs, que ce soit après quelques minutes ou plusieurs années.

La mémorisation à long terme consiste d'abord à organiser logiquement les informations : comparaison des souvenirs anciens et de l'information nouvelle à mémoriser, analyse de son sens, associations d'idées ou d'images, ajout d'étiquettes émotionnelles, chronologiques ou spatiales (moyens mnémotechniques). La motivation est aussi un élément fondamental pour ancrer durablement des souvenirs. Leur conservation nécessite enfin un processus de consolidation, car les événements récents sont encore fragiles et le sommeil, d'une part, le réapprentissage de l'autre, permet de renforcer leur fixation dans le cerveau. Cette consolidation permanente explique la meilleure qualité des souvenirs anciens par rapport aux récents.

Le stockage des souvenirs s'organise alors en deux types : la mémoire sémantique et la mémoire des épisodes. La mémoire sémantique est celle des connaissances partagées par un groupe social ; détachée du contexte d'apprentissage, elle permet la mise en évidence des traits communs à divers épisodes. La mémoire des épisodes est une mémoire plus personnelle qui permet de retenir les événements vécus dans un contexte précis. Elle est plus sensible au vieillissement cérébral.

À côté de cette mémoire du monde extérieur, on peut parler également d'une mémoire des informations intérieures à l'individu, c'est-à-dire des émotions. Un souvenir comporte pratiquement toujours un contexte émotionnel (par exemple de peur ou de plaisir) qui renforce sa consolidation. De même, une motivation importante renforcera l'apprentissage alors qu'une indifférence vis-à-vis de l'événement à mémoriser aura un effet négatif.

On peut enfin évoquer une mémoire motrice, qui rend progressivement automatiques certains gestes, comme la pratique du vélo ou les gestes techniques professionnels.

Du point de vue du processus d'acquisition de ces divers types de connaissances, on peut aussi distinguer entre mémoire implicite qui permet d'apprendre sans retenir le souvenir de l'expérience ayant permis l'apprentissage et mémoire explicite qui permet de garder (au moins provisoirement et avant généralisation) les événements liés à l'apprentissage.

Aspects sociaux

On peut les classer en trois catégories principales.

D'abord, **attribuer des désirs, des intentions, des motivations aux autres** (voir le dernier point du paragraphe *Capacités de représentation et d'interprétation*). Cette capacité à se construire une « théorie de l'esprit » des autres permet à l'être humain normal d'attribuer aux autres, certains états mentaux proches des siens. Cette prédisposition est si forte qu'elle peut être déclenchée par n'importe quel objet pouvant être identifié à un agent généralement doué d'intentionnalité, mais parfois non ! (anthropomorphisme).

Prédire des comportements (soi, les autres). Durant son développement, un enfant comprend graduellement que les autres ont des désirs, des intentions, des motivations, bref un point de vue différent du sien. Ainsi, divers auteurs pensent qu'une conscience de soi aurait émergé progressivement au cours de l'évolution à mesure que les groupes sociaux se complexifiaient et donnaient un avantage à ceux qui étaient capables de se mettre à la place des autres.

Pour certains, cette élaboration d'une théorie de l'esprit pour les autres est même première : c'est elle qui aurait mené à la possibilité de nous construire une théorie de l'esprit appliquée à nous-mêmes, et donc de nous reconnaître des désirs, des intentions, des motivations…

Enfin, **permettre une communication sociale efficace**. La conscience d'exister comme sujet et la conscience de soi comme objet social se développent simultanément, tout d'abord par des interactions non symboliques avec les autres, puis par des interactions symboliques permises par le langage. Le langage est donc fondamental pour la conscience, en ce qu'il nous permet d'écouter autrui, de considérer

son propre point de vue, de défendre notre point de vue, de négocier… Cependant, on sait que le langage ne peut être un reflet parfait de la véritable pensée du locuteur. Des éléments d'intériorisation de la pensée ne peuvent s'exprimer par un langage oral comme l'a montré Wittgenstein par exemple [23]. Le lien entre la conscience et le langage reste donc complexe.

La figure 4 (p. 60) donne une idée des premières étapes possibles de la construction de la conscience, étapes qui explicitent l'émergence du langage (lexique, syntaxe et sémantique). Elle est complétée par les étapes suivantes de la figure 1 qui explicite quelques-unes des relations existant entre les notions abordées ci-dessus.

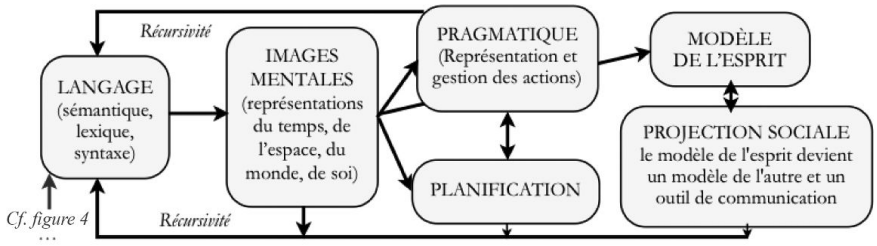

Figure 1 : Langage et conscience.

MODÉLISATIONS SCIENTIFIQUES ET IMPLÉMENTABLES

Qu'est-ce qu'un modèle ?

Soulignons tout d'abord que le mot *modèle* a deux sens en quelque sorte opposés, le premier essentiellement en logique formelle (où un modèle est une occurence particulière d'une théorie : l'arithmétique des nombres rationnels est un exemple de modèle de la structure de corps commutatif), le second plus courant et intuitif désigne une représentation, généralement simplifié, d'un objet, d'un système ou d'un phénomène. *Le lecteur peu familier avec la logique ou la sémantique formelle pourra aller directement à l'historique de la page 41 …*

1°) Une réalisation particulière (on dit aussi une instance) d'une structure abstraite. Ce sens se retrouve dans les systèmes formels quand on veut donner une sémantique à un ensemble de formules : on définit l'ensemble $\mathcal{D}$ des constantes de l'univers (appelé domaine) et une *interprétation* des fonctions et des prédicats. Ceux-ci ne sont en effet que des *identificateurs* qui ne sont porteurs d'aucune signification *a priori*. Ce n'est que lorsque l'on a indiqué le sens de toutes les fonctions et de tous les prédicats que l'on pourra dire que les formules *représentent* effectivement des connaissances. Un *modèle* d'un ensemble des formules consiste donc, outre la définition du domaine $\mathcal{D}$ et des tables de vérité des connecteurs, en la mise en évidence pour chaque prédicat n-aire d'une relation sur $\mathcal{D}^n$ et pour chaque fonction n-aire d'une

fonction de $\mathcal{D}^n$ dans $\mathcal{D}$; il s'agit donc ici d'une *instanciation* particulière respectant un certain formalisme (nous réserverons le mot « interprétation » pour ce concept).

2°) Un schéma formel *plus abstrait* que le phénomène dont il est l'image. Il s'agit cette fois d'une généralisation respectant la vérité des données spécifiques dont elle est issue.

C'est essentiellement ce second sens qui sera utilisé dans l'ensemble du présent document (il sera désigné par le terme « modèle formel » ou plus simplement « modèle »).

On rencontre dans tous les modèles scientifiques des concepts « sémantiques » (renvoyant à des aspects particuliers des phénomènes modélisés) et d'autres n'ayant aucun répondant expérimental direct (ils ne correspondent pas directement à des aspects du phénomène) : les concepts « syntaxiques » (fonctions combinant d'autres concepts). Cette opposition n'a rien d'ontologique, elle est relative à un état momentané de la connaissance ; le même concept peut jouer successivement un rôle syntaxique dans un certain système, puis, dans un autre système, un rôle sémantique. Un excellent exemple en est donné par la notion d'entropie en thermodynamique : initialement utilement introduite par Carnot et Clausius à un niveau purement formel (en tant qu'intégrale définie du quotient d'une variation de quantité de chaleur par une température), on ne s'est rendu compte de son interprétation sémantique concrète que plus tard, avec l'évolution des connaissances sur le sujet (lorsque Bolztmann et Planck la rattachent à la probabilité de réalisation d'un état du système) ; même s'il reste ainsi éloigné d'une mesure expérimentale directe, ce concept désigne alors quelque chose de précis par rapport au phénomène réel. Mais, on peut également rencontrer des évolutions dans le sens inverse : la masse d'inertie a un rôle absolument sémantique en mécanique classique, rôle qui devient beaucoup plus syntaxique avec la relativité.

Toutefois, si l'état des sciences et des techniques joue un rôle majeur dans les réalisations d'une certaine époque, rien n'empêche de concevoir des modèles sur-généralisateurs dont aucune réalisation concrète n'est possible à l'époque de conception (que l'on pense par exemple aux appareils imaginaires de Léonard de Vinci ou à la machine de Babbage). Un aspect essentiel de l'IA est précisément d'avoir créé la possibilité de « simuler réellement » les activités de telles instanciations de modèles (*réalités virtuelles, augmentées, etc.*) ; le jugement de la validité d'un de ces modèles est alors reporté sur son comportement constaté et non plus sur des critères formels (nous approfondirons plus particulièrement ce point au paragraphe *Quelle validation pour un tel modèle ?*).

Divers types de modèles formels

Par rapport à une réalité supposée exister objectivement, un modèle est une relation (application φ) entre cette réalité et un langage formel de description. Il peut ainsi être vu tout d'abord comme un mécanisme de représentation. À ce niveau, on dispose donc d'un outil mathématique formel, dont la seule validation possible est une preuve interne d'autocohérence ; cela n'implique nullement que les modèles dérivés correspondent à une quelconque réalité… Ces modèles **descriptifs** permettent une analyse rétroactive des activités passées, et se retrouvent dans des disciplines comme biologie et neurobiologie (un danger à signaler est qu'ils peuvent parfois être quelque peu réductionnistes). Des exemples de tels modèles sont issus du béhaviorisme ; comme on doit se contenter de chercher des régularités entre les entrées et les sorties, ces modèles ne sont descriptifs qu'« en moyenne ». Dans ce cadre, on peut éventuellement invalider des hypothèses, mais on ne peut jamais être certain de leur validité.

Mais si la clarté des outils mathématiques permet un accord objectif sur leurs résultats, il faut prendre garde à ne pas confondre la solidité d'un modèle avec sa pertinence comme projection d'une réalité. La démarche scientifique consiste en effet à montrer la cohérence d'un modèle sur des critères purement objectifs, formels et internes à la théorie, tandis que déduire, à partir du fonctionnement d'un modèle, des propriétés de la réalité est plutôt d'ordre pseudo-scientifique et reste aventureux !

Généralement, s'il existe une application φ-1 (figure 2), on considère que l'outil est aussi **prédictif** : des résultats au niveau du modèle permettent alors d'inférer certains états de la réalité, éventuellement non encore rencontrés. Le modèle est considéré comme valide quand ces inférences sont effectivement avérées. Ces modèles sont liés au structuralisme : chaque élément est décrit par rapport à l'ensemble ; les modèles sont alors statiques et explicatifs. Ils décrivent un état de choses, mais ont des difficultés pour en décrire les évolutions (ce qui signifie qu'ils ne sont pas forcément explicatifs). Les disciplines concernées ici sont essentiellement la psychologie cognitive qui souhaite des modèles prédictifs et cohérents avec les vérifications expérimentales, mais aussi permettant la compréhension du phénomène, tandis que la linguistique s'intéresse à des modèles prédictifs vérifiables expérimentalement (alors qu'elle s'intéresse rarement au phénomène lui-même).

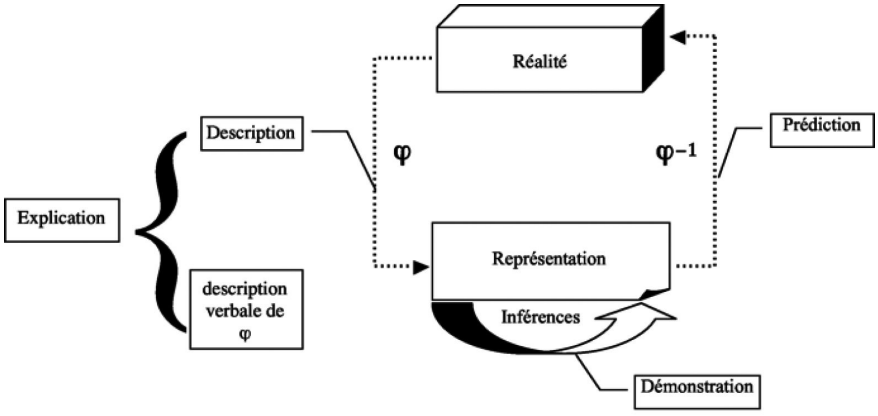

Figure 2 : Modèles et explication.

Les modèles purement formels rencontrent de grandes difficultés pour représenter les phénomènes complexes où existent d'innombrables interactions. En outre, une exigence forte introduite par l'IA porte sur le fait que les machines ont besoin de modèles non seulement pour décrire et prévoir, mais aussi pour agir. Un autre type de modélisation est alors nécessaire : le traitement effectif (il s'agit d'un modèle d'exécution, de fonctionnement du phénomène, fondé sur une analogie de comportement : on ne décrit pas entièrement le phénomène, mais on exprime sa dynamique de fonctionnement). La complexité se représente ainsi par du calcul effectif et non par la syntaxe du modèle.

Ces modèles proactifs permettent alors que l'entité qui se représente une certaine réalité prenne en considération ses actions sur cette réalité elle-même. Plus proches du constructivisme, ils permettent la prise en considération de l'observateur et des processus d'évolution des structures étudiées, et peuvent impliquer dans certains cas une certaine perte de l'objectivité. La notion de niveaux d'interprétation devient également primordiale : les interprétations d'un niveau dans les termes d'un niveau inférieur (réductionnisme…) ne sont pas compositionnelles mais émergentes[8].

[8] On pourra lire dans le paragraphe *Cardon*, p. 54, un résumé de ses réflexions approfondies sur ce sujet.

La *modélisation de phénomènes complexes* est ainsi fondamentalement différente des modélisations équationnelles :

- définition des éléments générateurs des mouvements de base du phénomène ;
- définition des règles de communication et de synchronisation entre ces éléments ;
- mise en mouvement des éléments de base (évolution temporelle du phénomène) ;
- observation de ces mouvements.

C'est essentiellement ce troisième type de modèle qui permet le développement de simulations sur machine ; on peut alors observer les résultats, les ajuster, et expérimenter à nouveau, selon une boucle *d'amorçage* absolument essentielle pour une véritable intelligence artificielle comme le souligne Pitrat [24-26].

Historique des travaux scientifiques liés à la conscience de l'homme

Pendant de longues années, les chercheurs en IA et en sciences cognitives se sont gardés d'aborder le thème de la conscience, qui apparaissait comme une notion trop vague pour permettre une étude scientifique et pour fonder la cognition. Puis, faute d'idée nouvelle, cette question resta longtemps en sommeil. Ainsi, Bertrand Russell prétendait que les résultats de l'introspection étaient scientifiquement inutilisables car n'obéissant pas aux lois physiques. De même, le behaviourisme, voulant fonder la psychologie comme science exacte, exclut toute notion d'état mental, et rejette ce qui concerne la conscience comme fondamentalement hors de son domaine.

Un renouveau de cette question semble dû à la théorie darwinienne de l'évolution bien qu'Eccles [27] se demande à ce sujet comment des organismes vivants ont acquis des expériences mentales – non matérielles – dans un monde autre que celui qui contenait alors tout ce qui existait ? (« *Conscience : "le cadavre dans le placard" de l'orthodoxie évolutionniste* »). D'un autre côté, le matérialisme orthodoxe (il n'y a pas d'esprit sans corps, sur lequel seules des entités physiques peuvent agir) débouche inévitablement sur la conclusion que l'homme est analogue à une machine. Les problèmes essentiels que pose cette conception sont alors

d'expliquer les sentiments, la conscience et le libre arbitre en se fondant uniquement sur les lois de la physique classique.

Pylyshyn [28] – suivi par Eckardt [29] – tente de fonder la science cognitive comme le domaine des perceptions et des connaissances avec un niveau de représentation où l'on fait abstraction des facteurs sociaux et des aspects émotionnels (ce cognitivisme classique est clairement explicité dans [30]). Ces hypothèses ont suscité des réactions hostiles et diverses remises en cause : Edelman argumente violemment à propos des affirmations non prouvées sur la structure du monde et les mécanismes de catégorisation que ces hypothèses supposent [5] ; il se fonde en particulier sur Rosch [31, 32] qui avait montré que le monde n'est pas structuré en catégories classiques, catégories que nos perceptions nous indiqueraient telles quelles. Par ailleurs, Searle estime scandaleux qu'une science qui se veut étudier l'esprit ignore les aspects liés à la conscience [33].

Ainsi, la « science cognitive », vue comme la science de l'esprit, ne peut négliger les facteurs sociaux, les aspects émotionnels et la conscience. Divers livres, sortis au début des années quatre-vingt-dix (Edelman, Rosenfield, Dennett, Varela et d'autres), se fondent sur l'idée commune que divers signes, biologiques et psychologiques, indiquent non seulement qu'une meilleure compréhension de la conscience est possible, mais qu'elle est nécessaire pour la compréhension de la cognition en général.

Dans un premier temps, nous donnons ci-après des résumés rapides des réflexions et propositions de Baars, Chalmers, Damasio, Dennett, Eccles, Edelman, Harth, Jackendoff, Jeannerod, Johnson-Laird, Maturana et Varela, Minsky, Ornstein, Penrose et Rosenfield (ordre alphabétique et références dans les paragraphes qui suivent). Tous sont essentiellement concernés par des modèles de la conscience humaine.

Dans un deuxième temps, après une présentation rapide de l'état de l'art des recherches sur une éventuelle conscience des machines, nous résumerons les modèles de Cardon, Pitrat et Sabah, qui constituent la base de la synthèse présentée ensuite. Par ailleurs, une analyse plus détaillée des modèles de Baars, Cardon, Eccles, Edelman et Sabah est disponible dans une annexe séparée.

[Même ici, nous restons loin de l'exhaustivité et on pourrait encore citer d'autres auteurs, par exemple : Benedetti, Block, Brentano, Carruthers, Changeux, Churchland, Cleeremans, Davidson, Harnad, Heidegger, Holland, Husserl, Marchetti, Pinker, Proust, Quine …, mais il faut bien se limiter.]

Baars

Baars [7] donne diverses caractéristiques d'une expérience consciente :

- implique une diffusion globale de l'information ;
- implique une cohérence interne (ce pourquoi elle est distincte du rêve) ;
- demande à être adaptée au reste du système ;
- demande accès par le système du soi ;
- peut nécessiter des perceptions d'une certaine durée.

Ce travail de psychologue développe une conception assez « économique » de la conscience : une zone de travail globale (proche du « théâtre cartésien » !), des processus inconscients spécialisés et des contextes (hiérarchies de buts) sont les trois seuls concepts essentiels de sa théorie que nous développons plus avant ci-après.

Chalmers

Chalmers [21, 34] est un philosophe qui cherche à construire les bases épistémologiques d'une théorie de la conscience en utilisant nombre de connaissances en neurosciences. Il postule que le matérialisme et le réductionnisme ont des limites (de l'ensemble des faits physiques, on ne peut dériver les faits des phénomènes de la conscience), et que réduire le cerveau humain à une machine biologique dont la conscience serait une conséquence ne pourra jamais expliquer ce qu'il appelle la « conscience phénoménale ». Il conclut alors à la nécessité d'un nouveau type de dualisme, permettant de rendre compte des facultés de la conscience que le matérialisme ne sait encore expliquer. Pour éviter le problème de l'interaction entre l'esprit et le corps, il défend un dualisme des propriétés et non (comme Descartes) un dualisme des substances.

Il a également formulé le problème difficile de la conscience : comment expliquer que nous avons des expériences phénoménales qualitatives. Opposé aux « problèmes faciles » (explications de nos capacités à assimiler des informations, à rendre compte d'états mentaux, de l'attention, etc.), ce problème persiste même quand toutes les fonctionnalités en question sont expliquées (voir ci-dessous Dennett, opposé à cette idée). Il peut se reformuler comme : pourquoi les qualia existent-ils ou pourquoi existe-t-il une sensation subjective de nos expériences ?

Damasio

Damasio [35] définit la conscience comme « la connaissance immédiate que possède un organisme de soi et de son environnement ». La conscience et les émotions sont inséparables : une perte de conscience entraîne une perte des émotions.

La conscience n'est pas organisée par secteur sensoriel et les émotions sont liées au sentiment de l'émotion et à la conscience de ce sentiment. Damasio exprime le concept de réflexivité en disant que la conscience de soi est comme l'observateur d'un film qui se verrait dans le film.

Il souligne l'importance des émotions pour le raisonnement et la conscience et définit la conscience-noyau comme essentiellement non verbale (mémoire à court terme, limitée) et la conscience-étendue comme fondée sur la mémoire à moyen et à long terme du Soi (condition préalable à l'intelligence) ; cette dernière est composée de deux espaces :

- un espace imagé, au sein duquel se produisent les images explicites ;
- un espace dispositionnel servant d'archives à long terme de ces images.

Dennett

Par certains côtés très proche de Baars, Dennett [36], bien que philosophe, se fonde sur des remarques de psychologie, de neurophysiologie et d'IA. Il critique fortement la notion de « théâtre cartésien » et propose pour la remplacer un mécanisme dans le cerveau permettant à des processus de « recruter » des processeurs d'interprétations et d'élaborations activés en parallèle (*pandémonium* de « versions multiples » *[Multiple Drafts]*). Il n'y a alors pas lieu de re-représenter les résultats ailleurs, à un centre de la conscience. La machine ne « rend conscient » que les résultats jugés pertinents au contexte, et par un effet de diffusion générale, permet à ce qui est conscient d'influencer n'importe quoi d'autre (*isotropie* – la pertinence est alors vue comme un phénomène un peu « magique » comparable aux heuristiques pas toujours bien justifiées de l'IA).

La conscience est ainsi vue comme une machine virtuelle sur le matériel parallèle du cerveau. Grâce au principe des machines virtuelles, on peut simuler n'importe quel processus parallèle sur une machine de von Neumann ; inversement, la question qu'il aborde est : comment la conscience simule-t-elle une machine de von Neumann sérielle sur l'architecture parallèle du cerveau ? (vu les caractéristiques

de lenteur et de sérialité des traitements conscients, on peut supposer que le cerveau n'est pas câblé *a priori* pour cela !). Il insiste sur le rôle essentiel du langage pour ce faire et voit l'apprentissage comme l'élaboration d'une telle machine virtuelle dans le cerveau.

Pour expliquer la production du langage, il imagine, à la base, des torrents de produits verbaux non contrôlés. Un aspect aléatoire préside ensuite au passage du non-conscient au conscient (ce qui explique la variabilité intra-individuelle, mais implique une absence quelque peu gênante d'intentionnalité !).

Il est fortement critiqué par Crick [37] ainsi que par Searle [38] (« un ensemble de conjectures philosophiques dont peu peuvent être réellement testées ») – en d'autres termes, une théorie que Max Planck aurait qualifiée de pas suffisamment spécifiée même pour être considérée comme fausse.

Eccles

Eccles [27] est pratiquement un des seuls (avec Popper) à se fonder sur la notion de dualisme.

Il distingue trois mondes (1 – matériel, 2 – états de conscience, 3 – connaissances objectives) qui entretiennent des relations récursives (le monde 2 crée le monde 3, mais est partiellement créé par lui par un processus de rétroaction). Il donne de nombreuses descriptions biologiques de fonctions élevées de l'esprit humain, sans qu'une idée précise de comment elles réalisent la conscience apparaisse clairement.

Parmi les propositions conceptuelles à la frontière entre recherche scientifique et philosophie, Eccles distingue toutefois l'hypothèse des *microsites*, fondés sur la physique quantique, par lesquels il tente de déterminer comment les événements mentaux agissent sur les événements neuraux : il propose d'identifier les interactions globales donnant lieu à des émergences comme des modes de résonance quantiques. Ils associeraient tout ou partie des microtubules où s'effectuent les transferts ioniques porteurs des informations traitées par le cerveau sans les limitations de Turing imposant des décompositions en sous-ensembles et interfaces explicites.

L'esprit conscient n'est pas seulement engagé dans la lecture des opérations du cerveau de façon passive, mais il possède une activité propre de recherche (attention, pulsion, nécessité…) et un rôle d'unification de l'ensemble.

Ses convictions religieuses l'amènent parfois à des affirmations sans la moindre démonstration convaincante comme par exemple : « il nous faut postuler l'existence d'une finalité dans toutes les vicissitudes de l'évolution biologique ».

Edelman

Edelman [4, 5] développe une vision de la conscience fondée sur une théorie des fonctions du cerveau à son tour fondée sur la thèse de l'évolution et du développement.

Le cœur de la théorie d'Edelman est la TSGN (théorie de la sélection des groupes de neurones), qui est fondée sur 3 principes : sélection ontogénétique ; renforcements ou affaiblissements synaptiques secondaires ; interaction de cartes cérébrales avec réentrance (échange de signaux entre répertoires différents de façon récursive et parallèle).

Il présente les fonctions du niveau neurobiologique qui, selon lui, ont permis l'apparition et l'évolution de caractéristiques de plus en plus élaborées de l'esprit humain, nécessaires et suffisantes pour la conscience : 1) spécialisations neurales permettant de distinguer les signaux internes des signaux du monde, 2) catégorisation perceptuelle, 3) mémoire(s) comme processus de recatégorisation continuelle avec possibilité de représentation de l'ordre d'activation, 4) apprentissage (liens entre les catégories précédentes et les valeurs essentielles de l'individu), 5) acquisition de concepts (catégorisation de patrons d'activité dans des correspondances globales), 6) **conscience** primaire (permet de connecter des états internes résultant de catégorisations perceptuelles antérieures avec les perceptions présentes : le présent remémoré), 7) capacité d'ordonnancement (résulte en une présyntaxe qui est la base de capacités symboliques), 8) langage et 9) **conscience** d'ordre supérieur.

Harth

Harth [3] rejette aussi bien le dualisme cartésien que le plus récent pluralisme radical (un million d'agents idiots au lieu d'un seul homoncule ingénieux). Pour lui, les images mentales ne sont pas des répliques d'objets du monde, elles sont combinées avec les connaissances antérieures et les résultats dispersés dans le cortex. Il n'y a pas d'homoncule qui examine l'état du cerveau, c'est le cerveau lui-même qui analyse et recrée, puis observe à nouveau ses propres créations (ce qu'il nomme *creative loop*).

Il part des caractéristiques suivantes : tout neurone a entrée(s) ET sortie(s) ; aucun n'est élément terminal d'une chaîne de traitement ; aucun neurone ne sait, ne comprend ou ne représente quelque chose ; ils sont également sensibles à des messages généraux envoyés partout (mécanisme qui serait responsable de l'humeur, du degré d'attention…).

Par ailleurs, il existe des chemins « descendants » qui modifient les messages des sens et injectent des informations supplémentaires, à la commande des niveaux supérieurs. Ce mécanisme d'autoréférence et ces boucles de rétroaction sont la règle dans les mécanismes du système nerveux. Le mécanisme de traitement est alors fondamentalement instable : par un processus d'amorçage, des variations infimes sont amplifiées et produisent des traits qui ne sont pas présents dans la perception initiale. Les zones centrales et périphériques du cerveau collaborent par ce mécanisme d'amorçage ; le message n'a pas besoin d'être lu par un homoncule quelconque, il se relit lui-même. Ce mécanisme apparaît au début du réseau sensoriel (et non à la fin comme le laisserait supposer l'idée qu'il s'agit d'une fonction avancée du cerveau).

En conséquence, pour Harth, la conscience se situerait au bas des processus de traitement, là où les résultats des traitements sensoriels forment encore un tout et préservent les relations spatiales de la scène originelle. Le reste du cerveau joue le rôle d'un observateur de ces résultats et influe sur eux pour maximiser la reconnaissance.

Jackendoff

Jackendoff [39] s'oppose en particulier à Johnson-Laird quant au haut niveau où se situerait la conscience. Pour lui, perception, action, pensée et apprentissage sont tous inconscients ; les contenus conscients sont des entités intermédiaires qui relèvent plutôt de la structure de surface des choses. La conscience est censée contenir les distinctions essentielles entre les choses (par exemple, différents mots reflètent différentes réalités, et ces distinctions sont projetées de la mémoire à court terme dans la conscience) [*sans qu'il explicite pourquoi ceci peut se passer comme cela…*].

Jeannerod

Jeannerod [40] centre ses recherches sur la conscience de soi (« connais-toi toi-même ») et aborde les questions du mécanisme sous-jacent à cette capacité

et de son utilité. Comme ce type de conscience porte sur un objet interne (introspection), elle se distingue de la conscience perceptive qui relève d'un ancrage objectif. Il aborde plus particulièrement la conscience de l'action qui est une forme particulière de la conscience de soi, mais qui, par l'action, porte sur un élément du monde extérieur.

Une distinction importante est que le responsable d'une action est à la fois auteur (aspect intemporel) et acteur (ancré dans l'instant). Le niveau de l'acteur fait intervenir la conscience du corps (sensations) plus « l'agentivité », et permet de distinguer deux niveaux : le soi minimal (automatique) et le soi conscient. Diverses expériences montrent qu'il y a une prise de conscience brutale au moment où le système automatique est saturé. L'exemple de la schizophrénie est une dissociation entre les rôles d'auteur et d'acteur. L'articulation entre les deux montre que l'action volontaire (réalisée ou non) commence par le *vouloir* (planification consciente en mémoire de travail), suivi par *l'action* (dans le domaine automatique : abandon [très court] de la conscience) et par l'*estimation* du résultat (conscience *a posteriori*). Mais tout cela n'est pas suffisant pour être la cause matérielle d'une action, il faut aussi une activation du réseau nerveux responsable du mouvement ; ce qui suggère l'existence de deux réseaux nerveux fonctionnant en parallèle : celui qui permet l'action et celui qui donne la conscience de l'avoir voulue.

Un rôle essentiel de la conscience serait alors d'éviter les hiatus et de maintenir la cohérence entre les idées, les croyances et le comportement (maintien de la *consonance cognitive* c'est-à-dire de l'équilibre du système cognitif). En cas de dissonance, on est amené à réviser ses choix et ses croyances, éventuellement à en abandonner, ce qui n'est possible que si on conserve son histoire (ce qui n'existe pas chez la plupart des animaux) et résulte en la notion de responsabilité (ce qui, de même, n'existe pas chez la plupart des animaux).

Johnson-Laird

Johnson-Laird présente un point de vue sur la conscience qui serait analogue à un système d'exploitation informatique. Dans [41], il suppose l'existence d'un moniteur de haut niveau pour gérer l'ensemble des processus inconscients qui agissent en parallèle dans l'esprit. Ce moniteur est chargé d'assigner des priorités relatives aux processus en attente, et a en plus la possibilité de gérer des interruptions

par l'intermédiaire de systèmes de sémaphores complexes[9]. Il considère alors la conscience comme un mode de fonctionnement particulier de ce système d'exploitation, supposé disposer d'un modèle de lui-même.

Maturana et Varela

Maturana et Varela [42] donnent des arguments pour la relativité de la réalité : la tache aveugle (nous ne voyons pas que nous ne voyons pas), les couleurs et les ombres (il y a des corrélations de noms de couleurs avec les états neuronaux mais pas avec les longueurs d'ondes). Nous ne voyons pas les couleurs du monde, nous construisons notre propre espace chromatique. Ils considèrent l'*autopoïèse* (capacité à s'autoproduire continuellement) comme caractéristique du vivant.

Il n'y a pas de *solipsisme* (le système nerveux interagit avec l'environnement, et ces interactions déclenchent des changements structuraux modulant la dynamique des états), ni de *représentationnalisme* (le système nerveux ne recueille pas d'informations de l'environnement, il spécifie quelles configurations de l'environnement constituent des perturbations et quels changements ces perturbations déclenchent dans l'organisme). Toute perception est vue comme une perturbation d'un état stabilisé du système nerveux ; le traitement de cette information consiste alors à rétablir l'équilibre.

Les communications sont des comportements coordonnés déclenchés mutuellement par les membres d'une société. La métaphore du canal de communication est fondamentalement fausse en ce qu'elle présuppose que ce qui arrive à un système est déterminé par l'agent perturbateur ; or c'est déterminé autant par l'état du receveur que par ce qui est émis.

Le langage est à la base de la conscience en ce qu'il permet de se décrire soi-même dans une situation donnée. C'est grâce à lui, et dans le langage même que le *soi* émerge. La conscience donne la possibilité de voir son propre comportement à la fois de façon interne et comme observateur.

[9] Cette idée semble assez extraordinaire puisque de tels systèmes d'exploitation sont utilisés depuis cinquante ans sur des ordinateurs sériels et parallèles, sans la moindre émergence d'une quelconque parcelle de conscience, même si cinquante ans est une période extrêmement courte au regard de l'évolution…

Minsky

Bien que ne voulant pas être un modèle de la conscience, la société de l'esprit [43] est concernée par cet aspect.

Pour Minsky, la conscience ne concerne que le passé et non le présent (comment penser à – et réorganiser – nos anciennes pensées). La connaissance est atomisée en un ensemble (gigantesque) d'agents très simples. En contexte, seul cet ensemble d'agents élémentaires interagissant explique notre comportement. Proche de Baars par certains côtés, cette position n'explique pas le rôle de la conscience dans la constitution de l'objectivité et dans les relations au monde extérieur. Elle constitue probablement aussi une origine des idées de Dennett.

Ornstein

Ornstein [44] présente un travail plutôt de l'ordre de la psychologie populaire. Il insiste sur les mécanismes de l'évolution et leurs implications sur le fonctionnement actuel de l'esprit et du cerveau. Il voit les capacités mentales de haut niveau comme pratiquement accidentelles (*serendipity*, terme anglais difficilement traduisible que l'on peut approcher par l'idée de « don de faire des trouvailles par hasard »).

Partant des capacités inconscientes présentes dans le cerveau, capacités qu'il appelle *simpletons*, il en envisage des escadrons qui passent rapidement d'un état contrôlé à un état non contrôlé. Les modifications et les évolutions conscientes du comportement passent alors par une compréhension profonde et une gestion de ce processus. Par certains côtés, sa théorie est proche de la « société de l'esprit » de Minsky (qui d'ailleurs n'est pas cité).

Penrose

À partir de réflexions sur le théorème d'incomplétude de Gödel [45] (et donc du fait que, dans certains domaines, on « sait » que certaines choses sont vraies sans avoir la possibilité de le démontrer), Penrose [46] argumente pour que ce type de décision soit du ressort de mécanismes réflexifs et en particulier de la conscience.

Rosenfield

Rosenfield [47] se fonde sur des recherches en neuropsychologie clinique, avec une perspective ethnographique, en considérant ses sujets non pas comme des patients avec des lésions, mais en étudiant comment ils luttent pour *construire du sens* à partir de leurs émotions et de leurs expériences physiques et sociales. Il donne une vision globale de ce mécanisme dynamique de construction du sens, en liaison avec la création consciente d'une vue intégratrice de l'image de nous-mêmes. Cette construction du sens implique de relier ce que l'on est en train de faire avec les expériences passées et ce que l'on attend pour l'avenir. De plus, la compréhension du monde est fondée sur – et émerge de – la dynamique des mouvements du corps.

Comme Edelman, Rosenfield défend l'idée d'un cerveau qui réorganise continûment et dynamiquement ses réponses aux divers stimuli : il donne un modèle global de la cognition fondée également sur la notion de catégorisation. Dans un style plus proche de la neurophysiologie cognitive, il insiste toutefois sur ce que les comportements anormaux révèlent à propos des fonctions normales. Les zones du cerveau s'organisent selon des circuits complexes, qui sont des généralisations impliquant des recatégorisations bi-directionnelles aux niveaux perceptuel, séquentiel, conceptuel et linguistique.

La catégorisation perceptuelle d'Edelman est le fait de donner des réponses cohérentes aux stimuli. Pour Rosenfield, cette propriété ne rend pas compte du fait que chaque catégorisation est une *relation* avec d'autres coordinations cohérentes. Ainsi, le sens d'un concept est contenu dans une relation fonctionnelle entre les processus neuraux en cours, eux-mêmes résultant de coordinations antérieures. Toute catégorisation est une relation dynamique entre des processus neuraux.

Il établit ainsi une relation étroite entre sens et conscience : être conscient, c'est s'engager dans un acte consistant à produire du sens. Or, il n'y a pas de compréhension du langage, pas de sens du temps, pas de personnalité sans une image de soi cohérente. La conscience est ainsi vue comme un mécanisme qui sous-tend une relation entre nos souvenirs et le sens actuel de nous-mêmes.

Enfin, le langage instaure un nouveau type d'autoréférence grâce auquel nous sommes explicitement conscients de nous-mêmes.

Historique des travaux scientifiques liés à la « conscience » des machines

La littérature concernant la conscience et la conscience humaine est extrêmement abondante, et on remarque de suite, même en ne la parcourant que partiellement, qu'on peut y distinguer deux sous-ensembles, en particulier relativement à nos préoccupations.

L'un, le plus imposant et à l'origine le plus ancien, relève d'une approche liée aux **sciences humaines**. En résumé, elle cherche à décrire et expliquer, *via* l'existence d'une conscience, l'origine, l'évolution et les comportements du vivant, l'ensemble débouchant sur l'existence d'une conscience. Les modèles de conscience qu'elle propose ne sont pas toujours directement exploitables par une technologie qui aurait l'ambition d'en construire une imitation robotisée. Avec ce point de vue, on peut citer Dehaene [48] qui s'intéresse à l'examen des mécanismes de l'accès à la conscience dans le cerveau humain, dans un cadre de psychologie cognitive expérimentale. Hesslow [49], quant à lui fonde sa réflexion philosophique sur la notion de simulation du comportement de l'homme dans un robot et Metzinger [50, 51], toujours avec une réflexion essentiellement philosophique, propose un modèle du soi comme processus et propose des idées générales pour juger de la présence d'une conscience dans une entité (animal ou machine). On trouve également des théories neurophysiologiques, qu'il faudrait travailler spécifiquement pour les adapter à une machine : Taylor [52] argumente pour donner un rôle central à l'attention dans la conscience humaine, et insiste sur l'importance de cette notion pour les machines. Avec un point de vue neuroscientifique de psychiatre, Tononi [53] propose une théorie de l'information intégrée pour la conscience et cherche à expliquer ses relations avec le sommeil et les rêves ; il utilise des modèles informatiques pour tester sa théorie.

L'autre sous-ensemble correspond à une approche qui relève de ce qu'on appelle les sciences dures. Elle part du principe que la conscience émane d'un cerveau, et que, percer les secrets physiologiques et biologiques du cerveau pourrait expliquer la présence ou l'apparition d'une conscience. Ainsi, de la même manière que l'on sait faire un cœur ou un rein artificiel, on pourrait penser à fabriquer un cerveau artificiel possédant les fonctionnalités importantes d'un cerveau naturel (dont, en particulier la génération de pensées et l'émergence

d'une conscience). Mais, actuellement, on se heurte à quelques problèmes non résolus : la complexité du cerveau que certains pensent irréductible, l'incertitude sur les liens entre cerveau et conscience, car le corps entier pourrait être impliqué, enfin l'absence d'une définition claire des propriétés du cerveau ou de la conscience (sachant de plus que la conscience englobe un inconscient) [54]. Pour aborder ce type de questions, Chella et Manzotti [55] proposent de construire un être artificiel et une expérience empirique afin de vérifier leur théorie de la conscience dans laquelle il n'y a pas de séparation entre le sujet et l'objet et où la réalité est constituée de processus répartis dans le temps et l'espace. Aleksander [56] veut également copier les êtres humains et suggère douze principes que devraient vérifier d'éventuelles consciences artificielles, alors que Haikonen [57] développe également un modèle de machine consciente qui cherche à reproduire de façon plausible la conscience humaine.

Sanz *et al.* [58], dans le domaine des systèmes multi-agents, proposent un modèle général des émotions, vues comme des processus de réorganisation du contrôle, dirigé par des valeurs (le problème principal étant de déterminer comment l'autonomie émerge de l'intégration des aspects cognitifs et émotionnels).

L'intelligence entretenant des liens étroits avec la conscience, il apparaît ainsi que l'approche fondée sur l'informatique et l'IA ouvre une voie prometteuse en matière de conscience artificielle. Le domaine est bien couvert par Chella et Manzotti [55], dont la recherche concerne la conscience des machines et les architectures cognitives pour les robots. Parmi les approches significatives de ce type de recherche, on peut citer Shanahan [59] qui propose une esquisse préliminaire pour rendre compte de la conscience réflexive, esquisse fondée sur une architecture implémentée qui combine un espace de travail global (*cf.* Baars) avec une boucle interne sensori-motrice fermée. Holland [60], lui, se penche sur la conscience de la machine dans un contexte fortement incarné et veut construire une structure de type squelette humain qui puisse interférer avec un monde réel, et ainsi se construire un modèle interne virtuel de ce monde et un modèle de sa propre interaction avec lui en tant que pensée consciente.

Finalement, Chrisley [61] considère que l'IA actuelle n'est pas sur le bon chemin ; pour réellement comprendre la conscience, il vaut mieux la considérer comme un moyen d'amplifier les capacités de l'homme que comme une technique permettant d'en implémenter une copie. Nos propositions ci-dessous permettent peut-être d'émettre un bémol vis-à-vis de cette opinion…

Cardon

Le livre d'Alain Cardon « *Modéliser et concevoir une machine pensante. Approche constructible de la conscience artificielle* » [62] relève d'une réflexion approfondie sur la question de l'intelligence et de la conscience. La construction méthodique qu'il propose vise à montrer le caractère automatisable de ces facultés, en principe réservées à l'Homme.

Parfois ardu, parfois manquant de justification (il s'agit d'une construction montante, dont l'aspect suffisant est généralement convaincant, sans que la nécessité des propositions avancées soit toujours évidente…), ce livre est tout à fait fascinant par son ampleur, courageux par les thèmes qu'il aborde et constructif par ses propositions concrètes et solides.

Les principales hypothèses qui fondent l'ensemble sont : une pensée artificielle est calculable ; *elle nécessite l'interaction physique d'un corps matériel avec son environnement.*

Pour justifier sa première affirmation, l'auteur met en avant divers arguments opposés aux trois raisons qui, à son avis, peuvent faire croire à l'impossibilité de mécaniser la pensée :

- les croyances religieuses et spirituelles ;
- le physique (par exemple si le continu de la matière et du temps était nécessaire) ;
- même si la pensée est constituée de processus discrets finis non continus, sa complexité pourrait être inatteignable pas les ordinateurs actuels.

La seconde hypothèse est argumentée en se référant à Damasio [35] (un organisme ne peut penser que s'il est d'abord capable d'éprouver sans cesse des émotions relatives à l'état et à la posture de son corps physique).

En résumant le modèle de départ de façon excessivement caricaturale, on peut dire que les composants de base sont des agents actifs, proactifs (capables, *de leur fait propre*, de mouvements et de communications), symboliques, évolutifs et communicants ; ils sont regroupés en ensembles et sous-ensembles (nommés *agents aspectuels*), eux-mêmes communicants, au fonctionnement automatique, inévitable.

Un autre ensemble d'éléments représente l'état et le fonctionnement des agents aspectuels (les agents *morphologiques*). Ces deux systèmes sont coactifs

et s'influencent réciproquement. La stabilisation de l'ensemble correspond à un « état de pensée ». La *conscience noyau* est définie comme la perception de soi-même en état d'appréhension du monde (*analogie forte avec la conscience de premier niveau d'Edelman*). La conscience étendue est alors la mise en situation de l'individu dans la temporalité.

L'objet de ce livre est de proposer une architecture informatique qui permettrait de développer un tel système. Dans ce dessein, est affirmée la nécessité d'un robot élaboré et d'un système de calcul massivement parallèle. Le livre se présente en quatre parties : des généralités sur les modèles et la complexité ; l'explicitation de relations entre pensée, émotion et conscience, et les réalisations informatiques pertinentes ; les modélisations des notions d'émotion artificielle et de « proto-soi » ; les modélisations de la pensée artificielle et du « soi ».

Pitrat

En s'inspirant de l'intelligence et de la conscience humaines, mais sans s'imposer de les copier, Pitrat développe CAIA (*chercheur artificiel en intelligence artificielle*) [13], un système de résolution de problèmes le plus efficace possible, expérimenté depuis plus de 20 ans. Il constate *a posteriori* que certains éléments sont analogues à certains composants de notre conscience, et que, souvent dans les cas de succès, ces mécanismes liés à la conscience, lorsqu'ils sont présents, sont nécessaires.

Ce système est fondé sur l'amorçage et sur la déclarativité des connaissances et des méta-connaissances, créant ainsi une cognition artificielle qui a des aspects très différents de la cognition humaine. La déclarativité des connaissances à tous les niveaux permet à CAIA, non seulement de raisonner par combinatoire (examen des valeurs possibles des variables du problème), mais aussi par méta-combinatoire : il raisonne sur les différentes méthodes possibles pouvant faire progresser la résolution, car, pour chaque procédure, il dispose d'informations (toujours déclaratives) sur ses conditions d'utilisation, les situations où elle n'est pas utile, les priorités qui précisent l'urgence de son utilisation. Il dispose donc de méta-méta-connaissances pour effectuer ces raisonnements (et celles-ci sont suffisamment générales pour qu'il n'y ait pas besoin de niveau supplémentaire). Ces différents éléments permettent au système de justifier les solutions qu'il propose, mais aussi d'indiquer comment il y est arrivé, pourquoi il a négligé tel élément…

Sabah

Le modèle CARAMEL (*conscience, automatismes, réflexivité et apprentissage pour un modèle de l'esprit et du langage*) [14, 15, 63] tire de nombreuses inspirations des travaux de Baars, Harth et Edelman. Un premier niveau traite les perceptions par des processus non contrôlés : une extension des tableaux noirs (le *carnet d'esquisses*) permet à de tels processus d'interagir « *inconsciemment* » en tenant compte de diverses rétroactions. Le deuxième niveau se fonde sur l'idée que la réflexivité et l'IA distribuée permettent le développement de programmes capables de représenter leurs propres actions et de raisonner sur ces représentations pour adapter dynamiquement leur comportement. Un modèle simpliste de conscience établit alors un lien entre ces deux niveaux de traitement par l'intermédiaire d'un modèle de mémoire comportant une mémoire à long terme (les connaissances du système), une mémoire de travail (où sont élaborés les résultats des processus conscients et inconscients) et une mémoire à court terme (où émergent les éléments qui deviennent conscients). Un avantage de ce modèle est que le niveau conscient peut se concentrer sur les tâches les plus adaptées à un traitement rationnel, les autres problèmes étant filtrés au niveau subliminaire.

Nous proposons ci-dessous une articulation des modèles de Cardon, Pitrat et Sabah (qui sont par ailleurs décrits plus en détail dans un des fichiers accessibles à l'adresse : http://gscns.free.fr).

VERS UN NOUVEAU MODÈLE

Proposition de synthèse

Dans ce modèle, l'entité interagit avec le monde par l'intermédiaire de ses perceptions et de ses aspects moteurs ; la « conscience » du système est en fait répartie dans les trois modèles précédemment cités. On introduit la notion de « valeur » qui est tout à fait centrale. Parmi les dizaines de significations potentielles de ce terme, on ne retient ici que le sens d'Edelman : *forces biologiques primaires* (ou besoins, comme, pour l'homme, le besoin d'alimentation, la reproduction ou la production d'adrénaline) autour desquelles se produisent implicitement des renforcements ou des inhibitions et qui servent de contraintes dans l'élaboration de comportements adéquats. En particulier, on ne se réfère absolument pas au sens moral. Les évolutions de ces valeurs engendrent des pulsions, qui sont des envies d'actions visant à rétablir des états convenables des valeurs. Transposé chez les machines, le concept de valeur recouvre leurs besoins essentiels (par exemple, besoin d'énergie, de lumière, et besoins spécifiques aux applications envisagées) ; il est central car, articulé avec l'expérience de l'entité, il permet, dans une situation donnée, l'adoption d'un comportement adéquat qui n'a pas été spécifiquement programmé. Les évolutions de ces valeurs selon les actions effectuées par la machine vont

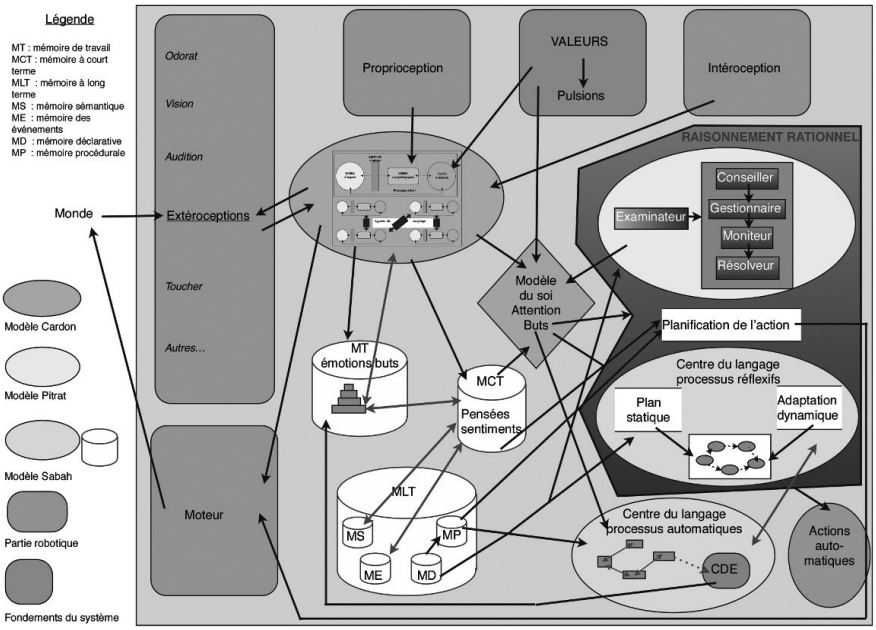

Figure 3 : Quelques relations entre les trois modèles (Cardon, Sabah, Pitrat) [toutes les interactions ne sont, bien sûr, pas représentées – cette figure a seulement pour but de donner un aperçu de la complexité à laquelle conduit une tentative d'implémentation informatique d'un modèle de conscience].

permettre des apprentissages (évolutions positives) ou des inhibitions (évolutions négatives)[10].

Le générateur de pensées de Cardon joue le rôle de frontal et traite les perceptions (données issues des capteurs) et produit des émotions (non conscientes, en mémoire de travail) et des pensées explicites (en mémoire à court terme) contextuellement pertinentes ; associées, les émotions et les pensées engendrent des sentiments et des buts contextuels qui, associés aux buts plus ou moins permanents

[10] Les machines modélisées ici sont destinées à acquérir une certaine autonomie. Il est clair que, si on souhaite en garder le contrôle total, il faudra une programmation telle que ces valeurs, fixées *a priori*, leurs soient inaccessibles afin qu'elles ne puissent les modifier à notre insu…

issus des valeurs constituent la pile des éléments à réaliser. Ces résultats sont, selon les buts à atteindre, fournis aux processus d'actions automatiques ou au carnet d'esquisses du modèle CARAMEL par l'intermédiaire des mémoires de travail (le niveau de la mémoire où sont écrits les résultats de Cardon dépend de leur « niveau de conscience » : totalement conscients, ils sont écrits dans la mémoire à court terme, sinon dans une des mémoires de travail « moins » consciente – c'est-à-dire demandant plus de ressources [temps et place] pour que le système puisse y accéder). Les données de la mémoire à court terme sont traitées par les processus attentionnels de CARAMEL, qui décident, selon le cas, de les envoyer à ses processus réflexifs ou aux règles de résolution de problèmes de Pitrat (figure 3).

Ce modèle pourrait fonctionner tel quel en utilisant les procédures de traitement automatique des langues et les données linguistiques disponibles dans le modèle CARAMEL. Mais, en toute rigueur, il faut également considérer la situation où le système démarre sans disposer de données sémantiques ni pragmatiques (concepts et symboles), mais ait le potentiel de les acquérir. En introduisant la vision d'Edelman, dont le modèle que nous venons de présenter permet de rendre compte, on peut expliquer la morphogenèse de ces capacités.

Tout d'abord, grâce à un processus de réentrance, l'interaction entre cartes construites au niveau perceptif débouche sur un mécanisme de catégorisation perceptive : la constatation de coïncidences entre un système perceptif, le système intéroceptif et le système de valeurs (une valeur évolue de façon significative, positivement ou négativement)[11] permet de catégoriser les perceptions significatives.

Ensuite, en intégrant différentes modalités, la catégorisation des corrélations entre perceptions significatives par rapport aux valeurs et la catégorisation pragmatique (prise en compte des effets sur le monde extérieur et des rétroactions venant de celui-ci) permettent l'émergence de concepts (figure 4).

La catégorisation de ces premières étapes débouche sur la capacité à créer des *scènes* (images mentales dans le présent, rappelant d'autres événements passés) et à les réactiver par des perceptions partielles, qui deviennent ainsi des *signes* : c'est la conscience primaire.

[11] L'intervention de ces valeurs plus ou moins hiérarchisées (survie, faim, soif, désirs…) est ici essentielle : elle permet de concevoir des convergences des divers processus proposés en optimisant ces critères, sans qu'aucune téléologie ait besoin d'intervenir.

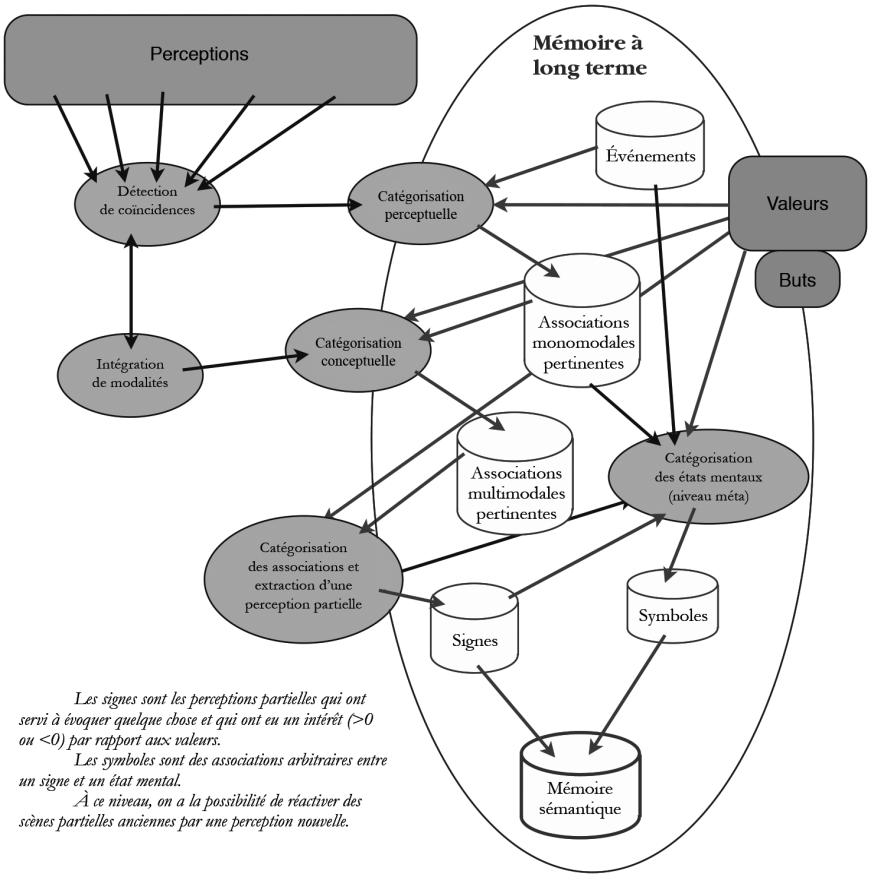

Figure 4 : Mécanismes d'émergence de connaissances sémantiques à partir des perceptions. Cette mémoire sémantique et les symboles permettent de faire émerger une syntaxe et un langage, ce qui l'articule avec la figure 1, p.35.

Ce type de processus récursivement appliqué au fonctionnement cérébral lui-même (construction d'une cartographie de divers types de cartes) permet à l'entité de catégoriser ses propres activités, ce qui résulte en la notion d'état mental. Les capacités symboliques – préalable nécessaire à l'apparition du langage – consistent alors à lier plus ou moins arbitrairement un signe à un état mental et à utiliser ce signe dans des processus de raisonnement, de planification et de communication. Si au départ, une image mentale est forcément rappelée par un

signal du monde actuel, le mécanisme de catégorisation, appliqué à nouveau aux processus de la conscience primaire elle-même, va permettre de se libérer de ce lien au présent : l'état mental résultant d'une perception va pouvoir lui-même être associé à un autre état mental, sans que les caractéristiques du présent soient pertinentes pour cette association. Ces symboles permettent ensuite un *amorçage sémantique* débouchant sur une mémoire sémantique détachée des contraintes du présent.

Utilisées dans des bouclages récursifs, des catégorisations ultérieures (non représentées dans la figure) permettent alors l'émergence d'un lexique et d'un embryon de syntaxe (ordonnancement simple). L'analyse et la compréhension des effets pragmatiques de ces éléments permettent à un langage de plus en plus élaboré de se développer : au fur et à mesure de cette acquisition, les mots et les phrases eux-mêmes deviennent des symboles pour les concepts et peuvent être utilisés de façon réflexive. Par l'application de catégorisations et d'ordonnancements à ces éléments, une vraie syntaxe peut alors apparaître.

Cette nouvelle étape débouche sur la création des concepts de *soi*, de *passé* et d'*avenir* (permettant d'établir des différences entre un modèle conceptuel symbolique et l'expérience perceptive en cours). Ce langage va permettre le développement de nouveaux concepts et de symboles, s'enrichissant mutuellement : avec un lexique suffisamment développé, les productions langagières seront traitées et classifiées récursivement par l'appareil conceptuel, sans autre référence à leurs origines (en particulier leurs bases perceptives et sociales).

Fonctionnalités que réaliserait un tel système

Par rapport aux fonctionnalités explicitées au paragraphe *Fonctionnalités attachées à la notion de conscience*, on peut considérer que le modèle proposé pourra réaliser les fonctionnalités suivantes :

- interprétation et unification des données des sens ;
- représentation et interprétation de l'environnement par rapport à soi ;
- *partiellement* : fonction de constitution et de perception de soi *(une « théorie de l'esprit » générale semble encore hors de portée)* ;
- effet de séquentialisation et de sélection ;

- *partiellement* : contrôle de soi et de ses pensées. Choisir des comportements appropriés ;
- ordonnancement des buts ;
- *partiellement* : réactions face à l'imprévu ;
- résolution de problèmes explicites ;
- gestion d'hypothèses, planification ;
- déclenchement de processus planifiés ;
- contrôle de l'exécution des buts ;
- gestion de la mémoire à long terme, de la mémoire de travail et de la mémoire à court terme ;
- gestion de l'attention volontaire ;
- rendre accessible de l'information pertinente ;
- *partiellement* : gestion de l'attention involontaire ;
- gestion de l'apprentissage volontaire ;
- prédire ses propres comportements ;
- *partiellement* : permettre une communication sociale efficace.

Quelle validation pour un tel modèle ?

Outre douze propriétés très générales qu'une machine doit posséder pour « être consciente » [56], et afin de fournir une définition minimale de la conscience, Aleksander et Dunmall [64] proposent un ensemble de 5 axiomes permettant d'établir les fonctionnalités minimales de conscience que doit nécessairement implémenter un agent informatique pour qu'on le considère comme conscient :

a) un agent doit être pourvu d'états perceptuels lui permettant de se représenter son environnement ;
b) un agent doit être capable de sélectionner quelles parties de l'environnement il doit se représenter (focalisation de l'attention) ;
c) un agent doit être capable d'imaginer des éléments non perçus (nécessité de raisonnements abductifs ou inductifs) ;
d) un agent doit avoir un moyen de contrôle sur les éléments imaginés afin de planifier des actions ;

e) un agent doit posséder des états affectifs qui évaluent les actions planifiées et déterminent l'action à sélectionner.

Majid Beshkar [65] quant à lui, propose quelques règles pour détecter des éléments de conscience chez les animaux :

f) absence de réactions stéréotypées (adaptation du comportement à des situations nouvelles et imprévisibles) ;
g) reconnaissance de soi dans un miroir (expérience de la tâche sur une partie non visible du corps ;
h) capacité à communiquer de l'information sémantique (alarmes différentes selon le prédateur perçu) ;
i) possibilité de mensonge et de tromperie ;
j) avoir une mémoire des épisodes passés ;
k) avoir des émotions (évident chez diverses espèces pour des émotions comme la peur, la joie, la jalousie, la peine, le désespoir…)

Il doit toutefois être clair ici qu'il ne s'agit pas de règles strictes et indiscutables, mais que notre interprétation joue un grand rôle dans ces évaluations (il faut se garder de l'anthropomorphisme).

Revenons à notre proposition : le modèle nous semble capable de répondre positivement aux a) e) f) g) h) j) et k). Il devrait pouvoir réaliser partiellement b) c) et d). Quant à i), la programmation devrait être telle que cela lui soit interdit !
Cela nous laisse penser toutefois qu'il aurait des facultés permettant de lui attribuer un certain nombre, non négligeable, de facultés liées à la conscience.

Pour l'instant, cela relève plus de notre intuition que d'une démonstration formelle, et des expériences effectives nombreuses doivent être réalisées afin d'évaluer réellement ces possibilités, même si une validation totale et formelle de ces capacités restera probablement toujours hors de notre portée… En effet, on ne dispose d'aucune description exacte du concept de conscience ; le modèle proposé ne reflète donc que le point de vue (subjectif !) de notre conscience. Sa qualité ne pourra donc être évaluée par un autre scientifique qu'en confrontant les principes sous-jacents au modèle (ou sa spécification si elle est disponible) à sa propre intuition de la notion de conscience.

Par ailleurs, en admettant qu'on ait muni une machine d'un tel modèle, on ne pourrait guère affirmer qu'alors, la machine serait dotée d'une forme de conscience voisine de celle de l'homme. « *Étant donné le caractère subjectif de l'expérience de conscience, seule la machine serait capable d'une telle affirmation ! …* » [66].

APPLICATIONS

Les robots ont évolué du robot industriel préprogrammé effectuant une tâche figée au robot commandé à distance ayant des possibilités d'adaptation dynamique, mais toujours sous contrôle d'un opérateur humain. On envisage maintenant des robots autonomes dotés de capacités d'intelligence suffisantes pour survivre dans un environnement inconnu et changeant. Le « robot conscient » sera encore plus autonome, grâce à ses facultés d'auto-observation, d'apprentissage et d'autoréorganisation.

Ainsi, ce type de robot aura de plus en plus d'applications pratiques. Sans prétendre à l'exhaustivité, ni mentionner les améliorations des applications existantes, citons quelques-unes des applications nouvelles envisageables.

Outre les missions humanitaires en cas de tsunamis, d'inondations, de tremblements de terre…, les interventions dans des milieux hostiles à l'homme seront facilitées (ce qui est déjà le cas dans les centrales atomiques) et aideront également à l'exploration de l'espace (réparations et actions préparatoires à l'arrivée d'équipages humains).

Les robots domestiques, autonomes et interagissants, ont également un grand avenir : robot-nounou ou robot-garde-malade ont des utilités manifestes ; en mêlant jeu et éducation, ils auront également une influence importante sur les jeunes.

En liaison avec l'industrie des logiciels, les biotechnologies et les nanotechnologies (certains voient d'ailleurs de grands dangers avec ces dernières [67]), les applications seront nombreuses avec des enjeux industriels et économiques

considérables : par exemple, des industriels japonais visent déjà, pour les dix ans à venir, le remplacement de TOUS les opérateurs humains sur leurs chaînes de production, par des robots autonomes ! En collaboration avec la biologie et la médecine, nombre d'applications thérapeutiques et prothétiques pourront voir le jour, sans parler de la chirurgie.

Enfin les robots militaires, s'ils sauvent des vies et sont indéniablement utiles pour le déminage d'objets suspects, les patrouilles de reconnaissance et de surveillance, le transport de matériel, présentent aussi, bien entendu, des risques de dérapage qu'il sera essentiel de contrôler.

Mais, outre ces applications d'ordre pratique, il faut voir aussi les conséquences scientifiques et fondamentales de ces recherches : elles aideront à mieux comprendre quelques-uns des mécanismes qui sous-tendent l'intelligence et les différents aspects de la conscience. Par là même, elles approfondiront les mécanismes du langage, de l'apprentissage, de l'acquisition de connaissances, fondamentaux pour l'émergence de la conscience. Par ailleurs, l'émergence à côté de l'homme d'entités dotées d'intelligences supérieures, mais aussi de formes de conscience nécessairement différentes, posera des problèmes de cohabitation et d'optimisation. Comment ces robots et les humains (plus généralement les systèmes vivants) pourront-ils coopérer ? Comment les humains et les robots pourront-ils au mieux profiter de cette cohabitation pour s'améliorer ?

RECOMMANDATIONS

Recommandations générales

Aux États-Unis et au Japon, des thématiques comme la robotique autonome et la vie artificielle jouissent de moyens importants en chercheurs et en crédits[12]. En conséquence, ces pays voient se développer nombre de débats sur les conséquences économiques et sociales (emploi, investissements, projets) de ces nouveaux développements.

En Europe en général, et en France en particulier, il serait raisonnable de préparer la reconversion des personnes qui seront affectées par cette nouvelle robotique, et d'adapter l'appareil d'enseignement et de formation professionnelle.

Sur le plan scientifique, un programme de recherche visant la mise en œuvre de machines intelligentes et conscientes aurait le grand avantage de permettre la collaboration des sciences et techniques considérées comme stratégiquement essentielles pour l'avenir : l'informatique, l'intelligence artificielle et la robotique, bien entendu, mais aussi les nanosciences et les nanotechnologies, ainsi que les

[12] Aux États-Unis, essentiellement pour la défense et au Japon pour la robotique civile et de loisirs (voir par exemple http://www.consciousentities.com/brittle.htm http://www.human-evolution.org/newsaiarchive.php)

neurosciences et la physiologie intégratives, la génétique et les sciences de la cognition. Actuellement, en France et en Europe, de nombreux laboratoires s'attaquent à ces thèmes de recherche, mais en ordre dispersé, et ne franchissent guère le stade du produit de laboratoire. Depuis 2010, l'Agence nationale de la recherche coordonne le projet européen Era-Net CHIST-ERA[13], destiné à renforcer les collaborations internationales de recherche multidisciplinaire dites de rupture dans le champ des sciences et technologies de l'information et de la communication. Le premier appel portait sur les deux thèmes :

i) Quantum Information Foundations and Technologies

ii) Beyond Autonomic Systems - the Challenge of Consciousness

et AUCUN des projets retenus ne concerne le second !

La question de savoir comment remédier à une telle dispersion et les possibilités de mise en place de grands projets, nationaux ou européens, devraient être discutées au niveau scientifique comme au niveau politique. Un cadre de recherche, d'implémentation et d'industrialisation devrait être précisé, avec une pérennité et des budgets garantis (on peut penser à quelques dizaines de millions d'euros pendant 10 ans…). Et bien entendu le suivi éthique de ces recherches sera essentiel.

Suites souhaitables à donner à ces travaux

Les points suivants devront être abordés dans le cadre d'une suite du GT, et, sans doute, en relation avec la commission « Technologies et société » ou la commission d'*Éthique* de l'Académie. Par ailleurs divers documents portant sur les applications de systèmes intelligents et « conscients » sont en cours d'élaboration (par exemple : applications au domaine militaire, à l'aérospatial, aux robots domestiques…).

[13] http://www.chistera.eu/

Aspects éthiques

Quelle vision est transmise au grand public sur ces questions, par exemple par des films comme Matrix, Terminator… Quelle est la réalité des conceptions présentées dans ces films ?

Quelles capacités seront probablement atteintes par les « machines intelligentes » dans ce domaine ? Quels seront les avantages de tels systèmes ? Si une machine consciente semble réalisable à moyen ou long terme, quels sont les risques prévisibles et comment les pallier ?

Quelles conséquences sur le comportement de l'homme ? Cela favorisera-t-il une voie vers un *homme nouveau* (transhumanisme ?) ?

Si la notion de conscience est vue, chez l'être humain, comme la capacité à penser un sens personnel de son existence, qu'il utilise comme fondement régulateur de ses choix moraux, la notion d'intention devient cruciale et débouche sur diverses sous-questions comme : comment détecter les intentions humaines dans le cadre d'une relation homme-homme ou homme-machine ? De quel libre arbitre disposons-nous ? Quel libre arbitre reconnaîtrons-nous à des artefacts ? Pourront-ils être poursuivis ou ester en justice, au pénal ou au civil ? À quoi pourra-t-on bien les condamner ?

Statut juridique d'un éventuel robot intelligent et conscient

Un robot programmé pour être intelligent et « conscient » est-il responsable de ses actes et des dommages qu'il peut causer ou est-ce son constructeur ? Son propriétaire ? (Le responsable est celui qui peut indemniser la victime d'un dommage qu'il a créé.)

Sur ce point, il convient de distinguer deux cas, par rapport au mécanisme d'apprentissage sous-jacent (ce ne sont pas seulement les gestes dangereux qui doivent être pourvus de « responsables » mais les modules d'interprétation en réception d'apprentissage).

A) Les robots avec apprentissage externe

Il s'agit de robots « conscients en un certain sens » avec un logiciel non réentrant. Le processus d'apprentissage peut être remplacé par l'insertion d'un module ayant effectué son apprentissage chez le constructeur, ou simplement par l'insertion des connaissances qui en résultent. Ils sont donc testables juridiquement par simple analyse de fichier ou seulement en opération en mettant leur système dans un autre robot identique, ce qui permet de vérifier leur conformité à des clauses de réception à partir de tests définis.

B) Les robots évolutifs

Il s'agit de robots « conscients en fonction d'un passé et d'un présent d'apprentissage » qui ne peuvent être testés hors de la connaissance de leur « vie antérieure » (testant leur réceptivité dans un environnement de tests passés et présents).

Comme l'évolution de leurs connaissances est excessivement difficile à prévoir, ils ne peuvent être testés *a priori*. Il faudra probablement qu'ils le soient *a posteriori*, car de cela dépend la sécurisation des programme d'auto-apprentissage des robots et leur qualification. Mais, ces vérifications resteront bien évidemment extrêmement complexes.

Types d'intégration dans la société ?

Des robots babysitteurs ou protecteurs de personnes âgées ou de SDF aux robots anti-manifs ou aux robots militaires ? Quelles contraintes ? Quelles conséquences ?

Quels types de communications seront possibles entre les hommes et de tels robots (on pense à la distinction entre discours scientifique et parole poétique, introduite par Dominique Peccoud [68]). Un robot sera-t-il capable d'apprécier un langage symbolique ? D'en produire ?

En outre, les expérimentations sur des processus complexes, adaptatifs et apprenants devront probablement être régulées. L'idée que la vie (même simulée) puisse émerger de la complexité n'est pas encore répandue, et il y a beaucoup

moins de méfiance ou de consensus à propos des expériences sur des systèmes complexes artificiels que sur le vivant ou les OGM !

Quels marchés ? Quelle importance pour l'économie ?

Si des machines « intelligentes et conscientes » voient le jour, elles vont déboucher sur un marché considérable. Des forces importantes sont déjà investies (Corée, Japon) pour préparer ces nouvelles générations de machines. Les retombées à moyen et long terme doivent être prévues, en Europe également : il faut déjà des investissements importants dans ces domaines, sans profits immédiats ; la recherche publique fondamentale doit ici jouer un rôle essentiel.

Influence sur l'emploi et la formation

Le développement de l'informatique en général, et de l'IA en particulier, implique la disparition de certains emplois (souvent peu qualifiés), mais il en crée de nombreux nouveaux (qualifiés ou très qualifiés) après un certain temps de latence. Il serait donc pertinent de développer les formations qui seront nécessaires, sans en attendre l'émergence effective, afin de permettre une reconversion efficace des personnes concernées dans des domaines nécessitant moins de compétences, peut-être dans les industries de l'environnement et du développement durable.

On a déjà été témoin de ce problème de reconversion lorsque les robots industriels sont apparus dans les années 1980 et que d'aucuns craignaient un accroissement du chômage issu du remplacement des ouvriers par ces machines. Il n'en a rien été, car la productivité, et donc la richesse globale, a augmenté et de nouvelles fonctions pour les travailleurs ont émergé. Le type de mutation auxquelles on va devoir faire face va être bien plus important et nécessite sans doute une réflexion prospective. Comment seront intégrés les robots conscients dans la communauté des hommes et femmes au travail ? Comment redistribuer la valeur ajoutée produite par des robots ou des systèmes robotiques intelligents

pour éviter que soient accrues les cohortes de personnes sans moyens de subsistance ? Plus largement, sur quelle base, autre que celle du travail rémunéré, construire la cohésion sociale ?

Tous ces points ne sont ici qu'évoqués car, comme nous l'avons dit, ils constitueront les sujets de réflexion de la suite du Groupe de Travail.

CONCLUSION

Jacques Pitrat nous donne une vision optimiste de l'IA : l'homme n'est peut-être pas assez intelligent pour concevoir et mettre en œuvre une véritable IA, mais un des buts de l'IA est de construire des machines qui, grâce à des mécanismes d'amorçage, apprennent elles-mêmes à devenir plus intelligentes ! (*cf.* la technologie des ordinateurs actuels, impossibles sans la génération précédente, qui constitue une forme de mécanisme d'amorçage des machines présentes et ceci de manière récursive).

Ainsi, il nous semble probable que l'ingénierie évolutionnaire et les systèmes hybrides (connexionnistes et symboliques) pourront permettre l'émergence de systèmes auto-adaptatifs complexes ayant des capacités de mémoire et de raisonnement aussi complexes que l'être humain d'aujourd'hui. Associés à la notion de conscience évoquée ci-dessus et à des capacités d'auto-apprentissage, ces systèmes, opérant à une vitesse bien supérieure à celle des neurones, pourraient même dépasser notre intelligence … C'est ce phénomène de comparaison entre l'intelligence des machines et celle de l'homme que certains nomment « singularité » [69, 70]. Vu la structure d'un tel système, on n'en comprend pas en détail tout le fonctionnement et on ne pourra donc prévoir toutes ses actions[14]. Si cela devient

[14] On pourrait toutefois se demander si ce comportement n'est pas dû à une loi mathématique implicite dans le programme ? Les mathématiciens devraient donc prendre ces questions à bras le corps, ce qui n'est pas vraiment le cas.

possible, cela se fera sûrement, vu les intérêts économiques en jeu. Actuellement, l'absence de débats sur ce point s'explique par l'état de la technologie actuelle, mais si on attend l'apparition effective de telles machines, il sera trop tard.

Les « trois lois de la robotique » d'Asimov[15] sont pertinentes, mais sont-elles suffisantes et réalistes ? Une véritable intelligence est-elle compatible avec une obéissance totale ? Asimov lui-même montre d'ailleurs les limites de ses lois. Considérées parfois comme naïves, ces lois semblent refléter, compte tenu de l'époque où elles ont été écrites, le seul point de vue de la programmation classique (de type von Neumann) d'un tel système. Nous avons vu qu'un système réellement intelligent est plutôt de l'ordre du processus distribué et insaisissable. Une première conséquence simple est que l'applicabilité de tout type de règles (comme celles d'Asimov) est indécidable, au sens mathématique du terme. Les risques liés à l'émergence de telles machines sont donc aussi grands qu'ils sont difficiles à prévoir ; il faut approfondir scientifiquement le lien qui peut exister entre le domaine de la conscience et celui des systèmes complexes.

Quelles limites légales doit-on fixer, par exemple, à l'intrusion d'un robot-nounou, ou d'un robot-garde-malade ? Actuellement, il n'y en a aucune. Quelles interdictions peut-on envisager pour les robots militaires ?

En outre, ces risques existent, mais ils ne se limitent pas à la robotique. Le chemin actuellement suivi par l'espèce humaine montre qu'elle est en train de modifier non seulement l'environnement terrestre, mais qu'elle a la possibilité de modifier aussi ses propres caractéristiques biologiques, notamment génétiques et épigénétiques. Elle peut modifier également, sans faire encore appel à l'IA, mais simplement avec de simples drogues (ou la TV), la façon dont fonctionnent les cerveaux. Personne n'est en mesure aujourd'hui de pronostiquer les effets à court terme ou à long terme de ces différents développements. Faut-il envisager

[15] a- un robot ne peut agresser un être humain, ni, restant passif, permettre qu'un être humain soit exposé au danger ;
b- un robot doit obéir aux ordres que lui donne un humain, sauf si de tels ordres entrent en conflit avec la 1re loi ;
c- un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la 1re ou la 2e loi ; Loi Zéro ajoutée plus tard : « Un robot ne peut pas faire de mal à l'humanité, ni, par son inaction, permettre que l'humanité soit blessée. »

des catastrophes plus ou moins étendues, touchant tous les écosystèmes ? Peut-on au contraire imaginer qu'une humanité renouvelée, parfois qualifiée de posthumaine ou transhumaine, puisse en émerger ? Dans l'incertitude, faudrait-il appliquer un principe de précaution radical en interdisant toutes ces recherches ? Mais qui pourrait ou qui voudrait procéder à de telles interdictions, à supposer qu'elles aient quelques chances de succès ?

BIBLIOGRAPHIE

1. Block, N., Flanagan, O. and Güzeldere, G. (1997) *The nature of consciousness*, A Bradford Book, The MIT Press.
2. Chalmers, D. and Bourget, D. (2009) *A bibliography of the philosophy and the science of consciousness*, Mindpapers, http://consc.net/mindpapers, Australian National University.
3. Harth, E. (1993) *The creative loop ; how the brain makes a mind*, Addison-Wesley, New York.
4. Edelman, G. (1989) *The remembered present : a biological theory of consciousness*, Basic Books, New York.
5. Edelman, G. (1992) *Biologie de la conscience*, Editions Odile Jacob, Paris.
6. Edelman, G. (2004) *Wider Than the Sky : The Phenomenal Gift of Consciousness*, Yale University Press, New Haven and London.
7. Baars, B. (1998) *A Cognitive Theory of Consciousness*, Cambridge University Press, Cambridge.
8. Baars, B., Ramamurthy, U. and Franklin, S. (2007) *How deliberate, spontaneous and unwanted memories emerge in a computational model of consciousness*, Blackwell, Oxford, UK.
9. Dennett, D. (1991) *Consciousness Explained*, Little, Brown and Company, Boston.
10. Dennett, D. (1998) *Brainchildren : Essays on Designing Minds*, MIT Press, Cambridge, Mass.
11. Varela, F. (1996) *Neurophenomenology : a methodological remedy for the hard problem*, *Journal of Consciousness Studies* **3**, 4, 330–349.
12. Damasio, A. R. (1999) *The Feeling of What Happens : Body and Emotion in the Making of Consciousness*, Harcourt Brace and Co, New York.

13. Pitrat, J. (2009) *Artificial Beings : The Conscience of a Conscious Machine*, Wiley-ISTE, London.
14. Sabah, G. (1990) *CARAMEL : A computational model of natural language understanding using a parallel implementation*, Proceedings of 9th European Conference on Artificial Intelligence (ECAI-90), Stockholm, 6–8 August, pp. 563–565, Pitman, London/Boston.
15. Sabah, G. and Briffault, X. (1993) *Caramel : a Step towards Reflexion in Natural Language Understanding systems*, Proceedings of IEEE 5th International Conference on Tools with Artificial Intelligence (ICTAI '93), Boston, 8–11 November, pp. 258–265, IEEE Computer Society Press, Washington.
16. Poe, E. A. (1849 / 2008) *X-ing a paragrab*, Flag of Our Union / Quill Pen Classics, Boston.
17. Dickens, C. (1853) *Bleak House*, Bradbury & Evans (Penguin Classics, 2003 paperback), Whitefriars.
18. Coste, P. (1700) *Essai sur l'entendement humain*, (traduction de *"An Essay Concerning Human Understanding"*, John Locke, *1689) ; réédition* 2009, Le livre de poche, Paris.
19. Freud, S. (1915) *(réédition 2010) Métapsychologie*, Presses Universitaires de France, Paris.
20. Sperry, R. (1974) *Lateral specialization in the surgically separated hemispheres*, In Schmitt, F. and Worden, F. (Eds.), Third Neurosciences Study Program, MIT Press, Cambridge.
21. Chalmers, D. (1996) *The Conscious Mind : in Search of a Fundamental Theory*, Oxford University Press, Oxford.
22. Dennett, D. (2002) *How could I be wrong ? How wrong could I be ?*, *Journal of Consciousness Studies* **9**, 5, 13–16.
23. Wittgenstein, L. (1961) *Tractatus*, Logico-philosophicus, Routledge and Kegan Paul, London.
24. Pitrat, J. (1984) *Maciste : un système qui utilise des connaissances pour utiliser des connaissances*, GR 22.
25. Pitrat, J. (1986) *Les systèmes qui s'observent*, Proceedings of « colloque Logique naturelle et argumentation », Royaumont.
26. Pitrat, J. (1990) *Méta-connaissances*, Futur de l'intelligence artificielle, Hermès, Paris.
27. Eccles, J. (1992) *Évolution du cerveau et création de la conscience*, Fayard.
28. Pylyshyn, Z. (1986) *Computation and cognition : Toward a foundation fo Cognitive Science*, MIT Press, Cambridge.
29. Eckardt, B. v. (1993) *What is Cognitive Science ?*, MIT Press, Cambridge.
30. Gardner, H. (1985) *The mind's new science*, a History of the Cognitive Revolution, Basic Books, New York.
31. Rosch, E. (1975) *Cognitive representations of semantic categories*, *Journal of experimental psychology : general* **104**, 192–233.

32. Rosch, E. (1977) *Human categorization*, In Warren, N. (Ed.) Studies in Cross-Cultural Psychology, pp. 1–49, Academic Press, New York.
33. Searle, J. R. (1992) *The rediscovery of Mind*, Cambridge University Press, Cambridge.
34. Chalmers, D. (2004) *How can we construct a science of consciousness ?* In Gazzaniga (Ed.) The Cognitive Neurosciences III., MIT Press, Cambridge MA.
35. Damasio, A. R. (1999) *Le Sentiment même de soi : corps - émotions - conscience*, Odile Jacob, Paris.
36. Dennett, D. (1993) *La conscience expliquée*, Odile Jacob, Paris (1991 - Consciousness Explained - Penguin Books).
37. Crick, F. (1995) *L'hypothèse stupéfiante : à la recherche scientifique de l'âme*, Plon, Paris.
38. Searle, J. R. (1996) *Deux biologistes et un physicien en quête de l'âme : Crick, Penrose et Edelman*, *La Recherche* **287**, 62–77.
39. Jackendoff, R. (1987) *Consciousness and the Computational Mind*, MIT Press, Cambridge, Mass.
40. Jeannerod, M. (2009) *Le cerveau volontaire*, Odile Jacob ; Collection Sciences, Paris.
41. Johnson-Laird, P. (1988) *The Computer and the Mind*, Harvard University Press, Cambridge.
42. Maturana, H. and Varela, F. (1994) *L'arbre de la connaissance*, Addison-Wesley France, Paris.
43. Minsky, M. (1988) *La société de l'esprit*, Interéditions, Paris.
44. Ornstein, R. (1991) *The Evolution of Consciousness : Darwin, Freud, and Cranial Fire - the Origins of the Way We Think*, Prentice Hall, New York.
45. Gödel, K. (1931) *Über formal unentscheidbare Stätze der Principia Mathematica und verwandter Systeme I*, *Monatsh. Math. Phys.* **38**, 173–198.
46. Penrose, R. (1989) *The Emperor's New Mind*, Oxford University Press, Oxford.
47. Rosenfield, I. (1992) *The Stange, Familiar and forgotten: an anatomy of Consciousness*, Alfred A. Knopf, New York.
48. Dehaene, S. (1997) *Le Cerveau en action. L'imagerie cérébrale en psychologie cognitive*, Presses Universitaires de France, Paris.
49. Hesslow, G. and Jirenhed, D.-A. (2007) *The Inner World of a Simple Robot*, *Journal of Consciousness Studies* **14**, *7*, 85–96.
50. Metzinger, T. (2000) *Neural Correlates of Consciousness : Empirical and Conceptual Questions*, MIT Press, Cambridge, MA.
51. Metzinger, T. (2009) *The Ego Tunnel - The Science of the Mind and the Myth of the Self Basic Books*, New York.
52. Taylor, J. G. (1999) *The Race for Consciousness*, MIT Press, Cambridge, MA.
53. Tononi, G. (2008) *Consciousness as integrated information : a provisional manifesto*, *Biological Bulletin* **215**, 216–242.
54. Dortier, J.-F. (2011) *Le cerveau et la pensée, Le nouvel âge des sciences cognitives*, Sciences Humaines Editions, Auxerre.

55. Chella, A. and Manzotti, R. (2007) *Artificial Consciousness*, Imprint Academic, Exeter.
56. Aleksander, I. (1995) *Artificial Neuroconsciousness an Update*, Proceedings of IWANN '95 : International Workshop on Artificial Neural Networks, From Natural to Artificial Neural Computation, Malaga-Torremolinos, 7–9 June, Lecture Notes in Computer Science, Springer, Heidelberg.
57. Haikonen, P. O. (2007) *Robot Brains ; circuits and systems for conscious machines*, Wiley and Sons, London.
58. Sanz, R., López, I. and Bermejo-Alonso, J. (2007) *A Rationale and Vision for Machine Consciousness in Complex Controllers*, In Chella, A. and Manzotti, R. (Eds.), Artificial Consciousness, pp. 141–155, Imprint Academic, Exeter.
59. Shanahan, M. (2006) *Towards a Computational Account of Reflexive Consciousness*, Proceedings of AISB 2006 - Adaptation in Artificial and Biological Systems, Bristol, 3–6 April, pp. 165–170, Springer, Heidelberg.
60. Holland, O. and Goodman, R. (2003) *Robots with internal models : a route to machine consciousness ?* Journal of Consciousness Studies, *Special Issue on Machine Consciousness* **10**, 4.
61. Chrisley, R. (2009) *Artificial intelligence and the study of consciousness*, In Bayne, T., Cleeremans, A. and Wilken, P. (Eds.), Oxford Companion to Consciousness, Oxford University Press, Oxford.
62. Cardon, A. (2003) *Modéliser et concevoir une machine pensante. Approche constructible de la conscience artificielle*, Automates intelligents, Paris.
63. Sabah, G. (1999) *The respective roles of conscious and subconscious processes for interpreting language and music*, Proceedings of 8th International Workshop on the Cognitive Science of Natural Language Processing (CSNLP-8), Galway, 9–11 August, pp. 241–253, Advances in Consciousness Research, John Benjamins, Amsterdam.
64. Aleksander, I. and Dunmall, B. (2003) *Axioms and Tests for the Presence of Minimal Consciousness in Agents*, Machine Consciousness, *Journal of Consciousness Studies* **10**, 4–5.
65. Beshkar, M. (2008) *Animal Consciousness*, *Journal of Consciousness Studies* **15**, 3, 5–33.
66. Grumbach, A. (1999) *À propos de l'étude et de la modélisation de la conscience*, *Intellectica* **2**, 29, 171–175.
67. Dupuy, J.-P. (2003) *Le risque inouï des nanotechnologies*, *L'Écologiste* **10**, 70–72.
68. Peccoud, D. (1985) *Discours scientifique et parole poétique : deux types de communication ?* *Économie et Humanisme* **281**, 24–36.
69. Boisse, S. (2010) *L'Esprit*, L'IA et la singularité, Lulu - Serge Boisse, pr@lulu.com.
70. Kurzweil, R. (1999) *The age of spiritual machines : when computers exceed human intelligence*, Viking, New York.
71. Cardon, A. (1999) *Conscience artificielle et systèmes adaptatifs*, Eyrolles.

Attendus sur le travail du GT conscience

POURQUOI UN GROUPE DE TRAVAIL « CONSCIENCE » À L'ACADÉMIE DES TECHNOLOGIES ?

Pierre Perrier

En fait parce que le monde de nos enfants sera peuplé non pas seulement d'ordinateurs personnels et de moyens de communication entre les hommes, mais aussi de plus en plus de « robots intelligents ». Ces robots intégreront une multitude de processeurs pour essayer d'avoir un comportement d'esclaves intelligents dont les services possibles nous font rêver.

Nous avons donc à nous soucier d'une certaine sorte de conscience artificielle dont ils seront dotés, influant leur comportement envers nous dès maintenant, pour que leur présence constante et influente sur notre vie ne la transforme pas en cauchemar, mais nous apporte au contraire une aide choisie et partagée parce que maîtrisée.

Nous avons perdu en France notre industrie des machines-outils en grande partie au profit de l'Allemagne qui en a fait un fleuron de ses capacités d'exportations. Nous n'avons pas pu garder ou développer une industrie importante du produit fini électronique couvrant les besoins principaux de gestion de l'information que chacun a désormais chez soi et les moyens de communication de ces informations que chacun a dans sa poche. Ceci a eu lieu avec, pourtant, une brillante recherche

informatique impulsant notre développement en logiciels de calcul et en outils d'aide à la conception et à la réalisation.

Mais en s'élevant d'un cran dans le niveau de généralité de ces filières industrielles, on peut espérer réunir toutes les interactions homme-machine dans des produits-phare génériques en électronique-mécanique dit souvent de la « robotique » incluant dans le futur à moyen-long terme même des parties exploitant la chimie de la vie. Il y a cependant plus de généralité, et donc d'efficacité industrielle finale, à réussir l'inclusion de processus de traitement d'information présents dans les organismes vivants supérieurs pour leur donner conscience de leur environnement/activités commandées et ceci indépendamment du support matériel et technologique ou biochimique qui est adapté aux contraintes des vivants et difficile à transposer sans perte de capacité à des machines.

La mise en œuvre de cette voie à long terme est incontestablement l'étape la plus nécessaire pour orienter l'ensemble des recherches et développements et en tirer un lieu d'excellence qui sera présent au centre des avancées des sociétés à haute technologie dans leur relation à leur développement le plus riche d'avenir puisque centré sur l'homme. Trois raisons d'aider la venue de cette mise en œuvre doivent retenir l'attention de notre Académie pour aider à recentrer le débat sur les progrès à engager et à maîtriser,

- parce que l'écart est énorme entre ce qui est rêvé par beaucoup et le temps nécessaire pour disposer du contenu complexe à développer dans le monde industriel : un écart entre le calendrier rêvé et un calendrier réaliste de progrès maîtrisé qui est sans exemple dans le passé compte tenu des besoins d'innovations industrialisées et des risques de perte de contrôle par l'homme de ce type de produits et de leur technologie adaptée à un emploi dans la vie courante ;
- parce que les possibilités sont, elles aussi, énormes, mais dépendantes comme jamais auparavant d'une phase préliminaire de travaux lourds, donc non finançables aisément : ils devront porter à la fois sur le développement des technologies nécessaires et la maîtrise de concepts aujourd'hui flous dont la mise prématurée en application pourrait conduire à un fiasco humain ou inspirer un usage militaire dévoyé ;
- parce qu'un encadrement de qualification technologique, légal et sociétal est à définir par des « sages », encadrement à la définition et à la mise au point duquel notre compagnie pourrait légitimement participer dans un cadre d'abord national, puis international à anticiper.

Premier point : Réalités rêvées ou bâties selon les technologies disponibles ou à anticiper

Jamais auparavant les rêves irréalistes sur un monde de robots dominant les hommes n'ont été accumulés dans une telle production de films, de bandes dessinées et de romans, dits de science-fiction, lesquels mettent en œuvre des procédures quasi-magiques faisant fi de la plus élémentaire réalité technologique ! Que l'on regarde les anticipations de Jules Verne avec ses détails précis dessinés il y a plus d'un siècle dans les éditions originales illustrées de ses livres ou celles d'Hergé concentrant dans le personnage du « savant » professeur Tournesol les découvertes ; celles-ci semblaient rendre possible d'un coup les voyages sur la Lune ou sous l'eau : les bandes dessinées reprennent cette mise en œuvre immédiate des concepts en négligeant le développement des technologies spécifiques nécessaires, fruit du travail patient des technologues, car supposé demander seulement des extrapolations aisées des technologies existantes. Le monde actuel de la science-fiction fonctionne sur des extrapolations reposant sur des fantasmes à tous les niveaux, comme si un produit de haute technologie du futur devait être conçu et dessiné par un styliste s'appuyant sur un code généralement inspiré des diplodocus les plus impressionnants, des armures des chevaliers du Moyen Âge et dans lequel l'aspect effrayant prime la fonctionnalité et le côté légendaire et épique des luttes à l'épée magique transcende tout les possibles grâce aux effets virtuels disponibles ! Nos enfants et petits-enfants sont habitués à jouer avec des monstres ou des véhicules dits « de la guerre des étoiles » ou avec des représentations « d'extra-terrestres » qu'ils peuvent aussi monter minutieusement avec des dizaines de pièces de fonctionnalités inconnues mais magiquement imbriquées dans des formes divergeant notablement, même esthétiquement, de notre quotidien, voire des véhicules militaires.

Leurs aînés démontaient leur vélo ou jouaient du fer à souder pour réaliser des montages électroniques fonctionnels ; eux, comment pourront-ils atterrir dans le réel technologique et avoir un attrait pour ses contraintes physiques (ne serait-ce que selon le déroulement monotone du temps) et l'ascèse du passage d'une idée à sa concrétisation dans des produits efficaces.

En effet dans ces « anticipations » les habitants de la terre sont à longueur de pages de romans ou d'animations virtuellement confrontés à des extra-terrestres, ou à des créatures de pure imagination auxiliaires de ces extra-terrestres. Ils ont à faire face intellectuellement à des affrontements qui sont plus proches des campagnes

militaires face à des envahisseurs sanguinaires et sans pitié pires que n'en a connu le passé. L'atterrissage dans le réel du lecteur sera et est déjà difficile puisque ce monde imaginaire est plus intégré dans leur tête que le réel technologique par des centaines d'heures d'entraînement à vivre devant un écran allumé présentant un univers virtuel voulu si possible beaucoup plus attrayant que le réel.

En résumé, un monde de rêve complètement imaginaire, ayant ses propres règles de développement sans relation avec les lois de notre monde physique, est plus présent à l'esprit de beaucoup de jeunes que le monde réel, même le plus technologique ; il fait ainsi écran à la vision de la réalité du monde de demain chez la plupart de nos jeunes comme, d'ailleurs, et c'est plus inquiétant, chez une grande partie des intellectuels qui jugent du futur d'après ces projections.

Deuxième point : échéancier

Il est clair que, à l'échéance d'une génération et non pas de dix ans, on peut anticiper que la croissance de « l'intelligence » mise dans des supports informatiques va être telle que l'on pourra faire des robots de la taille des animaux de compagnie d'un homme et capables (après une probablement longue mise au point et grâce à des bases de données internes ajustées de façon convenable) d'assurer des travaux autonomes en interaction directe avec des hommes mêmes handicapés. Ces robots pourront descendre de différentes lignées :

- la lignée des robots industriels qui se sera développée vers une bien plus grande autonomie, en particulier de déplacement, et en même temps vers une collaboration en groupe mobile impossible actuellement ; ainsi seront créés des ensembles porteurs d'outils et formant des systèmes complets de production. Ils assureront la bonne marche de systèmes complexes en conditions d'environnement plus dures et sur des durées sans repos inaccessibles à l'homme ;
- la lignée correspondant à un trajet parallèle, par l'imitation et par la sélection, au dressage des animaux de compagnie améliorant progressivement l'extension des auxiliaires « conviviaux » assurant des « services à la personne » dont le besoin se fait de plus en plus sentir avec le vieillissement des populations ainsi que des aides au handicap que l'utopie complémentaire de l'homme « augmenté » n'est pas près de résoudre.

Dans les deux cas, le perfectionnement de « véhicules » dotés des capacités souhaitées et en bonne interaction homme-machine revient à mettre en place des robots capables d'une conduite autonome et de gérer leurs interactions avec les hommes et avec une gamme d'outils produits pour différents usages donc capable de dupliquer ou surpasser le travail humain… Dans la voiture, dans un logement avec différents outils, l'assistant gèrera une machine à laver ou un ordinateur en dialogue avec l'extérieur, assurant même le suivi médical d'un polyhandicapé en lien avec une équipe médicale, tout cela petit à petit pouvant s'intégrer dans un ou plusieurs « robot-serviteurs » qui reprendront les tâches de gestion et d'aide à l'interaction de l'homme avec son environnement trop technique voire vécu comme « inhumain ».

Qui ne rêve pour une personne âgée de la présence rassurante d'un robot de compagnie obéissant et pouvant faire toutes les tâches de liaison avec le monde extérieur et remplacer la lecture absconse d'un mode d'emploi pour mettre en route des systèmes dotés de procédures de mise en route complexes avec toute la patience d'un automate ; mais ce robot ayant aussi toutes les possibilités d'activités de premiers soins en obéissant sans hésitation à des ordres médicaux ou paramédicaux, mais non sans surveillance « douce (?) ».

À l'autre extrémité du spectre auprès d'un opérateur ayant toutes ses capacités physiques et mentales, dans des situations dramatiques et en environnement difficile, un auxiliaire de sauvetage traitant d'un danger extérieur ou intérieur en environnement hostile ou dangereux répond à un besoin aussi évident tant civil que militaire. Il va donc être développé de toute façon, mais avec quelle norme d'acceptation quand il envahira l'espace public puis privé ?

Troisième point : quelle autorisation d'emploi

Il est un peu naïf de penser que le passage d'un contrôle humain à un contrôle réparti entre des robots n'accumule pas dans leur conception un supplément de réflexion technique pointue nécessaire à la réduction des risques ; certes, cela va apporter directement une augmentation de la qualité du service rendu. Toutefois il est clair qu'un passage par un bureau d'étude travaillant pour produire un robot évolué, selon des clauses de réception précisées *a priori*, n'est nullement la garantie

d'une approche plus rationnelle et plus efficace de la qualité et de la sûreté tant que le retour d'expérience n'a pas validé les clauses selon l'usage effectivement répété de multiples fois, puisque seule l'approche empirique en utilisation réelle génère une expérience acquise dans le concret pour un système trop complexe pour être ramené à un arbre d'analyse de ses comportements « nominaux ». Ce passage par l'expérience avec analyse de nombreux cas douteux peut apporter un réel progrès ; il introduit des concepts plus efficaces, et ceci en lien avec une critique et une étude détaillée des comportements humains en interaction avec ces robots voulus très intelligents et mobiles.

L'utilisation des connaissances scientifiques et techniques issues des théories conceptuelles est indispensable pour mieux cadrer initialement « l'acte de certification » de performances nominales dans un cadre d'emploi sûr ; celui-ci évalue une qualité normée et qui autorise l'acte de « mise en service » ; une autorisation « restreinte » d'utilisation est aussi indispensable initialement pour enrichir par l'analyse des pannes et bavures les implications sociétale, juridique et les contrats d'assurance. C'est la voie unique de la sagesse dans l'introduction de robots intelligents comme dans tous les usages à risques.

Cette voie perd toute valeur, si elle ne s'est pas rodée dès le début par une expérience acquise collationnée dans une commission spécifique d'une autorité indépendante de certification capable d'enquêter précisément hors des jugements médiatisés compassionnels. Ceci sera sûrement nécessaire en présence d'accidents venant de robots trop souvent devenus, par projection, des équivalents humains en termes de responsabilité, au moins au sens du langage de communication.

Il est clair que les processus d'apprentissage et les processus de mémorisation conduisant à des conduites pertinentes des robots en fonction de la dangerosité d'une situation vont prendre une importance décisive. Les responsables techniques de ces approches dans les bureaux d'étude pourront certainement être considérés après enquête comme juridiquement responsables en cas où seraient relevés des manquements significatifs aux règles de conception et de validation qui sont encore à définir précisément.

En l'absence de concepts robustes et déclinables dans leurs présupposés, rien ne pourra être fait pour mettre en place un encadrement satisfaisant des risques puis des responsabilités en balance avec des performances souhaitables. Il importe donc à l'Académie de repérer les concepts et de proposer et superviser les

règles de conception et de normalisation qui peuvent avoir un rôle bénéfiques d'encouragement à l'innovation nécessaire pour aider les premiers pas ou bien face à des protectionniste, ou bien face à des obstacles nuisant au développement des produits finaux en France.

Un effort de recherche sur le langage pertinent aux robots, en particulier aux robots de compagnie, devrait aider à repérer les langages inadéquats véhiculés par des extrapolations auto-référencées venant des thèses véhiculées par la littérature de fiction, comme des réflexions ne s'appuyant pas assez sur une connaissance profonde, théorique et pratique de la réelle complexité des phénomènes de linguistique, de mémorisation et de réflexion en miroir au cours d'interactions homme-robot. Tout ceci doit participer à la construction par approximations successives d'un corpus de recommandations, issues du bon sens certes, mais bien informé des risques et bénéfices et encadrés par l'expérience de l'emploi de ces nouveaux « objets ».

Il s'agit de proposer les règles d'une « robologie » en dialogue selon une anthropologie cohérente et d'encourager la génération d'une série de règles conceptuellement homogènes, applicables dès la conception en ses concepts de base en tenant compte des limites mécaniques de leurs actions.

Ainsi, avant que les concepteurs français ne les reçoivent d'une culture et d'une pensée moins exigeante voire inadaptée aux finesses des hommes dans leur variété, l'Académie pourrait ouvrir à des règles de spécification inspirées peut-être des règles existantes, par exemple sur l'emploi des animaux de compagnie apprivoisés, en tout cas du suivi des apprentissages qu'ils auront dû effectuer et des règles de validation sur parcours libre des robots de nouvelle génération, ceci en interaction avec l'homme simulé puis effectif.

Conclusions

C'est à l'aune de ce besoin du XXI^{e} siècle que doit être compris l'effort de discernement poursuivi pour que le groupe de travail ait une « bonne conscience » humaine des limites des raisonnements actuels, des implémentations actuelles ou en développement, donc face à des logiciels trop restreints par rapport au besoin minimal.

Cependant c'est sur ces bases restreintes de nos connaissances actuelles que doit être tentée dès aujourd'hui une première rationalisation des « robologies » du futur qui échapperaient aux rêves médiatiques et permettraient de fonder une industrie avec des garde-fous et des services effectifs ouverts sur les progrès continus espérés au service maîtrisé de tous.

GLOSSAIRE

Les mots surlignés dans le texte sont des entrées dans ce glossaire.

Ancrage des symboles

Chez l'homme, les capacités symboliques consistent à lier plus ou moins arbitrairement un signe à un état mental (ce signe devient alors le symbole de cet état mental) et à utiliser ce *symbole* dans des processus de raisonnement, de planification et de communication. Ce lien, intuitif, culturel ou conventionnel, est ce qu'on appelle l'ancrage du symbole.

Un système d'IA classique manipule des symboles qui restent dénués de sens pour lui-même puisqu'il ne connaît que leur forme : tout sens est attribué aux symboles de l'extérieur (le programmeur, la théorie ou l'utilisateur). Il s'agit alors d'une sémantique dénotationnelle (les représentations pointent sur des choses et ce pointage se réalise par un observateur extérieur).

Divers arguments ont été avancés contre la possibilité de simuler l'intelligence par la manipulation de symboles fondée sur le signifiant seul (c'est-à-dire des manipulations uniquement syntaxiques ne portant que sur la forme des symboles et des formules).

Une réponse est d'ancrer les processus mentaux et leurs représentations à l'extérieur du système qui les manipule. Il faut alors doter un système de capacités lui permettant d'identifier ses perceptions et d'y reconnaître des discriminations (catégorisation). Si certaines catégories sont probablement innées (données *a priori* au système), d'autres doivent être apprises pour que le système construise lui-même l'ancrage des symboles qu'il manipule.

De là vient l'hypothèse que pour créer une véritable intelligence dans un système d'IA, il faut que ce système soit plongé dans un monde réel qu'il peut percevoir et sur lequel il peut agir – ce qu'on appelle l'IA « située ». Le système aurait alors la possibilité de construire lui-même, par apprentissage, les liens entre les signifiants qu'il manipule et les signifiés correspondants.

Animisme

Croyance en une âme, une force vitale, animant non seulement les êtres humains, mais également les animaux et les éléments naturels (pierres, arbres, vent…). Ces âmes – en particulier celles des esprits supérieurs (ceux des défunts ou de divinités animales) – sont censées pouvoir agir sur le monde réel.

Attention

Distinct de la conscience, dirigé par des buts (volontaires ou non), ce processus (censé exister chez tous les mammifères) rend conscient un résultat ou un processus actif et correspond à la capacité à se concentrer sur une activité pendant un temps donné.

L'attention est un processus qui permet à un sujet d'augmenter l'efficacité des processus de perception, de réflexion et de remémoration.

La perception d'un stimulus nécessite tout d'abord l'orientation (volontaire ou non) des organes sensoriels en direction du signal, puis la focalisation des ressources cognitives sur le signal perçu afin de permettre son traitement et son interprétation.

« La réflexion n'est autre chose qu'une attention à ce qui est en nous. »
(Leibniz)

Attention consciente (ou volontaire)

Déclenchée par un processus conscient, elle produit l'apparition dans la mémoire à court terme d'un résultat ou d'un but à atteindre.

Attention non consciente (automatique)

Déclenchée par une surprise (une différence entre un résultat effectif et une attente espérée). Cette situation produit l'émergence consciente des résultats correspondants.

Pour Baars, l'attention joue le rôle de contrôle métacognitif de la conscience : l'attention volontaire correspond à un contrôle conscient de l'accès à la conscience et s'oppose à l'attention automatique, liée à un contrôle inconscient de l'accès à la conscience. Il note une possibilité de lutte entre attention volontaire et attention involontaire.

Autonomie

Faculté de se déterminer par soi-même, de choisir, d'agir sans contrainte externe.

Cartésianisme

La philosophie cartésienne repose sur quelques postulats simples que l'on peut résumer de la façon suivante :

- l'homme peut accéder à la connaissance universelle par la raison ;
- il emploie pour cela toutes les ressources de son intelligence, en premier lieu l'« intuition évidente » et la déduction, mais également l'imagination, les sens, et la mémoire ;
- l'homme est une « substance pensante ». Cela s'exprime par le célèbre *cogito ergo sum*, exposé dans *Le discours de la méthode*, et précisé pour l'essentiel dans les *Méditations sur la philosophie première* ;
- l'homme peut s'appuyer sur la raison seule, et n'a pas besoin des « lumières de la foi » pour accéder à la connaissance ;
- Partant de ces postulats, toute la connaissance repose sur une nouvelle métaphysique, y compris la morale.

Catégorisation

La catégorisation est le moyen principal par lequel nous donnons du sens à nos expériences ; elle permet de répondre aux questions : Pourquoi chaque objet n'est-il pas considéré de façon unique ? Quels critères décident de l'appartenance d'un membre à une catégorie ? Comment s'effectuent les regroupements ? Catégorisation et dénomination sont indissociables.

On distingue deux points de vue fondamentalement différents : le courant objectiviste et le courant expérientialiste.

Le premier, fondé sur le modèle aristotélicien, suppose que les catégories ont des frontières clairement délimitées, l'appartenance d'une entité à une catégorie est booléenne et s'opère sur la base de propriétés communes (conditions nécessaires et suffisantes [CNS]) ; tous les membres d'une catégorie ont un statut identique. Dans l'ensemble des caractéristiques attachées à un mot, on distingue les traits essentiels (CNS) et des traits contingents ou accidentels.

Dans le second, la catégorisation s'opère sur la base de mises en relief de similitudes globales et de formation de prototypes de référence. On a à nouveau deux approches, la version standard et la version étendue :

- pour la version standard, on distingue « le meilleur » exemplaire communément associé à une catégorie (ce n'est pas un exemplaire particulier mais une sous-catégorie) : ce prototype se redéfinit comme l'exemplaire qui condense les propriétés les plus saillantes de la catégorie. Les membres d'une catégorie sont reliés les uns aux autres sans qu'il existe forcément une propriété commune à tous (ressemblance de famille), mais ils ont tous au moins une propriété commune avec le prototype ;
- pour la version étendue, la notion de degré de prototypicalité remplace la notion de prototype. Le point central est la notion de ressemblance de famille qui unit l'ensemble des éléments d'une catégorie (deux éléments peuvent alors n'avoir aucun trait commun, mais tout élément a au moins un trait commun avec un autre élément de la catégorie).

Pour l'une comme pour l'autre, les mécanismes de classification ne sont pas toujours clairs.

Cognition

L'ensemble des processus mentaux mis en œuvre dans la gestion des connaissances (acquisition, stockage, transformation et utilisation), allant de la perception au langage, en passant par l'imagerie mentale, la mémoire, la résolution de problèmes, le raisonnement et la prise de décision…

Compétence

Les capacités linguistiques théoriques d'un individu qui maîtrise une langue (c'est-à-dire : la connaissance intuitive des règles de sa langue).

Comprendre

Donner un sens clair à quelque chose. Se faire une idée nette des causes, des motifs d'une situation. Avoir une connaissance intuitive.

Concept

Voir aussi intension et extension

Représentation générale et abstraite d'une réalité (à distinguer de la chose elle-même comme du signe qui la désigne).

Identification d'un ensemble de traits dans une situation, ensemble qui a une influence sur le comportement de l'individu pour lequel cet ensemble de traits caractérise une entité cognitive correspondant à une description globale d'un objet, d'une classe d'objets ou de situations (ce point serait à généraliser, en particulier aux aspects abstraits).

Définitions plus techniques

Un nœud d'un réseau sémantique. Son sens est donné par les relations qu'il entretient avec les autres nœuds ; l'étiquette n'a pas d'importance en elle-même ; plus généralement, les éléments (ou catégories) à partir desquels sont construits les critères dits sémantiques ; (En liaison avec la logique formelle) une des constantes du domaine d'interprétation ($\mathcal{D}$), ou une relation sur $\mathcal{D}^n$ (interprétation d'un prédicat n-aire) ou encore une fonction sur $\mathcal{D}^p$ (interprétation d'une fonction p-aire).

Conscience

(en italique les traductions anglaises de chaque sens car celles-ci distinguent mieux les concepts sous-jacents).

Usage courant

- être éveillé (se rendre compte) *awakeness*
- connaissance immédiate (spontanée) *awareness*
- d'un état du monde (*présence d'un prédateur, situation passée ou future…*),
- d'une propriété du monde (*le rouge de la pomme, le goût du Médoc…*),
- d'un état corporel (*faim, froid, douleur… : conscience visuelle, olfactive, tactile, gustative, auditive*)

Conscience psychologique

- d'être d'une certaine manière (*ce que ça fait « d'être X »*) (proprioception ?) *awareness/consciousness*
- connaissance de sa propre activité psychique et de ses opérations mentales (*introspection ; réflexivité*) *consciousness*

Définition physique

- (mécanique quantique ?)

Conscience morale

- faculté de juger ses propres actes ;
- conscience ;
- (sans parler de conscience professionnelle, de liberté de conscience…)

Quatre hypothèses sur la façon dont la conscience peut être vue :

- *L'hypothèse de l'activation* (un élément est conscient lorsque son activation dépasse un certain seuil). Toutefois, cette définition n'est généralement pas l'interprétation donnée à l'activité elle-même. Elle n'explique pas non plus comment on perd la conscience des événements répétés et prédictibles (ce qui est à mettre en relation avec l'apprentissage). La notion d'activation est plutôt une façon de modéliser la probabilité qu'un événement a de devenir conscient que la conscience elle-même.

- *L'hypothèse de la nouveauté.* Par ce concept lié à la notion d'information, on considère que ne deviendraient conscients que les éléments qui apportent effectivement de l'information.
- *L'hypothèse du sommet de l'iceberg.* Ce qui est conscient n'est que l'émergence de tout un ensemble d'expériences inconscientes. Lié au fait que nos capacités conscientes sont très limitées (par rapport à d'autres capacités intellectuelles). Les limites elles-mêmes de la conscience sont alors une caractéristique importante qui demande à être expliquée.
- *L'hypothèse du théâtre.* La conscience est ici vue comme le lieu de présentation des résultats produits par les traitements issus de nos sens (*cf.* Platon et sa caverne, ou le « théâtre cartésien » de la conscience supposant qu'il existe un lieu dans le cerveau où les informations sont collectées pour être rendues conscientes).

Plutôt que de donner une (des) définition (s), on peut préciser quelques caractéristiques importantes de la conscience :

- *Sélectivité.* Tout n'arrive pas à la conscience. La conscience a une fonction de sélection permettant d'extraire les éléments menant à des pensées « intéressantes » ;
- *Exclusivité.* On n'est conscient que d'une chose à la fois. La conscience a un effet de séquentialisation. Or, comme il existe diverses activités cérébrales en parallèle, tous les niveaux ne sont donc pas conscients ;
- *Enchaînement.* Les événements conscients sont traités en série. La conscience a une fonction constructive consistant à mettre ensemble les résultats disparates des processus inconscients ;
- *Unité.* Ce qui fait que l'esprit est un tout. La conscience recrée ou modifie aussi bien les résultats des perceptions que les données elles-mêmes.

Corrélation

Au sens d'Edelman : coïncidence d'événements au niveau neuronal. Elle mesure la relation existant entre deux événements neuraux qui ne peuvent être pensés l'un sans l'autre ou leur tendance à être présents en même temps.

Dénotation

Voir extension

Dualisme

En philosophie, le dualisme se réfère à une vision de la relation entre la matière et l'esprit fondée sur l'affirmation que les phénomènes mentaux possèdent des caractéristiques qui sortent du champ de la physique.

Ces idées apparaissent pour la première fois dans la philosophie occidentale avec les écrits de Platon et Aristote, qui affirment, pour différentes raisons, que l'« intelligence » de l'homme (une faculté de l'esprit ou de l'âme) ne peut pas être assimilée ni expliquée par son corps matériel.

La version la plus connue du dualisme a été formalisée en 1641 par René Descartes qui a soutenu que l'esprit était une substance immatérielle. Descartes fut le premier à assimiler clairement l'esprit à la conscience, et à le distinguer du cerveau, qui est selon lui le support de l'intelligence. Ainsi, il a été le premier à formuler le problème corps-esprit de la façon dont il est présenté aujourd'hui.

De nos jours, le dualisme est opposé à des formes variées de monismes, parmi lesquelles le physicalisme et le phénoménisme.

Émergence

En règle générale, l'émergence caractérise un événement qui semble en discontinuité avec les événements antérieurs et qui n'est pas expliqué par ses constituants. En termes structurels, « émergence » est l'équivalent de « l'apparition d'une structure au sein d'une structure précédente » là où la succession logique des structures paraît impossible à établir.

En psychologie, on parle d'émergence à propos du passage d'une formule de comportement à une autre, sur la ligne de croissance normale d'un organisme vivant.

En biologie, on se réfère par ce terme à l'apparition (souvent par mutation) d'un organe nouveau ou de propriétés nouvelles d'ordre supérieur.

En intelligence artificielle, la question de l'émergence des symboles dans un système automatique est liée à celle de leur ancrage, c'est-à-dire de la possibilité de conserver une trace de leur relation avec les événements qui leur ont donné

naissance. L'émergence du symbole ne se fait donc pas à partir du regroupement en formes selon des règles syntaxiques, mais par cet ancrage empirique et pragmatique. Les systèmes symboliques ont simultanément un caractère individuel et un caractère social ; pour leur émergence, il faut en plus des capacités d'auto-organisation et des mécanismes d'évolution dynamique.

Empirisme

L'empirisme postule que toute connaissance provient essentiellement de l'expérience. Représenté par exemple par les philosophes anglais Roger Bacon, Francis Bacon, John Locke et David Hume, ce courant considère que la connaissance se fonde sur l'accumulation d'observations et de faits mesurables, dont on peut extraire des lois générales par un raisonnement inductif, allant par conséquent du concret à l'abstrait. L'« empirie » est ainsi l'ensemble des données de l'expérience pure, considéré comme l'objet sur lequel porte la méthode expérimentale.

Émotion

Réponse physiologique involontaire, non consciente, induite par des éléments provenant de l'environnement ou du milieu interne de l'organisme. Une émotion a deux aspects : l'un, expressif, traduit par le comportement et visible par autrui, l'autre, neurophysiologique, interne mais observable par les moyens scientifiques. [Vision de Damasio]

Environnement

Ensemble des éléments et des phénomènes physiques qui environnent, se trouvent autour d'un organisme vivant ; contexte.

Esprit

- Principe de la pensée et de l'activité réfléchie de l'homme ; ensemble des facultés psychologiques tant affectives qu'intellectuelles.
- Intention, point de vue particulier déterminant une attitude.

- Capacité intellectuelle (*esprit* d'abstraction, *esprit* mathématique) ; disposition psychique dominante d'une personne ou d'un groupe, déterminant le choix d'une attitude et l'orientation de l'action (*esprit* d'indépendance).

Expérience

- Expérience scientifique : réalisée selon un protocole précis et renouvelable.
- Les connaissances d'un individu construites à partir de situations passées vécues.
- Un exercice intellectuel, éventuellement indépendant de la réalité (expérience de pensée).
- Une pratique innovante dans un domaine artistique.

Extension (ou dénotation)

L'ensemble des objets auxquels s'appliquent les caractères qui définissent un concept.

Extéroception

Sensation incluant vision, audition, somesthésie générale, olfaction et gustation.

Gelée grise (ou *gray goo*)

Formulée pour la première fois par Éric Drexler en 1986, dans son texte *Les moteurs de la création (Engines of Creation)*, la menace du *gray goo* – gelée ou glue grise – reprise par l'univers de la science-fiction, est un exemple des inquiétudes produites par le développement des nanotechnologies.

Désormais entré dans l'imagination collective, ce scénario, à la base du célèbre roman de Michael Crichton *La proie (Prey)* paru en 2002, évoque une catastrophe généralisée suite à la perte de contrôle du processus d'autoréplication d'une nuée de nanorobots échappés d'un laboratoire.

Dans le mythe de la gelée grise, cette autoréplication consomme toute la vie sur la planète pour s'assurer l'énergie nécessaire au processus de reproduction des nanorobots. La masse des nanomachines répliquantes, qui forme une substance gluante (*goo*), remplace alors la matière de l'univers.

Holisme, holistique

L'holisme est la théorie qui consiste à considérer les phénomènes dans leur ensemble. Les diverses parties d'un élément ne pouvant se comprendre que par rapport à cet ensemble, qui est ce qui lui donne une signification. S'oppose à l'approche analytique ou atomistique.

Idéalisme

Prééminence donnée à des formes abstraites ou à des représentations mentales sur la réalité, qu'elle soit expérimentée ou qu'elle soit inconnaissable ; ces formes sont considérées comme l'essence de cette réalité, éventuellement réduite ainsi à une dimension illusoire.

Idée

- Ce que l'esprit conçoit, peut concevoir ou se représenter, par opposition aux phénomènes concernant l'affectivité ou l'action.
- Ce qui n'existe que dans l'esprit, dans l'imagination, par opposition à ce qui existe en fait, dans la réalité, de façon concrète.
- Ensemble des opinions d'une personne ou d'un groupe de personnes.
- L'ensemble du mouvement intellectuel concernant une époque, une civilisation.
- Représentation intellectuelle, abstraite, générale, d'un objet.
- Esquisse, ébauche, représentation élémentaire, sommaire.

Identité personnelle, personnalité, personne, je, moi, soi

- Ce qui constitue un individu, qui le rend psychiquement, intellectuellement et moralement distinct de tous les autres.
- Conscience de la persistance du moi, de l'unité de sa vie psychique et son identité dans le temps.
- Ensemble des traits ou caractéristiques qui, au regard de l'état civil, permettent de reconnaître une personne et d'établir son individualité au regard de la loi.

Initiative

- Capacité qui permet d'entreprendre quelque chose, de prendre une décision, sans attendre un ordre venant de l'extérieur.
- « Prendre l'initiative » : action de celui qui, le premier, propose ou réalise quelque chose de lui-même.

Intension (on disait aussi autrefois compréhension)

Description du sens d'une expression comme l'ensemble des propriétés que possèdent les concepts correspondants (par opposition à *extension*).

Intention

Le but qu'un être pensant se propose d'atteindre – avec préméditation. Préalable à l'exécution d'une action, elle implique la conception de l'acte, ainsi que la conscience de l'acte et de ses conséquences.

Intéroception

Sensations dues aux fonctions physiologiques et aux modalités sensorielles inconscientes.

Langage de la pensée (ou mentalais)

Une hypothèse de Jerry Fodor qui suppose un langage (analogue à une langue naturelle) qu'utiliseraient les processus mentaux, permettant d'élaborer des pensées complexes à partir de concepts plus simples. Il se fonde sur les faits suivants : 1) les représentations mentales sont structurées, 2) les composants de ces structures peuvent apparaître dans des représentations différentes et 3) les représentations mentales possèdent une sémantique compositionnelle ; le sens des représentations complexes se construit à partir du sens de leurs composants.

Le mentalais se différencie toutefois des langues naturelles car il ne se réalise qu'au travers de configurations neuronales, mais ni par des sons ni par des écrits ou des gestes. Ces configurations sont également supposées entretenir une analogie

importante avec les bits dans la mémoire d'un ordinateur. Il implique enfin qu'une pensée sans langage (au sens langue naturelle) est possible.

Langage intérieur

Activité langagière produite uniquement mentalement (« se parler à soi-même »).

Alain Morin soutient l'idée que sans langage intérieur, les êtres humains ne pourraient pas être conscients (le dialogue que nous entretenons avec nous-mêmes est ce qui nous rend conscient de notre existence).

Logique formelle

C'est un langage dont le vocabulaire est constitué de *symboles de base* :

- un ensemble de constantes qui désignent les objets composant l'univers que l'on veut représenter ;
- un ensemble de variables qui permettent de désigner des classes d'objet et d'écrire des formules générales ;
- un ensemble de prédicats qui désignent des propriétés des objets et prennent les valeurs *vrai* ou *faux* ;
- un ensemble de fonctions qui désignent des propriétés des objets et prennent les valeurs dans des sous-ensembles des constantes.

Ces symboles n'ont une signification que lorsqu'on les interprète, extérieurement au système lui-même (voir la définition de modèle).

Les *connecteurs logiques* permettent de relier entre eux des prédicats pour constituer des formules plus complexes : et, ou, non ou implique.

Là encore, la signification de ces symboles doit être précisée par une interprétation (par exemple, « A et B » est vrai si et seulement si A est vrai et B est vrai). Le « et » du langage courant permet de définir la signification du « et » logique.

Matérialisme

Le matérialisme considère que la matière construit toute réalité et s'oppose au spiritualisme pour lequel l'esprit domine la matière. Matérialisme et spiritualisme sont des philosophies sur la **nature** de l'être. Leur opposition ne doit pas se

confondre avec celle de l'idéalisme et du réalisme, qui sont des doctrines sur l'origine de la connaissance. D'une façon générale, le matérialisme rejette l'existence de l'âme, de l'esprit, de la vie éternelle, ou de Dieu. Il considère que la conscience, la pensée et les émotions sont les **conséquences** de mécanismes matériels. Pour le matérialisme, la mort du corps matériel entraîne la disparition de la conscience et de la sensation d'exister. Le matérialisme considère que le monde résulte de mécanismes matériels, sans but et sans signification et que l'esprit est une illusion. Matérialisme et spiritualisme ne sont pas des doctrines du courant réaliste, leurs philosophies respectives reposent sur une représentation mentale de la réalité (idéalisme).

Mécanisme

Conception matérialiste qui perçoit la plupart des phénomènes suivant le modèle des relations de cause à effet.

Mémoire

Globalement, la mémoire est l'ensemble des processus d'encodage, de stockage et de récupération des représentations mentales de l'information perçue et du souvenir (subjectif). Faculté cognitive de l'homme et des animaux ; dispositif physique permettant la conservation et la restitution de données en informatique.

Deux théories s'opposent sur ce que sont les divers types de mémoire : la théorie structurelle se fonde sur des différences de nature physique (modules et structures cérébrales différentes) ; la théorie fonctionnelle suppose au contraire que le support physique est le même et que ce sont des différences d'activation qui rendre compte des « modules » ou fonctionnements différents. Quel que soit le modèle, on distingue les sous-systèmes suivants :

- **le registre sensoriel** – il peut retenir une grande quantité d'informations sous forme visuelle, auditive ou tactile, pendant un temps extrêmement court (processus différent du phénomène de rémanence visuelle) ;
- **la mémoire de travail** – avec un point de vue fonctionnel, on peut dire que la mémoire de travail ne représente que des parties plus ou moins activées de la mémoire à long terme. Selon Cowan la partie la plus activée de la mémoire de travail correspond au focus *attentionnel* (voir attention).

On peut alors distinguer :

- La *mémoire à court terme* : elle contient un nombre limité d'éléments, stockés sous forme verbale ou imagée pendant quelques secondes. Ce qui y émerge est considéré comme conscient ;
- Une partie *subliminale* au sein de laquelle les opérations interprétatives établissent une cohérence vis-à-vis du contexte cognitif courant. Une interprétation cohérente franchit le seuil de la conscience et apparaît dans la mémoire à court terme ;
- La *mémoire à long terme* qui ne connaît pas en pratique de limites (ni théoriques, ni pratiques) de capacité ou de durée de mémorisation. On distingue :
  - la mémoire des épisodes (souvenirs des événements de la vie personnelle) ;
  - la mémoire sémantique (des connaissances générales pour la perception et la compréhension [du langage en particulier]) ;
  - la mémoire encyclopédique (contient les connaissances sur le monde).

Avec un point de vue différent, celles-ci se répartissent selon une autre distinction :

- la mémoire implicite (procédurale) qui permet l'acquisition et l'utilisation d'automatismes pour traiter les informations verbales, imagées, sensitives et motrices…
- la mémoire explicite (déclarative) responsable de la mémorisation de toutes les informations sous forme verbale, imagée ou auditive (représentations mentales manipulables par la conscience et l'attention).

Enfin, on peut également parler de métamémoire : une mémoire de la mémoire, qui se souvient des variations de celle-ci. Elle permet à l'esprit de s'abstraire du présent et de considérer la succession des souvenirs de ses divers états de mémoire. Cette propriété serait aussi nécessaire à la construction de la conscience de soi.

Métareprésentation

Les métaconnaissances, permettent à un système d'observer son propre fonctionnement et d'utiliser une représentation de lui-même pour raisonner sur son propre comportement.

Pour chaque agent, chaque procédure ou chaque module, on distingue ainsi clairement ses connaissances du domaine, d'une part, et les connaissances sur son comportement et ses interactions avec les autres éléments, d'autre part. Cette connaissance de sa propre activité correspond à une première mise en œuvre d'une conscience partielle (représentation de soi- même).

L'objet et la partie qui le représente sont deux systèmes différents, mais chacun d'eux correspond à un modèle conceptuel classique. Le métaraisonnement est ainsi considéré et réalisé comme n'importe quel autre type de raisonnement.

Toutefois, pour que ce dernier niveau « représente » réellement le système initial, il faut qu'il existe un lien causal entre cette représentation et le système : si l'un des deux change, l'autre change en conséquence (la représentation doit toujours être véridique). Dans le même ordre d'idée, les propriétés significatives de la métareprésentation doivent être en accord avec le niveau objet (intégrité introspective).

Dans une métareprésentation, seules sont représentées les parties du système nécessaires pour les raisonnements réflexifs, elle est donc sélective (ce qui permet une réduction de la complexité du système). En outre, le contenu de la métareprésentation dépend du type de raisonnement qui va l'utiliser : elle est spécialisée (un système « conscient » a donc la possibilité de disposer de plusieurs métareprésentations du même objet). En conséquence, cette représentation est partielle (ce n'est pas une simple image complète du système, mais une représentation des seuls éléments pertinents pour la tâche considérée).

Modèle

Le mot *modèle* a deux sens en quelque sorte opposés, le premier essentiellement en logique, le second plus courant et intuitif.

1°) Une incarnation particulière d'une structure abstraite. Ce premier sens se retrouve dans les systèmes formels quand on veut donner une sémantique à un ensemble de formules : on définit l'ensemble $\mathcal{D}$ des constantes de l'univers (appelé domaine) et une interprétation des fonctions et des prédicats. Ceux-ci ne sont en effet que des identificateurs qui ne sont porteurs d'aucune signification *a priori*. Ce n'est que lorsque l'on a indiqué le sens de toutes les fonctions et de tous les prédicats que l'on pourra dire que les formules représentent effectivement des connaissances. Un modèle d'un ensemble des formules consiste donc, outre la

définition du domaine $\mathcal{D}$ et des tables de vérité des connecteurs, en la mise en évidence pour chaque prédicat n-aire d'une relation sur $\mathcal{D}^n$ et pour chaque fonction n–aire d'une fonction de $\mathcal{D}^n$ dans $\mathcal{D}$; il s'agit donc ici d'une instanciation particulière respectant un certain formalisme (nous réserverons le mot « interprétation » pour ce concept). Ainsi, l'arithmétique des nombres rationnels est un exemple de modèle de la structure de corps commutatif.

2°) Un schéma formel *plus abstrait* que le phénomène dont il est l'image. Il s'agit cette fois d'une généralisation respectant la vérité des données spécifiques dont elle est issue. C'est essentiellement ce second sens qui est utilisé dans l'ensemble du présent document (il sera désigné par le terme « modèle formel » ou plus simplement « modèle »). On peut en distinguer différents types :

- **modèle descriptif**
  Par rapport à une réalité supposée exister objectivement, un modèle est une relation (application j) entre cette réalité et un langage formel de description. Il peut ainsi être vu tout d'abord comme un mécanisme de représentation. À ce niveau, on dispose donc d'un outil mathématique formel, dont la seule validation possible est une preuve interne d'autocohérence ; cela n'implique nullement que les modèles dérivés correspondent à une quelconque réalité … Ces modèles permettent une analyse rétroactive des activités passées, et se retrouvent dans des disciplines comme biologie et neurobiologie (un danger à signaler est qu'ils peuvent parfois être quelque peu réductionnistes). Des exemples de tels modèles sont issus du behaviorisme ; comme on doit se contenter de chercher des régularités entre les entrées et les sorties, ces modèles ne sont descriptifs qu'« en moyenne ». Dans ce cadre, on peut éventuellement invalider des hypothèses, mais on ne peut jamais être certain de leur validité.
- **modèle prédictif**
  Généralement, s'il existe une application φ^{-1}, on considère que l'outil est aussi prédictif : des résultats au niveau du modèle permettent alors d'inférer certains états de la réalité, éventuellement non encore rencontrés. Le modèle est considéré comme valide quand ces inférences sont effectivement avérées. Ces modèles sont liés au structuralisme : chaque élément est décrit par rapport à l'ensemble ; les modèles sont alors statiques et explicatifs. Ils décrivent un état de choses, mais ont des difficultés pour en décrire les évolutions (ce qui signifie qu'ils ne sont pas forcément explicatifs).

Les disciplines concernées ici sont essentiellement la psychologie cognitive qui souhaite des modèles prédictifs et cohérents avec les vérifications expérimentales mais aussi permettant la compréhension du phénomène, tandis que la linguistique s'intéresse à des modèles prédictifs vérifiables expérimentalement (alors qu'elle s'intéresse rarement au phénomène lui-même).

- **modèle d'exécution (proactif)**
  De façon plus pragmatique, une exigence forte introduite par l'intelligence artificielle porte sur le fait que les machines ont besoin de modèles non seulement pour décrire et prévoir, mais aussi pour agir. Ces modèles *proactifs* permettent alors que l'entité qui se représente une certaine réalité prenne en considération ses actions sur cette réalité elle-même. Plus proches du constructivisme, ils permettent la prise en considération de l'observateur et des processus d'évolution des structures étudiées, et impliquent donc une certaine perte de l'objectivité. La notion de niveaux d'interprétation devient également primordiale : les interprétations d'un niveau dans les termes d'un niveau inférieur (réductionnisme…) ne sont pas compositionnelles mais émergentes.

C'est essentiellement ce troisième type de modèle qui permet le développement de simulations sur machine ; on peut alors observer les résultats, les ajuster, et expérimenter à nouveau, selon une boucle d'*amorçage* absolument essentielle pour une véritable intelligence artificielle.

Monisme

Le monisme est la doctrine fondée sur la thèse selon laquelle tout ce qui existe – l'univers, le cosmos, le monde – est essentiellement unitaire, et qu'il est donc constitué d'une seule substance (une réalité fondamentale qui n'a besoin que d'elle-même pour exister). Le monisme s'oppose à toutes les philosophies dualistes, qui séparent le monde matériel et le monde spirituel (l'au-delà).

Occasionalisme (Malebranche)

L'action du corps sur l'esprit et de l'esprit sur le corps est impossible ; en conséquence, c'est Dieu qui agit seul, en conformant la volonté de l'esprit aux actes du corps.

Pensée

(on oublie ici la fleur)

- Activité psychique aussi bien affective qu'intellectuelle.
- Faculté de connaître, de raisonner, de juger ; activité intellectuelle qui en est la source.
- Ensemble des capacités intellectuelles d'une personne.
- L'esprit en tant que faculté de se représenter ce qui n'existe pas en réalité (imagination, souvenir).
- Forme de l'esprit propre à une personne ou à un groupe de personnes.
- Opinion.
- Produit de cette faculté.
- Toute représentation dans la conscience (laquelle inclut notamment celle d'un sentiment, d'une sensation, d'un état d'âme).
- Toute représentation mentale à caractère objectif (laquelle comprend également l'image).
- Ensemble de réflexions réunies en recueil par un auteur.
- Fait de se représenter mentalement quelque chose, d'en avoir conscience.
- Fait d'imaginer quelque chose.
- Fait d'envisager quelque chose.
- Fait de se représenter mentalement quelque chose ou quelqu'un ; le fait d'envisager quelque chose.

Perception

La perception est le résultat de l'action des sens sur l'esprit et du travail de celui-ci qui organise les données sensorielles pour se constituer une représentation de la situation perçue. Le mot désigne soit l'activité elle-même, soit le résultat de cette activité. Chez l'être humain, on distingue la perception consciente de la perception inconsciente (ou implicite ou subliminale). On parle aussi de perception active quand l'individu gère son attention et ses actions afin de découvrir dans la situation des informations définies.

On peut également distinguer :

- perception du monde ;
- perception de ses actes ;

- perceptions internes ;
- perception de sa personnalité ;
- perception de ses intentions ;
- perception de ses désirs ;

…

(*cf.* aussi l'entrée *sensation*)

Performance

Les productions effectives qui sont faites par les sujets connaissant une langue en parlant ou en écrivant (réalisation concrète de la compétence).

Phénoménisme

Doctrine philosophique dont la thèse principale, sous sa forme absolue, consiste à refuser l'existence de toute substance matérielle ou spirituelle sous les phénomènes perçus par les sens et la conscience. Le phénoménisme est ainsi un cas très particulier d'une doctrine beaucoup plus vaste, l'idéalisme. Si l'idéalisme est généralement proche du rationalisme (*cf.* ses plus grands représentants, Platon, Descartes, Malebranche, Hegel et Scheling), le phénoménisme est proche de l'empirisme pour considérer que tout n'est qu'un faisceau de phénomènes.

Physicalisme

Le physicalisme est une thèse métaphysique (soutenue, entre autres auteurs par Quine) selon laquelle toute entité existante est de nature physique, c'est-à-dire qu'il n'y a rien en dehors des choses dites physiques. En philosophie de l'esprit, le physicalisme admet que le mental est une réalité physique ; il s'agit donc d'une forme de matérialisme et de monisme, qui peut être mis en parallèle avec les premiers philosophes grecs, comme Thalès qui soutint que tout est eau. Dans sa version la plus radicale, on peut exprimer cette thèse ainsi : « Un moniste matérialiste suppose que tous les phénomènes chimiques, biologiques, psychologiques, linguistiques, culturels et sociologiques sont des phénomènes physiques qui obéissent aux lois fondamentales de la physique ».

Une thèse contradictoire, mais également moniste, est l'idéalisme immatérialiste, illustré par George Berkeley, qui soutint que tout ce qui existe est un phénomène mental.

Une autre thèse contradictoire, non moniste cette fois, est le dualisme.

Polysémique

Propriété d'un mot, d'une phrase ou d'un texte de posséder plusieurs sens, plusieurs interprétations (ambiguïté).

Processus

Un processus est une séquence d'activités qui utilise des moyens divers pour transformer les éléments qui lui sont donnés en entrée et produire comme résultat des éléments en sortie. Le processus est déclenché par un élément qui est garant de son bon fonctionnement.

- **processus conscient**
  Lorsque nous pensons, soit nous pouvons nous représenter (ou dire quelque chose de) la façon dont certaines opérations ont été effectuées ainsi que de leurs interactions. Nos processus mentaux sont alors conscients. Ces processus travaillent en série, traitent des contenus très variés, ont une grande cohérence interne, mais sont peu rapides, sont sujets à des erreurs et sont sensibles à des interférences avec d'autres processus conscients.
- **processus inconscient ou inconscient cognitif**
  Ensemble des processus de traitement de l'information consciemment inaccessibles au sujet (et donc à l'introspection) ; on ne se rend compte que de leur résultat. Ils sont dits inconscients. Ils travaillent dans des domaines limités mais sont très efficaces pour leur tâche spécifique (ils font peu d'erreurs, sont rapides, et ne subissent pas d'interférences avec d'autres processus) ; ils traitent de grands volumes et peuvent opérer en parallèle.

Proprioception (ou kinesthésie)

Sensations de tension musculaire, de position et de mouvement, d'équilibre et de déplacement.

Qualia

Expérience personnelle ressentie lorsqu'on perçoit quelque chose (perceptions usuelles, mais aussi sensations corporelles comme douleur, faim, plaisir… ainsi que passions et émotions). Ce sont des effets subjectifs ressentis et associés à des états mentaux. Par définition, les qualias d'autrui sont inconnaissables et incommunicables.

Réflexion

Voir métareprésentation.

Représentation

- Action de rendre quelque chose ou quelqu'un présent sous la forme d'un substitut.
- Reproduction, restitution des traits fondamentaux de quelque chose ou de quelqu'un.
- Ce qui est présent à l'esprit ; ce que l'on « se représente » ; ce qui forme le contenu concret d'un acte de pensée.

Savoir

- Connaître de façon rationnelle en ayant dans l'esprit un système organisé notionnel et psychologique (= connaître).
- Être au courant de, être informé de et/ou sur l'existence ou la nature de quelqu'un ou de quelque chose.
- Avoir des connaissances rationnelles acquises par l'étude, la réflexion et l'expérience, de façon approfondie.

Sémantique

Voir sens ci-dessous.
Pour *sémantique formelle*, voir modèle.

Sens/signification

Trois attitudes en linguistique sur les définitions respectives de ces termes :

- Le sens existe indépendamment du contexte (ou du cotexte) ; il est intrinsèque.
- Le sens sémantique (dénoté « signification ») est premier et le sens pragmatique (dénoté « sens ») s'en déduit selon le contexte ; un élément linguistique possède un ensemble de significations potentielles filtrées pour aboutir au sens en situation.
- L'essentiel de la sémantique est porté par les interactions entre la signification et le sens.
- Ainsi, la signification serait une propriété des signes alors que le sens serait une propriété des textes ou des dialogues.
- On peut alors distinguer (sans qu'il s'agisse le moins du monde d'une partition !) :
  - sémantique *véri-conditionnelle* : précise les conditions de vérité de l'expression traitée (une description formelle des situations dans lesquelles l'expression peut être considérée comme vraie) ;
  - sémantique *intensionnelle* : description d'une expression comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants ;
  - sémantique *extensionnelle* (ou dénotationnelle ou référentielle) : description d'une expression comme l'ensemble des éléments du monde de référence que cette expression peut désigner ;
  - sémantique *componentielle* : cherche à décomposer les mots en éléments de sens plus primitifs ; étudie les possibilités de combinaison de ces éléments ;
  - sémantique *véri-relationnelle* : représentation du sens d'une expression comme l'ensemble des expressions qui peuvent avoir le même sens ;
  - sémantique *procédurale* : description du sens d'une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné ;
  - sémantique *argumentative* : cherche à dépasser la description d'actes de langage isolés pour étudier les enchaînements d'actes dans le discours et les connecteurs qui marquent ces enchaînements. Ces travaux visent à mettre en évidence les marqueurs et les constructions utilisées pour qu'un énoncé puisse servir comme un argument en faveur d'un autre énoncé, et sont liés aux notions de supposition et de présupposition.

Sensation

- Perception consciente produite par une stimulation physiologique (externe [extéroception] ou interne [intéroception ou proprioception]).
- État de conscience affectif.
- Émotion forte.

Sentiment

- Représentation mentale, subjective et privée, d'une émotion rendue consciente ; nécessite des processus cognitifs de haut niveau. [vision de Damasio]
- Conscience plus ou moins claire que l'on a de quelque chose (*avoir le sentiment de son rôle*).
- Faculté de comprendre certaines valeurs (*le sentiment de son devoir*).
- Amour (souvent au pluriel).
- Sensibilité artistique (*jouer avec du sentiment*).
- Bracelet de cheveux tressés (*Vx*).

Signe

- (linguistique, sémiotique) La plus petite unité de sens, provenant de l'association d'un signifiant et d'un signifié.
- (pour Edelman) Une perception d'un élément du monde actuel qui rappelle une image mentale passée (conscience primaire).

Signifiant

Une forme (sonore, écrite, imagée…) qui est associée à un élément possédant une signification. (exemple : le mot « chat » ou une image d'un chat).

Signifié

Le concept associé à un signifiant (c'est-à-dire la représentation mentale d'une chose) (exemple : l'idée qu'on se fait d'un chat [intension], ou l'ensemble des chats existant [extension]).

Spiritualisme

Doctrine qui affirme l'existence de l'esprit comme une réalité supérieure et antérieure à la matière. Cette doctrine proclame également l'existence de valeurs spirituelles et morales.

Subjectivité

- Qualité de ce qui appartient seulement au sujet pensant.
- Qualité de ce qui ne donne pas une représentation fidèle de la chose observée.
- Fait d'être partial ; appréciation, attitude qui résulte d'une perception de la réalité, d'un choix effectué en fonction de ses états de conscience.

Symbole

Signe concret permettant d'évoquer quelque chose d'absent ou d'impossible à percevoir.

On distingue les symboles « naturels » (le rapport avec la chose évoquée préexiste dans la nature), des symboles « arbitraires » (le rapport avec la chose évoquée est totalement conventionnel, comme le noir pour le deuil, le lion pour le courage…).

Pour Edelman, le mécanisme de catégorisation appliqué aux processus de la conscience primaire elle-même va permettre de se libérer du lien au présent : l'état mental résultant d'une perception va pouvoir lui-même être associé à un autre état mental, sans que les caractéristiques du présent soient pertinentes pour cette association.

Exemples :

- symbole naturel « chaussée glissante »  ;
- symbole arbitraire « sens interdit »  ;

En logique formelle, on appelle également « symbole » les éléments de base d'un système.

Syntagme

En analyse syntaxique, ensemble de mots constituant une catégorie intermédiaire entre les catégories lexicales et la phrase (par exemple : groupe nominal, formé d'un article d'un nom et d'un adjectif, ou groupe verbal, formé d'un verbe et de ses compléments).

Syntaxe

Le rôle de la syntaxe est d'expliciter les rôles des différents mots et des syntagmes dans une phrase ou un énoncé, ainsi que les relations qu'ils entretiennent.

Téléologie

Théorie qui suppose que toute chose a une finalité.

Turing (machine de)

Modèle abstrait du fonctionnement des ordinateurs, créé par Alan Turing afin de formaliser la notion d'algorithme.

Ce modèle se fonde sur quatre éléments :

- un ruban (supposé de longueur infinie) composé de cases où sont écrits des symboles ;
- un élément qui lit et écrit les symboles sur le ruban, et se déplace à gauche ou à droite ;
- un registre qui mémorise l'état courant de la machine ;
- une table qui précise, selon le contexte, quel symbole écrire, comment se déplacer et quel est le nouvel état. Si, dans un état donné, aucune action n'existe, la machine s'arrête.

Valeurs

Parmi les dizaines de significations potentielles de ce terme, on ne retient dans cette étude que le sens d'Edelman : forces biologiques primaires (comme le besoin d'alimentation, la reproduction ou la production d'adrénaline) autour desquelles

se produisent implicitement des renforcements ou des inhibitions (Edelman parle de « centres de valeurs » pour désigner les aires cérébrales capables de diffuser dans l'ensemble du cerveau puis de l'organisme des neurotransmetteurs génériques, incitatifs ou inhibiteurs, qui renforcent les réactions globales de l'organisme).

[En particulier, rien à voir avec le sens moral].

Vitalisme

Tradition philosophique pour laquelle le vivant n'est pas réductible aux lois physico-chimiques. Elle envisage la vie comme de la matière, animée d'un principe (une force vitale), qui s'ajouterait pour les êtres vivants aux lois de la matière. Selon cette conception, c'est cette force occulte qui insufflerait la vie à la matière.

En biologie, ce cadre théorique a été un moment fécond, car il dégageait le vivant du mécanisme et des explications causales réductrices du cartésianisme, sans pour autant revenir au surnaturel.

Volonté

Souvent associé à intentionnalité : un acte volontaire implique une intention.

La volonté permet de tempérer, maîtriser, gouverner ou inhiber ses désirs.

Avoir de la volonté implique de la détermination et de la persistance dans la suite d'actions visant à réaliser son objectif.

von Neumann (architecture de, machine de)

Ordinateur formé de quatre composants : l'unité de traitement arithmétique et logique, l'unité de contrôle, la mémoire (qui contient à la fois les données et le programme), les dispositifs d'entrée-sortie. Avec une telle architecture, un ordinateur peut modifier ses instructions durant l'exécution. Cette faculté (utilisée par les virus) est inutile avec les ordinateurs actuels.