Délibérer avec l'intelligence artificielle au service de l'intelligence naturelle.

Frédéric Alexandre, Thierry Viéville, Marie-Hélène Comte

Titre de l'ouvrage: Penser , calculer , délibérer.

Titre du chapitre : Délibérer avec l'intelligence artificielle au service de l'intelligence naturelle.

Ceci est le preprint du chapitre publié dans le livre.

Auteurs:

Frédéric Alexandre
Chercheur en neurosciences cognitives et en intelligence artificielle, Inria et Université de Bordeaux
Thierry Viéville
Chercheur en neurosciences computationnelles et en sciences de l'éducation, Inria et Université Côte d'Azur
avec la relecture et le conseil de Marie-Hélène Comte
Ingénieure pédagogique du Learning Lab Inria, coordinatrice de projets

Résumé : La numérisation de la société et le traitement automatique de l'information, y compris avec des techniques dites d'intelligence artificielle, induisent des ruptures dans notre façon de penser, calculer et délibérer. Mais comment fonctionnent ces fonctions cognitives et comment peuvent-elles être affectées par cette révolution numérique ? Pour comprendre en profondeur cela, nous allons d'abord prendre le temps d'expliquer ce que nous comprenons aujourd'hui de notre intelligence biologique qui pense, ce qui offrira un éclairage crucial sur ce qui se passe quand on utilise des machines qui calculent, avant de conclure en quoi cela aide à réfléchir sur comment délibérer. Car oser comprendre les aspects scientifiques et techniques de la pensée et du calcul est essentiel pour se donner les moyens de délibérer en toute conscience avec les outils intellectuels et numériques qui nous sont donnés.

Mots clés : intelligence artificielle, neuroscience cognitive,

I. Introduction **1**

II. Penser : comment marche le cerveau ? **2**
II.I Comportements dirigés par les stimuli 3
II.II Comportements dirigés par les buts 5

III. Calculer : comment coder et traiter mécaniquement l'information ? **7**
III.I Comment passer de nos connaissances à des données ? 7
III.II Que peut-on demander à une machine de calculer ? 11

IV. Délibérer: comment mettre l'intelligence artificielle à notre service ? **14**
IV.I De la différence entre penser et calculer 14
IV.II Allier pensée et calcul pour délibérer. 15

I. Introduction

Aujourd'hui, avec la numérisation de la société (allant de l'instrumentation et de la mise en ligne des entreprises, au développement d'internet et des réseaux sociaux, et à leur

pénétration dans nos foyers), le stockage et surtout le traitement massif des données s'impose comme une norme pour le traitement de l'information.
En particulier, les progrès récents dans ce qu'on appelle communément[1] intelligence artificielle ont permis la réalisation de dispositifs aux résultats impressionnants : reconnaissances d'objets visuels, traitement de la langue naturelle, outils logiciels d'analyse de données et d'aide à la décision, y compris sur des sujets critiques comme le diagnostic médical, les décisions militaires opérationnelles, ou les arbitrages de justice. Des tâches qui requéraient l'intelligence humaine sont maintenant dévolues à des algorithmes.

Cette efficacité, basée sur du traitement statistique massif et de grande efficacité pour des tâches très spécifiques, finit par imposer ce type de traitement comme la norme d'interaction entre un humain et un dispositif numérique, surtout avec sa généralisation dans notre environnement et sa présence dans tous nos objets numériques. Nous allons questionner cela en explicitant les limites de ces approches et en montrant des alternatives.

De plus, avec le terme d'intelligence artificielle[2], cela laisse à penser qu'il y a un certain degré de similarité avec notre intelligence naturelle. Cela peut aller jusqu'à faire un amalgame (penser par exemple que les machines vont devenir intelligentes au sens biologique du terme ou même qu'elles traitent déjà l'information comme le font les humains, voire que les humains devraient traiter l'information comme elles le font) et nous voulons aider à bien distinguer ce qui relève de la connaissance de ce qui relève de la croyance.

Pour cela, et pour aider à comprendre comment délibérer avec l'intelligence artificielle mise au service de l'intelligence naturelle, entre calculer et penser[3], donc, nous allons expliquer en première partie, avec l'apport des neurosciences et des sciences de la cognition, que nous avons une vision assez élaborée de ce qui se passe dans notre cerveau quand nous traitons de l'information et contrôlons notre comportement. Ainsi, le cerveau n'est pas un organe à traiter massivement des données de façon aveugle et systématique et fonctionne bien différemment des algorithmes statistiques en question ici.

Dans une seconde partie, nous rentrerons dans les détails de la façon dont nous codons les informations quand nous les faisons traiter par des algorithmes et expliciterons le fonctionnement des différentes formes de ce qu'on appelle intelligences artificielles (au pluriel, en évoquant d'autres formes existantes de traitement de l'information, au-delà du traitement statistique massif de type big-data).

Forts de ce double éclairage, qui va nous permettre de bien comprendre la différence entre penser (au sens de l'intelligence humaine) et calculer (sur des valeurs numériques ou symboliques, donc au sens de l'informatique), nous pourrons dans la troisième partie discuter dans quelle mesure, avec quels bénéfices et quelles limites, délibérer avec l'intelligence artificielle au service de l'intelligence naturelle.

II. Penser : comment marche le cerveau ?

---

[1] Ce qu'on appelle intelligence artificielle aujourd'hui est surtout le traitement statistique ``big data´´ : c'est-à-dire de grands volumes de données, de grandes variétés, approximatives, et ceci le plus souvent de manière opaque (traitement par une "boîte noire").
[2] Le terme "artificiel" signifie à la fois, (i) être le produit de l'activité et de l'habileté humaine (opposé à naturel), mais aussi (ii) quelque chose de "factice", c'est à dire imité, donc, qui ne l'est pas vraiment . Il serait plus correct de parler d'«intelligence simulée» qui correspond à la définition reconnue https://fr.wikipedia.org/wiki/Intelligence_artificielle.
[3] On distingue penser de calculer au sens expliqué ici : https://interstices.info/calculer-penser

Au fur et à mesure des progrès des neurosciences et de leurs techniques d'observation, on a pu affiner notre compréhension de ce qui se passe dans notre cerveau, dans différentes circonstances de la vie. Mesurer et décrypter les activités de notre cerveau nous renseigne sur ce qui se passe dans notre vie mentale et nos comportements. Les premières études ont porté sur **la perception et sur l'action** (Lettvin et al. 1968)[4], toutes deux reliées à des phénomènes extérieurs facilement objectivables (Goodale and Humphrey 1998)[5]. Puis on a pu aussi considérer des perceptions internes, dites intéroception (Craig 2003)[6], correspondant au traitement des sensations internes relatives au fonctionnement de notre corps ou relatives à la perception du plaisir et de la douleur. On peut aussi considérer des actions internes qui correspondent aux commandes que nous envoyons vers nos organes internes ou à certaines prises de décision dont les effets ne sont pas immédiatement visibles de l'extérieur (nous précisons cela plus bas). Ces études ont pu permettre de définir, dans notre cerveau, certaines structures comme étant sensorielles (codant et recueillant des informations sensorielles), motrices (codant et envoyant des commandes motrices) ou sensorimotrices (codant et représentant des transformations d'informations sensorielles vers des informations motrices) (M. M. Mesulam 1998)[7].

Ensuite, au-delà de l'immédiateté, le cerveau est aussi **le siège de notre mémoire** et une activité cérébrale peut refléter l'évocation de souvenirs passés ou l'exploitation de contingences apprises. Depuis les travaux fondateurs de E. Kandel (Kandel 2006)[8] sur les mécanismes de la mémoire et les études cliniques de M. Moscovitch et B. Milner sur la dissociation de systèmes de mémoires (Squire and Zola 1996)[9], de nombreuses connaissances ont été accumulées sur l'architecture et la dynamique mnésique. On fera en particulier la distinction entre deux types de mémoires: la mémoire explicite que l'on peut déclarer (« je sais que ») et la mémoire implicite que l'on sait réaliser (« je sais faire ») (Squire 1992)[10]. Ainsi pour le premier type, l'hippocampe nous permet de mémoriser très rapidement un épisode particulier, tel qu'on l'a vécu (on parle de mémoire épisodique), alors que le cortex va mémoriser des relations sémantiques, plus générales et abstraites, et doit être soumis à une répétition fréquente de contingences pour mémoriser lentement ces concepts. Pour le second type, deux structures forment, avec le cortex, ce que l'on appelle la mémoire procédurale. Ce sont les ganglions de la base, qui vont apprendre les comportements motivés ou encore le cervelet pour le contrôle moteur.

Sur la base de ces structures sensorielles, motrices et mnésiques, il est possible de rendre compte d'un certain nombre de comportements et de leur implantation cérébrale. Les comportements peuvent être dirigés par des buts établis par l'individu, mais ils peuvent être également dirigés par des stimuli, en réaction à l'environnement. Nous abordons d'abord cette situation.

---

[4] Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts. 1968. "What the Frog's Eye Tells the Frog's Brain." Edited by W. C. Corning and M. Balaban. The Mind: Biological Approaches to Its Functions, 233–58.

[5] Goodale, M. A., and G. K. Humphrey. 1998. "The Objects of Action and Perception." Cognition 67 (1–2): 181–207

[6] Craig, A. D. 2003. "Interoception: The Sense of the Physiological Condition of the Body." Current Opinion in Neurobiology 13 (4): 500–505. https://doi.org/10.1016/s0959-4388(03)00090-4.

[7] Mesulam, M. M. 1998. "From Sensation to Cognition." Brain 121 (6): 1013–52. https://doi.org/10.1093/brain/121.6.1013.

[8] Kandel, Eric. 2006. A La Recherche de La Mémoire. Odile Jacob.

[9] Squire, L. R., and S. M. Zola. 1996. "Structure and Function of Declarative and Nondeclarative Memory Systems." Proceedings of the National Academy of Sciences of the United States of America 93 (24): 13515–22. https://doi.org/10.1073/pnas.93.24.13515.

[10] Squire, L. R. 1992. "Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory." Journal of Cognitive Neuroscience 4 (3): 232–43.

II.I Comportements dirigés par les stimuli

Dans ce premier cas, on étudie ce qui se produit dans notre cerveau, suite à la perception d'un stimulus quelconque (un objet par exemple). Cette perception peut nous inciter à faire une action dirigée vers cet objet (on parle d'affordance, possibilité d'action suscitée par une perception, cf. (Cisek 2007)[11]). Si cette action est réalisée, cela induira une transformation locale du monde tel qu'on le perçoit (ou éventuellement tel qu'on l'imagine à travers notre mémoire) et tel qu'on le ressent (par intéroception). Par exemple, suite à l'action "saisir", l'objet auparavant sur la table est maintenant dans notre main droite et cela nous procure du plaisir car cet objet est comestible. Cette description nous conduit à préciser deux concepts importants, les réponses et les renforcements.

Une action, ou plus généralement une réponse (interne ou externe) de notre organisme, correspond en fait à une transformation locale du monde, que nous traduisons par une transition de son état sensoriel initial à son état résultant de la réponse. Il y a donc, pour représenter nos comportements et leurs effets sur le monde, une **dualité entre réponses et transitions d'états**. Cette dualité va être permise par la dualité de nos mémoires. Si ce que nous représentons (et pouvons déclencher) dans notre mémoire procédurale correspond à des réponses (en particulier motrices), ce que nous représentons et apprenons dans notre cortex en mémoire sémantique pour représenter nos comportements, est avant tout sensoriel, c'est-à-dire représenté en termes de perceptions externes ou internes (Bindra 1978)[12]. C'est en particulier le rôle du cortex frontal (la partie antérieure de notre cortex) qui va représenter nos comportements, des plus simples aux plus complexes. Nous savons comment cela se fait : le but immédiat de l'action motrice est représenté dans les cortex moteur et prémoteur (Graziano 2006)[13], et la décomposition spatiale et temporelle des tâches se fait plutôt dans le cortex préfrontal latéral (Fuster 2001)[14], tandis que leur motivation et leur but final sont représentés dans le cortex préfrontal médial (Wise 2008)[15].

Ceci nous conduit à parler maintenant de **renforcement** (récompense ou punition) que l'organisme peut recevoir suite au comportement sélectionné en réponse au stimulus. Cela génère des comportements motivés, ce cadre permettant de décrire de nombreux mécanismes d'apprentissage d'organismes vivants (Robbins and Everitt 1996)[16]. Nous évoquons alors deux composantes essentielles de ces phénomènes, **les émotions et les motivations** (Cardinal et al. 2002)[17] et leur rôle majeur dans notre système nerveux (Alexandre 2021)[18] : survivre (maintenir la structure de notre organisme, nous nourrir, nous reproduire, etc.). Pour cela, il est vital de savoir non seulement détecter, mais aussi prédire l'arrivée de stimuli biologiquement importants (d'où l'importance d'avoir un système de

---

[11] Cisek, Paul. 2007. "Cortical Mechanisms of Action Selection: The Affordance Competition Hypothesis." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 362 (1485): 1585–99. https://doi.org/10.1098/rstb.2007.2054.
[12] Bindra, Dalbir. 1978. "How Adaptive Behavior Is Produced: A Perceptual-Motivational Alternative to Response Reinforcements." Behavioral and Brain Sciences 1 (1): 41–52. https://doi.org/10.1017/S0140525X00059380.
[13] Graziano, Michael. 2006. "The Organization of Behavioral Repertoire in Motor Cortex." Annual Review of Neuroscience 29 (March): 105–34. https://doi.org/10.1146/annurev.neuro.29.051605.112924.
[14] Fuster, Joaquín M. 2001. "The Prefrontal Cortex—An Update : Time Is of the Essence." Neuron 30 (2): 319–33. https://doi.org/10.1016/s0896-6273(01)00285-9.
[15] Wise, Steven P. 2008. "Forward Frontal Fields: Phylogeny and Fundamental Function." Trends in Neurosciences 31 (12): 599–608. https://doi.org/10.1016/j.tins.2008.08.008.
[16] Robbins, T. W., and B. J. Everitt. 1996. "Neurobehavioural Mechanisms of Reward and Motivation." Current Opinion in Neurobiology 6 (2): 228–36.
[17] Cardinal, Rudolf N., John A. Parkinson, Jeremy Hall, and Barry J. Everitt. 2002. "Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex." Neuroscience & Biobehavioral Reviews 26 (3): 321–52. https://doi.org/10.1016/s0149-7634(02)00007-6.
[18] Alexandre, Frédéric.. 2021. "A Global Framework for a Systemic View of Brain Modeling." Brain Informatics, February. https://doi.org/10.1186/s40708-021-00126-4.

signalisation plaisir/douleur spécifique) et de pouvoir exprimer des émotions permettant d'anticiper et éventuellement prévenir ou accompagner la survenue de ces stimuli particuliers, qui pourront correspondre à des buts à atteindre ou à éviter car annonciateurs d'effets de l'environnement que l'on nomme parfois punitions et récompenses. Depuis I. Pavlov, ces circuits ont été bien décrits dans le cerveau (Ledoux 2000)[19].

Un autre mécanisme important pour aider à maintenir la structure de notre organisme est celui de nos motivations (O'Reilly 2020)[20] qui permettent justement de définir et de se donner les moyens d'atteindre des buts. Ces buts peuvent être externes (comme de la boisson) et on parlera alors de motivation extrinsèque, déclenchée par un besoin (la soif) qui donnera une importance plus forte à l'action de boire plutôt que d'autres actions qui auraient pu être suscitées par d'autres stimuli. Ces buts peuvent être internes (comme l'acquisition d'une compétence ou d'une connaissance) et on parlera de motivation intrinsèque, pouvant correspondre à la curiosité ou à la pratique assidue de certaines activités.

Ce cadre expliquant les comportements motivés permet d'expliquer certains comportements dirigés par les stimuli perçus. Parmi l'ensemble des réponses possibles à ces stimuli, celles qui, dans le passé, ont fourni le plus de renforcement sont privilégiées, ce que Thorndike a appelé **la loi de l'effet**. Il y a aussi la possibilité de **comportements réflexes**, automatiques, sélectionnés par l'évolution parce que bénéfiques ou automatisés par une pratique intensive (on parlera de comportements habituels, cf. (Boraud, Leblois, and Rougier 2018)[21]). Il y a enfin la nature et l'intensité (on parlera de **saillance**, c'est-à-dire du fait que cela attire l'attention) des stimuli eux-mêmes qui peuvent expliquer qu'on va y répondre en priorité. En résumé, notre attention sera attirée vers les stimuli qui se détachent des autres naturellement ou parce qu'ils appellent des comportements habituels, ou associés par apprentissage à des motivations ou des émotions.

II.II Comportements dirigés par les buts

Nous disposons aussi d'une approche téléologique de l'organisation de nos comportements, à savoir dirigés par des buts. Dans ce cas, un but est considéré en premier, fourni par une motivation ou par une consigne émanant de l'environnement. La sélection du comportement se base sur le choix d'une action parce qu'elle permet de réaliser ce but. Si aucune n'est directement accessible, un raisonnement à rebours permet de définir une série d'actions dont la première peut opérer depuis l'état présent, et passant par des sous-buts intermédiaires contribuant au but final . En contraste avec le cas précédent où la réponse est élaborée simplement en réaction à l'environnement, il est important ici de disposer d'un modèle interne du monde (qu'on pourrait se représenter comme un simulateur mental: on est capable de s'imaginer faire une action et d'en voir les conséquences), nous permettant d'anticiper les conséquences de nos actions. Dans ce cas, le choix de l'action n'est plus basé sur une analyse rétrospective de son efficacité passée (selon la loi de l'effet) mais sur une analyse prospective (selon un modèle interne du monde) de ce qu'une action pourrait nous apporter si on la déclenchait (Dolan and Dayan 2013)[22].

Si on peut supposer que dans ce cas, les mécanismes seront plus complexes que pour un comportement réactif simple, il est important de mentionner que ces mécanismes mentaux internes, en lien avec ce qu'on appelle **imagination**, utilisent les mêmes circuits cérébraux

---

[19] Ledoux, Joseph E. 2000. "Emotion Circuits in the Brain." Annual Review of Neuroscience 23 (1): 155–84. https://doi.org/10.1146/annurev.neuro.23.1.155.
[20] O'Reilly, Randall C. 2020. "Unraveling the Mysteries of Motivation." Trends in Cognitive Sciences 24 (6): 425–34. https://doi.org/10.1016/j.tics.2020.03.001.
[21] Boraud, Thomas, Arthur Leblois, and Nicolas P. Rougier. 2018. "A Natural History of Skills." Progress in Neurobiology 171 (December): 114–24. https://doi.org/10.1016/j.pneurobio.2018.08.003.
[22] Dolan, Ray J., and Peter Dayan. 2013. "Goals and Habits in the Brain." Neuron 80 (2): 312–25. https://doi.org/10.1016/j.neuron.2013.09.007.

(Schacter, Addis, and Buckner 2007)[23]. En fait, l'hippocampe est capable non seulement de mémoriser des épisodes vécus, séquences de perceptions et de réponses émises pour les contrôler en vue d'obtenir un renforcement, mais aussi de rappeler ces épisodes et les rejouer, c'est à dire réactiver les régions du cerveau qui avaient répondu lors de cet épisode, voire de générer des épisodes inédits combinant des épisodes existants (Pezzulo et al. 2014)[24]. En fonction de la situation, on pourra donc rappeler un épisode pertinent ou même plusieurs morceaux d'épisodes différents et on pourra ainsi s'imaginer faire quelque chose que l'on n'a encore jamais fait. Ce mécanisme de contrôle de l'imagination s'appelle la pensée (Pezzulo and Castelfranchi 2009)[25].

Dans les deux cas (la pensée ou l'action directe sur le monde), c'est le cortex frontal qui assure le travail de contrôle (Fuster 2001)[26], c'est à dire de sélection des indices importants à traiter (par l'attention sélective) et d'organisation dans le temps du comportement (par l'inhibition de comportements non adaptés et l'activation de règles de comportement adaptées à la situation, selon la gamme des tâches qu'il représente). Quand ceux-ci sont mis en œuvre par la pensée, ceci permet de **délibérer**, c'est-à-dire d'évoquer plusieurs possibilités et d'imaginer leurs conséquences potentielles, avant de décider. Cette capacité est particulièrement importante chez les humains avec le développement de notre cortex frontopolaire, particulièrement marqué dans notre espèce (Koechlin et al. 1999)[27].

Pour répondre à notre question initiale (que se passe-t-il dans notre cerveau quand nous pensons et agissons ?), nous avons évoqué ici les différents types de comportement que nous pouvons réaliser (Balleine and Dickinson 1998)[28]: **actions réflexes, réponses émotionnelles, comportements dirigés par un stimulus ou par un but**. Dans notre vie courante, ils vont s'articuler et parfois entrer en compétition pour définir concrètement notre comportement (Cisek 2012)[29]. Selon la compréhension actuelle de ces mécanismes, il semble que nous sommes en général mûs par un but, que nous cherchons à résoudre (O'Reilly et al. 2014)[30]. Pour ce faire, nous essayons d'abord d'utiliser des stratégies, qui se présentent comme des règles qui appliquent une action si des conditions sont remplies afin d'obtenir les résultats attendus. Au-delà de ces règles comportementales éprouvées, nous essayons de nous adapter et même de faire preuve de **créativité** si nous détectons que l'environnement a changé et que nos stratégies ne sont plus adaptées (Duverne and Koechlin 2017)[31].

---

[23] Schacter, Daniel L., Donna Rose Addis, and Randy L. Buckner. 2007. "Remembering the Past to Imagine the Future: The Prospective Brain." Nature Reviews Neuroscience 8 (9): 657–61. https://doi.org/10.1038/nrn2213.

[24] Pezzulo, Giovanni, Matthijs A. A. Van der Meer, Carien S. Lansink, and Cyriel M. A. Pennartz. 2014. "Internally Generated Sequences in Learning and Executing Goal-Directed Behavior." Trends in Cognitive Sciences 18 (12): 647–57. https://doi.org/10.1016/j.tics.2014.06.011.

[25] Pezzulo, Giovanni, and Cristiano Castelfranchi. 2009. "Thinking as the Control of Imagination: A Conceptual Framework for Goal-Directed Systems." Psychological Research PRPF 73 (4): 559–77. https://doi.org/10.1007/s00426-009-0237-z.

[26] Fuster, Joaquín M. 2001. "The Prefrontal Cortex—An Update : Time Is of the Essence." Neuron 30 (2): 319–33. https://doi.org/10.1016/s0896-6273(01)00285-9.

[27] Koechlin, E., G. Basso, P. Pietrini, S. Panzer, and J. Grafman. 1999. "The Role of the Anterior Prefrontal Cortex in Human Cognition." Nature 399 (6732): 148–51. https://doi.org/10.1038/20178.

[28] Balleine, Bernard W, and Anthony Dickinson. 1998. "Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates." Neuropharmacology 37 (4): 407–19. https://doi.org/10.1016/S0028-3908(98)00033-1.

[29] Cisek, Paul. 2012. "Making Decisions through a Distributed Consensus." Current Opinion in Neurobiology 22 (6): 927–36. https://doi.org/10.1016/j.conb.2012.05.007.

[30] O'Reilly, Randall C., Thomas E. Hazy, Jessica Mollick, Prescott Mackie, and Seth Herd. 2014. "Goal-Driven Cognition in the Brain: A Computational Framework." ArXiv:1404.7591, 2014. http://arxiv.org/abs/1404.7591.

[31] Duverne, Sandrine, and Etienne Koechlin. 2017. "Hierarchical Control of Behaviour in Human Prefrontal Cortex." In The Wiley Handbook of Cognitive Control, 207–20. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118920497.ch12.

La notion de créativité[32] (Guilford 1962)[33] a une définition assez précise au niveau de la cognition, c'est la capacité de produire un travail à la fois nouveau (c'est-à-dire original, inattendu) et approprié (c'est-à-dire utile, adapté aux contraintes de la tâche). Cela s'exprime dans un double processus de pensée divergente (chercher une nouvelle idée) puis convergente (vérifier qu'elle convient), comme analysé par exemple par (Dietrich 2004)[34]. Ces processus sont modélisés en cognition informatique (Alexandre 2020b)[35].

Ces liens entre activité cérébrale et comportement sont en particulier explorés et confirmés grâce à des expérimentations en imagerie cérébrale qui permettent d'observer l'activation de **réseaux à large échelle** dans notre cerveau, que l'on peut faire correspondre aux grands réseaux neurocognitifs élaborés à partir des théories de divers domaines des sciences cognitives (M. Mesulam 2008)[36]. C'est ainsi que l'on retrouve, pour les comportements dirigés par les stimuli, des réseaux relatifs au langage, à la cognition spatiale ou encore à la saillance (qui attire l'attention, donc). On retrouve par ailleurs des réseaux réalisant des comportements dirigés par le but, comme les réseaux dédiés à l'attention et au contrôle cognitif. Toujours dans ce cadre, on observe aussi le réseau en mode défaut, c'est-à-dire celui qui est activé par défaut, quand la consigne est de ne rien faire et de ne penser à rien. Dans ce cas, on voit un accès à nos sensations internes et à notre mémoire épisodique, pour évoquer des ressentis qui sont en particulier utilisés dans des phases de créativité (Dietrich 2004)[37].

III. Calculer : comment coder et traiter mécaniquement l'information ?

Regardons maintenant[38] ce que nous pouvons faire faire à une machine, à savoir après avoir codé l'information, faire traiter celle-ci par des algorithmes. Expliquer à chacune et

---

[32] C'est la capacité d'un individu à imaginer ou construire et mettre en œuvre un concept neuf, un objet nouveau ou à découvrir une solution originale à un problème. Cette originalité se définit dans un contexte donné. La démarche commence par la reconnaissance d'un problème. C'est à partir de là qu'un processus de divergence s'engage. Ce dernier correspond à une exploration à partir d'éléments connus (la créativité ex-nihilo n'a jamais été observée) de combinaisons ou de déformations inédites. Elle se termine, par convergence, dans une nouvelle solution du problème, par un mécanisme de vérification et d'évaluation de la nouvelle proposition.

[33] Guilford, Joy Paul. 1962. "Creativity: Its Measurement and Development." A Source Book for Creative Thinking, 151–67.

[34] Dietrich, Arne. 2004. "The Cognitive Neuroscience of Creativity." Psychonomic Bulletin & Review 11 (6): 1011–26. https://doi.org/10.3758/BF03196731.

[35] Alexandre, Frédéric. 2020b. "Creativity Explained by Computational Cognitive Neuroscience." In . https://hal.inria.fr/hal-02891491.

[36] Mesulam, M. 2008. "Representation, Inference, and Transcendent Encoding in Neurocognitive Networks of the Human Brain." Annals of Neurology 64 (5): 367–78.

[37] Dietrich, Arne. 2004. "The Cognitive Neuroscience of Creativity." Psychonomic Bulletin & Review 11 (6): 1011–26. https://doi.org/10.3758/BF03196731.

[38] Nous proposons ici une vision synthétique et invitons la lectrice ou le lecteur à approfondir ces notions en profitant de formations numériques attestées et librement utilisables:
- https://classcode.fr/iai qui introduit ces notions d'intelligence artificielle symbolique et numérique de manière très concrète mais sans aucun prérequis technique;
- https://classcode.fr/snt qui fournit une formation citoyenne de base en informatique qui correspond à ce que les élèves du secondaire apprennent désormais pour se former sur ces sujets;
et pour aller plus loin:
- https://www.elementsofai.fr permet de se former aux éléments plus techniques de l'intelligence artificielle;
- https://tinyl.io/3kpr permet de se former à l'intelligence artificielle symbolique utilisée, entre autres, au niveau du web sémantique.
Cette section est une synthèse de ces éléments en vue de répondre à la question posée ici.

chacun ces grandes idées de l'informatique va nous permettre de comprendre le fonctionnement de ce qu'on appelle intelligence artificielle.

III.I Comment passer de nos connaissances à des données ?

Nous parlons usuellement de connaissances[39] lorsque notre esprit assimile un contenu objectif. Ce contenu est préalablement traduit en signes et en idées. C'est donc une possession symbolique des choses, qui permet l'émergence du sens. Mais comment coder cela dans une machine ? Autrement dit comment réduire ces connaissances à des données ? Regardons cela.

***Nommer les choses* :** rien n'existe si il n'est pas nommé. C'est bien le cas en informatique, à tous les niveaux, et tout particulièrement avec l'omniprésence d'internet.
Chaque "ressource" (on nomme ainsi tous les objets immatériels qui sont considérés) a un identifiant unique, ces fameuses constructions telles que :

https://fr.wikipedia.org/wiki/Internationalized_Resource_Identifier

On parle d'« identificateurs de ressources internationalisés » (IRI), d'aucun parle -improprement- « d'adresse internet ».
Avoir un nommage unique de chaque ressource permet de la distinguer des autres, donc de pouvoir s'y référer de manière bien définie quand on la manipule.
De plus, quand cela est pertinent, cet identificateur correspond à son emplacement sur Internet, ce qui fournit les éléments pour y accéder. On parle alors de « localisateur uniforme de ressource » (URL en anglais).
Cet identificateur a aussi parfois des paramètres[40], qui permet de piloter son fonctionnement ou interagir avec elle : par exemple l'adresse internet de votre chauffage connecté permet de régler à distance la température de la maison.

Ces IRI sont un peu plus que des "noms". Ce sont des locutions d'un petit langage qui explicite le processus pour interagir avec la ressource, on parle de protocole, et toute autre information utile pour y accéder (la machine où elle se trouve et le chemin sur cette machine) ou la contrôler (par exemple les paramètres pour envoyer un message).

Cette entreprise humaine collective est immense :

+ elle est universelle au sens où tous les alphabets du monde sont utilisables pour former ces identificateurs;
+ tous les pays du monde, même en guerre les uns contre les autres, adoptent ces mêmes *standards* pour encoder, les lettres, puis les mots, puis les locutions.

Imaginez un groupe qui imaginerait réinventer ses propres standards : coupé du reste du monde, il n'existerait tout simplement plus.

Nommer ces ressources est indispensable quand elles sont accessibles par internet, mais c'est aussi le cas pour des informations traitées localement au sein d'un logiciel. En effet, les entrées et sorties de ces traitements ne feraient pas de sens si elles n'étaient pas correctement nommées.

---

[39] https://fr.wikipedia.org/wiki/Connaissance#Définition_de_la_connaissance

[40] Essayez vous-mêmes ! Par exemple entrez :
https://fr.wikipedia.org/wiki/Internationalized_Resource_Identifier?action=raw
dans un navigateur avec le paramètre ``*action=raw*´´ en plus de l'adresse de la page, vous allez voir un affichage étonnant : vous venez de donner l'ordre au site web de fournir le contenu de la page, mais sous la forme qu'utilisent les personnes qui écrivent dans wikipédia pour l'éditer, en utilisant une syntaxe bien définie https://www.mediawiki.org/wiki/Help:Formatting.

Interagir avec les objets numériques **:** cette démarche concerne aussi les objets connectés qui nous entourent (éléments robotiques d'une voiture, pacemaker de notre voisin, robot cuisinier qui nous guide pour faire de bonnes recettes), par exemple, cet identificateur :

http://192.168.1.137/?action=start&temperature=20&time=17:00

pourrait, disons, lancer mon chauffage domestique à partir de mon smartphone. Ces objets deviennent accessibles et utilisables, car ils sont *nommés*. Même si nous n'utilisons pas directement ces identificateurs, lorsque nous cliquons sur des "menus" d'applications, en fait, ce sont de tels codes qui sont générés.

Voilà une première forme d'intelligence artificielle qui nous entoure et qui est *déjà là* : notre quotidien se remplit d'objets connectés auxquels nous déléguons plein de petites tâches cognitives et sensori-motrices élémentaires spécifiques et parfois rudimentaires (faire garer sa voiture toute seule, soulager sa mémoire d'un rappel de rendez-vous, transformer notre parole en texte, traduire d'une langue humaine à une autre) et c'est cela que l'on appelle aujourd'hui couramment intelligence artificielle. Nous utilisons cela sur nos smartphones, bien entendu, mais cela existera aussi de plus en plus au sein de beaucoup de biens manufacturés.

De l'information aux nombres **:** Une monumentale entreprise de numérisation s'est appliquée à toutes les informations humaines, tout est numérique désormais. Pour cela, nous avons standardisé la façon de coder toutes les lettres des alphabets, attribuant à chaque lettre un code numérique, ce nombre étant lui-même représenté en "binaire" c'est-à-dire par une suite de '0' et de '1'. Nous avons fait de même avec les images découpées en points (dits "pixels") assez petits pour donner à l'œil humain une impression de continuité, la couleur de chaque point étant codée par quelques nombres. Il en va de même des sons, l'intensité sonore étant découpée en une séquence de valeurs numériques, donc une myriade de '0' et de '1'. Et en combinant ces éléments, des vidéos aux textes illustrés, voilà toutes nos bibliothèques et médiathèques devenues numériques.

Pourquoi en binaire ? La raison profonde est que si un paramètre peut prendre deux valeurs possibles ('0' ou '1' ou bien 'oui' ou 'non', peu importe) alors cela crée un *atome* d'information. S'il n'y avait qu'une seule valeur, on ne pourrait rien dire. Ensuite, en combinant plusieurs éléments binaires (bits) on peut coder n'importe quelle liste d'éléments, par exemple, avec deux bits on codera les trois couleurs primaires:

rouge -> '00', vert -> '01', bleu -> '10'

et avec plus de bits toute une palette de couleurs, et au-delà toute information. Pour s'en convaincre il suffit de jouer au "portrait" (appelé aussi "qui-est-ce"), à deviner tout ce qui est possible au fil de simples réponses par oui ou non.
Une autre belle propriété du codage binaire est que si vous encodez les nombres en binaire:

0 -> '0000', 1 -> '0001', 2 -> '0010', 3-> '0011', …

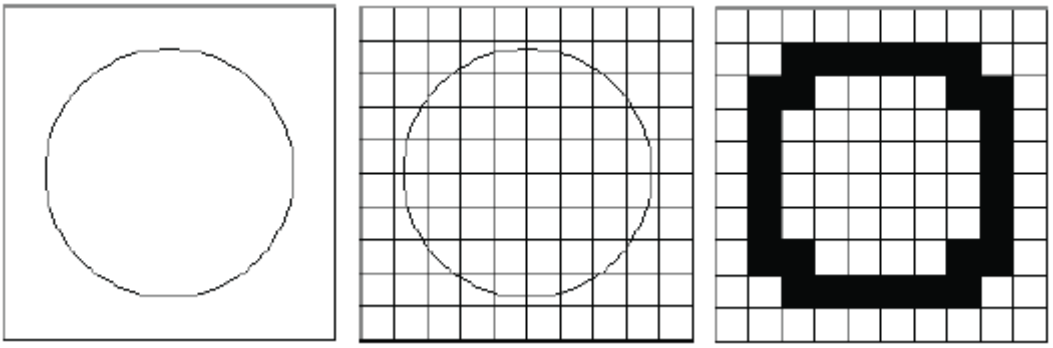
alors tous les calculs numériques usuellement faits en numération décimale fonctionnent exactement pareil en binaire, avec des règles de calcul tellement plus simples qu'il est extrêmement efficace de les câbler électriquement, même s'il faut environ quatre bits pour coder un chiffre.

Le fait que toutes ces informations se codent en binaire a plusieurs conséquences:
+ toutes les opérations générales qui ne dépendent pas du contenu, comme mémoriser, transmettre, compresser, chiffrer (pour les rendre inaccessibles sans une clé), dupliquer ces données, se font avec des algorithmes universels, et à très faible coût;
+ le fait que dupliquer ces données est une opération à coût quasi nul, bouleverse l'économie de ces biens informationnels: on achète plus un livre, par exemple, mais un droit à lire un texte.

***Information symbolique et numérique* :** Il y a deux grandes façons de coder une information, par exemple un dessin ou bien un air de musique:

- en approximant *numériquement* l'image du dessin (avec des pixels) ou bien l'enregistrement sonore de cet air de musique (les intensités sonores successives);
- en définissant *symboliquement* ce dessin, par exemple en spécifiant les formes à dessiner, ou cet air de musique en fournissant la partition de musique à jouer.

Regardons l'exemple d'un cercle ci-après, qui est codé sur une grille de pixels, en noircissant chaque pixel qui rencontre le cercle : on observe que le rendu est peu ressemblant, il faudrait une grille de pixels bien plus fine, pour que l'approximation numérique soit plus fidèle. On pourrait aussi coder ce cercle en donnant la position de son centre et de son rayon, ce qui le définit parfaitement et permettrait de le tracer. On dit alors que le cercle est défini "symboliquement" par les éléments symboliques en question.



Dans le premier cas, numérique, l'information est forcément approximative mais nous pouvons encoder des formes complexes de la même manière que des formes simples, dans le second cas symbolique, on offre une description sans approximation, mais limitée. Par exemple, pour l'air de musique, ce sera sans les nuances de l'interprète.

***Information humaine et informatique* :** L'information, avec sa numérisation, est devenue une quantité physique qui se mesure,

- elle a un "poids" c'est le nombre de bits pour la coder de manière non redondante (en cherchant à la compresser au maximum de manière réversible);
- lorsque nous générons cette information avec un programme (par exemple un calcul) on définit aussi sa "complexité" c'est-à-dire la taille du programme dont on a besoin à minima pour la générer et sa "profondeur" c'est-à-dire le temps de calcul minimal que mettra ce programme pour la produire.

Ce sont des notions complètement inédites qui font basculer la notion d'information telle que nous l'entendons dans la vie courante vers une notion formalisée.

Quand on utilise des techniques de codage de l'information comme par exemple le codage binaire, cela ne signifie en rien que l'on en capture le sens. Par exemple, tous les idéogrammes chinois sont bien codés en binaire au niveau lexical, cela ne signifie pas que toute la subtilité de la langue a été formalisée. Des outils de représentation de connaissances comme les ontologies informatiques permettent de formaliser partiellement le sens de concepts; les résultats sont parfois très performants, mais c'est aussi en regardant les limites de ces spécifications que notre intelligence humaine peut aller plus loin dans l'analyse du sens des choses étudiées.

Un cran plus loin, nous sommes amenés à distinguer langue (humaine) et langage (formel) au sens de Dowek (Dowek 2019)[41]. Les deux utilisent des symboles, mais une langue formelle a volontairement un vocabulaire contrôlé limité et une grammaire minimale, et

[41] Dowek, Gilles. 2019. Langues et langages : Ce dont on ne peut parler, il faut l'écrire. Le Pommier. https://www.editions-lepommier.fr/ce-dont-ne-peut-parler-il-faut-lecrire.

s'applique à un champ sémantique précis, avec un objectif opérationnel précis, différent de ceux d'une communication humaine. Il y a des langues formelles non informatiques : une partition de musique, par exemple, s'écrit dans un *langage* musical formel, pour faire fonctionner un instrument de musique, tandis que l'interprétation musicale va dépasser ce qui est codé dans la partition, pour devenir l'expression d'une *langue* musicale.

Le langage juridique cherche à être le plus formel possible, et se heurte ainsi à deux grandes limites : (i) tant qu'il sera écrit en langue humaine, il pâtira des ambiguïtés de cette langue (à l'inverse d'une poésie qui en joue et parfois s'en joue); (ii) dans la mesure où il s'applique à des actions humaines, son objet est impossible à appréhender de manière exhaustive et hors de portée d'une spécification complète. Le but du droit est justement de prendre en compte la singularité même du cas instruit, au-delà d'appliquer une règle.

À l'inverse, "informatiser" est un processus qui vise à *réduire une information à des données*. Sur wikipédia, par exemple, il y a un texte qui va décrire en langue humaine une ressource, mais il y aussi wikidata qui extrait de ce texte toutes les informations, on parle parfois de métadonnées, qui peuvent se réduire à une liste de faits. Une personne, par exemple, pourrait -selon les besoins- être réduite aux éléments de son état civil (nom, adresse, profession), aux compétences de son curriculum-vitae et la liste de ses activités socioprofessionnelles, à ses antécédents judiciaires, à des traits de caractère prédéfinis, aux éléments du dossier médical, aux relations formelles avec son entourage (fille-de, proche-de, en-procès-avec, …). Juger (dans tous les sens, on en discutera à la section IV.II. du terme) une action sur ces éléments a quelque chose d'équitable et de rigoureux, mais peut sembler terriblement inhumain : on propose en fin de chapitre des lectures pour approfondir cet aspect.

C'est donc en premier lieu par la façon de coder les données, les choix de représentation que nous faisons, ce que nous prenons en compte et ce que nous négligeons, que nous transformons, en les rendant artificielles, les connaissances de notre intelligence naturelle.

III.II Que peut-on demander à une machine de calculer ?

Nous voilà avec des données factuelles, que pouvons nous en faire ? Pour cela il faut maintenant représenter et spécifier le traitement de ces données. Nous allons regarder maintenant comment traiter les informations quand elles se présentent sous forme de symboles, par exemple les trier, tirer des conséquences des faits, etc…

Coder le l'information symbolique pour la traiter **:** prenons ce fait « محمد بن موسى » :

Muhammad est-le-fils-de Mūsā

Nous énonçons une connaissance minimale de la forme "sujet prédicat objet", et il est effectivement important de structurer et de modulariser au maximum nos informations, plutôt que de les livrer amalgamées à la machine. De ce fait, on en déduit plusieurs choses : si il est le *fils* de Mūsā et non la fille, alors c'est un homme, donc un humain, donc un animal, puisque tous les humains le sont. Et si on sait par ailleurs que Mūsā est le fils de Aïcha alors on en déduit par transitivité que Muhammad est le petit-fils de Aïcha, etc. Cela se code par des règles de la forme:

?x est-le-fils-de ?y ⇒?x est-du-genre masculin ET ?x est humain

?x est-le-fils-de ?y ET ?y est-le-fils-de ?z ⇒?x est-le-petit-fils-de ?z

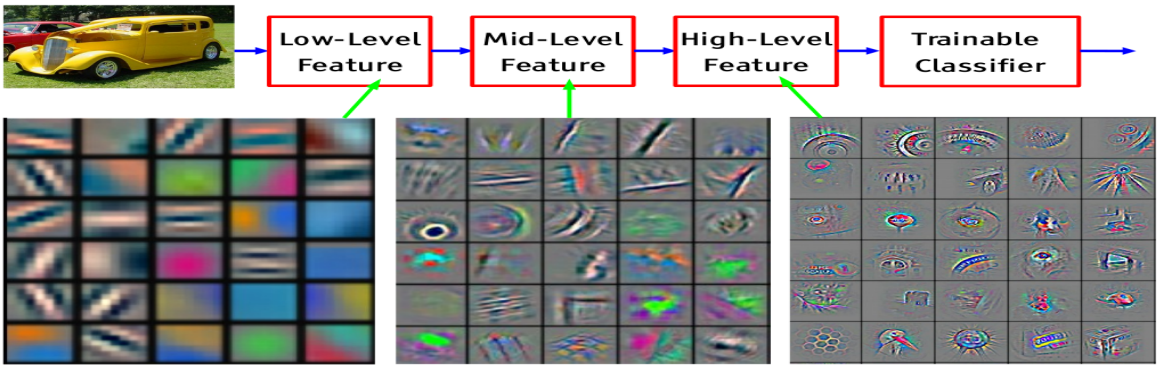
Ce qui est important pour nous ici est que nous pouvons coder des règles et que l'application de ces règles permet de déduire d'autres faits. Cela permet de trouver les conséquences d'une hypothèse, d'expliciter des informations implicites qui découlent des données en entrée, de vérifier si un ensemble de faits ne génère pas une contradiction, d'interroger une base de données sur ce qui a été explicitement défini et tout ce qui en découle.

Cela signifie que nous pouvons *mécaniser* dans une certaine mesure des raisonnements logiques, qui se réduisent à des calculs sur des symboles. On peut alors mener des raisonnements qui dépassent de beaucoup la capacité cognitive humaine. Il y a de multiples langages et de multiples outils en informatique pour cela. On démontre même des résultats mathématiques avec des systèmes de déduction automatiques, ce qui permet de vérifier qu'il n'y a pas eu d'erreur humaine dans les déductions (il peut y en avoir au niveau des règles et des faits qui ont été définis). Les plus modernes et efficaces qui nous concernent sont ceux du Web sémantique : on parle d'ontologie informatique pour nommer cette formalisation des connaissances.

Mieux encore, on sait aussi comment mécaniser des calculs "incertains", par exemple si on pense qu'"il est relativement possible" que Muhammad soit le fils de Mūsā, alors il est relativement possible aussi que ce soit un homme. On quantifie très souvent cette certitude relative par des probabilités, mais il est plus réaliste de formaliser cela par des notions plus proches de la cognition humaine, en utilisant les notions de nécessité et de possibilité.

On connaît aussi les limites intrinsèques à ces mécanismes : certaines formulations peuvent donner lieu à des calculs qui ne finissent jamais ou dont le temps de calcul est prohibitif (facilement plus que l'âge de l'univers, disons); on tombe aussi sur des formules qui ne peuvent être ni démontrées, ni infirmées : elles sont indécidables. Plus important encore est le fait qu'il est démontré que pour obtenir des résultats très sophistiqués (de grande profondeur logique, pour utiliser le terme exact), il faudra un temps de calcul très long.

Paramétrer le traitement numérique de l'information **:** la méthodologie précédente marche bien pour des informations qui se formalisent aisément. Mais si on doit, par exemple, distinguer un chat d'un chien dans une image, alors il sera bien difficile de "formaliser" cela avec des règles symboliques en tenant compte de toutes les variantes possibles. Il y a alors une toute autre famille de solutions, basées sur ce que l'on nomme réseaux de neurones artificiels.



Source: Zeiler, Matthew D., and Rob Fergus. Visualizing and understanding convolutional networks. Computer Vision ECCV 2014. Springer International Publishing, 2014. 818 833

Prenons une image en entrée, comme illustré ci-dessus[42]. Accumulons des petits calculs entre les pixels voisins, par exemple des moyennes ou des différences, pour ne considérer que les valeurs les plus importantes du résultat obtenu. Puis, à la sortie de cette couche de calculs, considérons d'autres calculs qui combinent les premiers de la même manière, et ainsi de suite, jusqu'à un calcul dont la sortie sera une valeur entre 0 et 1 pour dire si c'est plutôt un "chat" qu'un "chien". On parle de "**neurone**" pour désigner le petit calcul élémentaire, et de réseau "profond" de neurone pour expliciter le fait que l'on accumule une pile profonde de couches de calcul.

[42] Reproduit avec autorisation de Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014. Springer International Publishing, 2014. 818–833

Un neurone combine donc plusieurs entrées dans des proportions ajustables, ces proportions définissent les paramètres du calcul.

Une telle architecture jouit d'une propriété remarquable : c'est une transformation qui peut réaliser de manière approximative grâce aux statistiques, *une très grande famille de transformations.* Pour cela on ajuste les paramètres de chaque neurone, à partir d'exemples : on fournit des images de chiens et de chats, et chaque erreur d'estimation donne lieu à une correction progressive des paramètres. Après quelques millions (plutôt que milliers) d'exemples on observe que cet **apprentissage statistique** fonctionne : il y a peu d'erreurs parmi les animaux utilisés pour l'apprentissage, mais *aussi* parmi des animaux jamais utilisés mais qui sont statistiquement cohérents avec les exemples fournis. On dispose de résultats empiriques qui permettent d'estimer la robustesse d'un tel calcul.

Le résultat n'est pas "miraculeux" : nous avons toutes et tous l'habitude de la "déraisonnable efficacité" des statistiques, quand -par exemple- on prédit le résultat d'une élection à partir d'un sondage sur un petit échantillon d'une population. C'est une situation similaire ici, on a en quelque sorte fait un sondage sur un échantillon de chiens et de chats et la prédiction s'applique alors à toute une population.

Ce que ces résultats montrent par exemple en matière de reconnaissance d'images, c'est que le problème n'est pas si compliqué que cela. Prenons, par exemple, la transcription de la parole en mots: les sons de la parole humaine, par exemple, se décomposent en une cinquantaine de sons élémentaires (on parle de phonèmes, qui correspondent plus ou moins aux sons des voyelles ou des consonnes) et se combinent pour former environ 1000 à 10000 mots selon le niveau de langage. Ce sont des quantités assez raisonnables au regard des millions de calculs que peut faire une machine; il n'est donc pas surprenant, à condition d'utiliser assez de données, de pouvoir transformer la parole en texte, bien entendu sans *rien* en comprendre, ni faire la moindre analyse linguistique du contenu

Ces systèmes sont programmés "par les données". Personne n'explicite les traitements à faire. On se contente de définir une architecture en choisissant les nombre de couches, le nombre et le type de calculs élémentaires par couche, puis de lancer l'ajustement des paramètres en fournissant avec des méthodologies assez précises les données en très grand nombre. En ce sens, le système est un mécanisme d'**apprentissage automatique**.

Ce sont ces algorithmes qui permettent, à ce qu'on appelle le **big data**, d'offrir aujourd'hui des résultats spectaculaires en matière de reconnaissance visuelle ou sonore, et de prédiction de toutes sortes, y compris sur des données peu modélisables. Ainsi commence-t-on à prédire des décisions de justice, comme discuté dans la troisième partie.

Bien entendu un tel système de transcription ne comprend rien, le texte transcrit ne fait aucun sens pour lui, c'est juste un calcul. De plus, ces calculs sont des "boîtes noires" : il est assez hasardeux en regardant l'intérieur des calculs de pouvoir expliciter grâce à quelles caractéristiques on différencie, par exemple, un chien d'un chat. Par ailleurs, ils sont d'autant plus efficaces que leur tâche est très spécifique : distinguer un chien d'un chat sera bien plus efficace qu'un chien de tous les animaux du zoo, et si le système a appris à distinguer un chat d'un chien, le travail d'apprentissage est à refaire si on change d'animal. Pire encore, on sait créer des images "antagonistes" c'est-à-dire prendre une image de chat et identifier les quelques pixels qu'il faut modifier, pour le faire prendre pour un chien. Et de façon étonnante, ces pixels peuvent être très peu nombreux et les humains ne pas se laisser flouer par ces changements qui ne modifient que la distribution statistique des pixels et pas leur organisation structurelle.

Ce sont des algorithmes “universels” déraisonnablement efficaces, mais difficiles à interpréter, très coûteux en données, spécifiques d’une tâche cognitive donnée, et très fragiles en matière de performance.

***Quelques autres splendides algorithmes* :** on vient de voir comment faire faire à la machine des calculs symboliques ou numériques qui auraient été qualifiés d’intelligents s’ilsavaient été fait par une personne humaine. Dans ces deux cas, on ne programme pas explicitement des instructions à faire exécuter par la machine, mais on fournit des connaissances symboliques ou des données numériques qui vont permettre à l’algorithme de résoudre le problème posé. Il y a en fait plusieurs autres façons de produire des calculs sophistiqués. On peut par exemple programmer “par contraintes” : fournir les données du problèmes, les contraintes à respecter et le but à obtenir, par exemple, pour planifier une suite d’actions ou résoudre un problème d’ordonnancement (exemple: gestion de stocks ou de transport d’objets).

Bien entendu, on peut aussi programmer de manière “impérative[43]” c’est-à-dire expliciter manuellement les instructions du calcul. Tous les langages et outils de programmation usuels, par exemple ceux qui permettent de s’initier à l’informatique comme Scratch[44], disposent de cette base impérative. Elle permet de programmer des algorithmes parfois utilisés à de très grandes échelles, par exemples les algorithmes de tri (manipuler un ensemble de données triées est souvent bien plus efficace qu’en vrac) ou de calcul d’un plus court chemin (qui peut servir à la fois à un itinéraire routier ou à trouver comment gagner à un jeu en trouvant “le chemin” vers une situation dominante, etc…). Dans tous les cas, pour réaliser l’interface entre tous ces mécanismes, il faut impérativement savoir programmer de manière impérative pour faire le lien entre tous ces outils.

IV. Délibérer: comment mettre l’intelligence artificielle à notre service ?

IV.I De la différence entre penser et calculer

Dans la première partie consacrée à l’intelligence naturelle, nous avons d’abord parlé de la nécessité, pour un être vivant, d’interagir avec son environnement à travers ses perceptions, ses actions et les traces mnésiques qu’il en conserve. C’est avec ces éléments qu’il peut élaborer des comportements qui peuvent permettre de répondre aux stimulations de l’environnement aussi bien que d’obéir à ses buts internes, principalement relatifs à sa survie. Ce que montrent ensuite les neurosciences et en particulier au cours de l’évolution, c’est la montée en gamme des stratégies utilisées pour réaliser des traitements de plus en plus complexes. Le fait qu’un cerveau ait “plusieurs millions d’années” d’expériences accumulées[45] est probablement la cause première de ces performances, on montre en effet que pour produire des résultats sophistiqués, un calcul ou un processus quel qu’il soit doit réaliser un nombre d’étapes vertigineusement grand. De ce fait, on a pu associer ces traitements à des phénomènes comme la pensée, la créativité ou même la conscience, mais

---

[43] Par impératif on entend le fait qu’il permettent de définir des séquences d’instructions, d’affecter des valeurs à des variables, d’exécuter des instructions conditionnelles ou des boucles, Ils disposent aussi d’autres abstractions comme des fonctions, des objets informatiques, qui “encapsulent” des instructions pour les manipuler plus efficacement.

[44] https://scratch.mit.edu permet à chacune et chacun d’apprendre en manipulant et en jouant les bases de l’informatique, le langage https://www.python.org est lui très utilisé au niveau scolaire. Il est bien établi que pour comprendre comment marchent tous ces mécanismes, y compris ceux que l’on manipule avec des données, il faut des bases en informatique ce qui est maintenant le cas pour tous nos enfants, heureusement.

[45] Voir par exemple https://www.epi.asso.fr/revue/articles/a1302b.htm pour une discussion grand public sur les différences cerveau / ordinateur.

finalement le principe reste le même: l'être vivant et singulièrement l'humain, traite les données qu'il perçoit pour les transformer en informations utilisables pour contrôler son environnement et en connaissances pour mieux les généraliser et les transmettre à ses congénères[46].

Dans la seconde partie consacrée au traitement automatique de l'information, nous avons indiqué que la force principale de cette approche est l'accès aux régularités cachées dans les données par leur traitement statistique. C'est ici que ces techniques peuvent supplanter l'humain, par leur capacité de calcul mécanique à grande échelle et à haute fréquence que nous ne pouvons pas atteindre. Mais c'est aussi leur limite: c'est un traitement massif de données où le cas particulier ne peut pas être pris en compte et issu d'un calcul purement mécanique qui n'a pas accès au sens ou à l'interprétation des concepts manipulés.

Par opposition au traitement automatique de l'information qui, par le calcul, cherche à extraire les régularités statistiques de l'environnement, le but pour le vivant est d'en extraire son sens, pour les dimensions qui concernent sa survie et son intérêt propre: on ne doit pas lire des milliards de phrase ou voir des millions d'images de chat pour apprendre à parler ou reconnaître un animal. On essaie simplement d'en extraire ce qui est suffisant pour notre compréhension du monde et de ses règles, afin de mieux pouvoir l'utiliser pour notre bénéfice. On passe ainsi notre temps à (très bien) reconnaître (interpréter plutôt) des scènes ou des phrases qu'on n'a encore jamais vues ou entendues, avec un système pouvant "comprendre" car ayant déjà "vécu" des choses similaires. On utilise des "données" en nombre très raisonnable, mais dans un espace de représentation interne qui est extrêmement grand et structuré. La complexité d'un cerveau est de l'ordre de grandeur de tous les ordinateurs du monde[8].

Pour approfondir cette différence entre penser et calculer[3] John Searle, propose l'expérience de pensée, dite de la chambre chinoise[47]. Dans cette chambre, un opérateur (humain ou artificiel) pourrait appliquer des règles syntaxiques pour répondre à des questions écrites en chinois, sans connaître cette langue. Il pourrait faire illusion pour un observateur extérieur mais pour autant, il ne saurait pas parler chinois car il n'aurait pas accès au sens des phrases qu'il construit. C'est bien ici l'importance de la sémantique et sa dissociation d'une représentation symbolique qui est questionnée.

IV.II Allier pensée et calcul pour délibérer.

Délibérer[48] ? « Examiner, peser tous les éléments d'une question avec d'autres personnes, ou éventuellement en soi-même, avant de prendre une décision, pour arriver à une conclusion, en pesant le pour et le contre, de façon à décider par un débat ». On voit à la fois la dimension intrinsèquement humaine du concept, mais aussi la dimension *collégiale* : il s'agit dans la plupart des cas de délibérer au pluriel de la pensée de plusieurs cerveaux.

Comment délibérer au mieux aujourd'hui, où nous vivons à l'ère numérique et au temps des algorithmes[49], comme le discutent (Abiteboul and Dowek 2017)[50] ? Ce qui a été partagé

---

[46] Cette compréhension est d'ailleurs extrêmement fructueuse pour concevoir d'autres mécanismes d'intelligence artificielle, bien différents de ce qui se fait avec ce qu'on appelle le "big data", comme mis en lumière ici (Alexandre 2020a), déjà cité.
[47] Voir https://fr.wikipedia.org/wiki/Chambre_chinoise pour une présentation de cette expérience et https://plato.stanford.edu/entries/chinese-room pour une discussion complète y compris critique de cette expérience de pensée, le choix de la langue chinoise n'étant peut-être pas le plus pertinent compte tenu de la structure de la langue.
[48] Cité de http://atilf.atilf.fr à consulter pour une belle définition très complète : https://www.cnrtl.fr/definition/deliberer
[49] On pourra lire (Abiteboul and Dowek 2017) le « temps des algorithmes » pour une réflexion plus poussée sur ces mutations par deux grands collègues en science informatique.
[50] Abiteboul, Serge, and Gilles Dowek. 2017. Le Temps Des Algorithmes. Le pommier.

précédemment offre un double éclairage : à la fois sur ce qui peut-être délégué à la machine, avec ses limites, et sur ce qui reste l'apanage de la pensée humaine.

On comprend aisément que ces algorithmes peuvent servir "d'extension cognitive" comme les applications de nos smartphones qui complètent notre mémoire, nous aident à nous orienter géographiquement, ou transcrivent en texte notre parole. Ce sont des outils qui se mettent dans la boucle de notre interaction avec le monde, comme lorsque nous prenons une perche pour rallonger la portée de notre bras. C'est à nous de ne pas inverser les rôles.

Pourrions nous arriver à une situation où un outil hyper-puissant va rendre son verdict, c'est-à-dire le résultat de ses calculs sur les données engrangées et l'humain n'aura plus qu'à s'exécuter ? Pourquoi pas … mais ce serait notre choix. Imaginons bien plus simplement s'en remettre au verdict[51] d'un tirage à pile ou face. Et il y a des situations extrêmes (par exemple de survie) où cela reste le dernier outil pour départager toute autre raison écartée. Dans ce cas c'est bien l'humain qui décide de s'en remettre à ce mécanisme, ce n'est pas le tirage au sort qui est responsable. S'en remettre à une "IA" quelle qu'elle soit, est un choix *humain*. Nous faire croire que "c'est pas de notre fait c'est l'IA" est un oxymore.

Cela ne veut pas dire que l'occurrence de ces nouvelles technologies est sans influence sur notre façon de penser, bien au contraire. Comme l'a très bien théorisé Gilbert Simondon[52], la réalité technique fait partie intégrante de la réalité humaine et par les outils qu'elle génère transforme la société humaine dans laquelle elle est conçue et donc transforme sa culture, et la pensée humaine elle-même.

Ainsi, ce qu'on nomme "intelligence artificielle" ne serait pas liée au fait que les machines vont devenir "intelligentes comme nous" mais (au-delà de leur déraisonnable efficacité pour des tâches précises), induirait le fait que notre *vision[53] de l'intelligence humaine* doit évoluer. Voyons deux aspects.

L'informatique nous force à réfléchir et à concevoir la meilleure façon de représenter nos connaissances, mais cela serait vrai même si nous ne confions pas leur traitement à une machine, mais à notre intelligence biologique, on parle alors de pensée informatique. On peut aussi penser qu'une bonne partie du travail est faite quand l'humain a réfléchi à la manière de présenter le problème à la machine: c'est souvent cet effort de codage qui permet de présenter le problème sous un angle qui le rend simple à traiter. Nous voilà donc en train de penser non plus directement à la solution d'un problème mais à la meilleure façon de la calculer, manuellement ou algorithmiquement.

De même que nous déléguons à la machine le soin de nous transporter, ou faire les calculs numériques, en laissant le soin à notre cerveau de choisir comment faire faire cela, et quoi faire du résultat, nous devons procéder de même avec ces nouveaux outils. Avant il fallait penser puis délibérer, maintenant il faut penser deux fois : penser comment calculer, et délibérer en repensant face au résultat du calcul.

---

[51] Philosophiquement, le hasard est un mécanisme absolument impartial, sans aucune subjectivité, ne prenant en compte aucun facteur discutable, au contraire de tout autre algorithme qui est sujet à débat car il est le résultat de choix.

[52] On lira à ce propos l'excellent texte introductif de Gérard Giraudon https://www.lemonde.fr/blog/binaire/2020/04/05/informatique-culture-et-technique-le-schisme-de-simondon sur cette pensée philosophique.

[53] Si ces "processus" sont juste là pour "optimiser l'organisation du travail humain", alors comme cela a déjà été observé lors de la révolution industrielle cela va juste appauvrir le travail humain, et nous transformer en machine, inversant le rôle entre outil et artisan.

Cela impose une chose importante : former chacune et chacun (qui délibèrent et à propos de qui on délibère) aux fondements du numérique, donc à l'informatique, pour comprendre ce qui a été fait et vérifier aussi comment cela a été fait. Voilà un moyen pour que l'intelligence artificielle nous rende collectivement plus intelligents.

Pour aller plus loin.

Voici quelques références supplémentaires qui traite des liens entre justice et algorithmes

- Un algorithme capable de prédire les décisions des juges : vers une robotisation de la justice ? (Barraud 2017)[54] https://www.cairn.info/revue-les-cahiers-de-la-justice-2017-1-page-121.htm
- Comment va se rendre la justice au temps des algorithmes ? https://www.lemonde.fr/blog/binaire/2019/11/25/comment-va-se-rendre-la-justice-au-temps-des-algorithmes/
- La justice prédictive et l'égalité devant la loi. https://www.lemonde.fr/blog/binaire/2019/11/25/comment-va-se-rendre-la-justice-au-temps-des-algorithmes/
- Magie numérique et défis juridiques. https://www.lemonde.fr/blog/binaire/2021/02/05/magie-numerique-et-defis-juridiques/

Bibliographie.

Frédéric Alexandre : Directeur de Recherche Inria, en sciences du numérique, responsable de l'équipe Inria Mnemosyne, hébergée à l'Institut des Maladies Neurodégénératives sur le NeuroCampus de Bordeaux. Dans son parcours en Intelligence Artificielle puis en neuroscience cognitive, il cherche à déchiffrer les mécanismes de notre architecture cognitive, en considérant en particulier comment nos différentes formes de mémoires contribuent à nous faire appréhender le monde et à élaborer notre pensée.

Thierry Viéville : Directeur de Recherche Inria, membre de Mnemosyne et du laboratoire LINE de l'INSPÉ de Nice en sciences de l'éducation, ancien chargé de mission Inria pour la médiation scientifique et la formation des enseignantes et enseignants en science informatique et fondements du numérique dans le secondaire. Il contribue à modéliser de manière pluri-disciplinaire l'apprentissage humain ou mécanique.

Marie-Hélène Comte : relectrice et conseillère sur ce chapitre, ingénieure pédagogique et documentaliste, co-autrice de formations citoyennes en ligne à la pensée informatique et d'initiation à l'intelligence artificielle citoyenne.

Références.

Abiteboul, Serge, and Gilles Dowek. 2017. *Le Temps Des Algorithmes*. Le pommier.

Alexandre, Frédéric. 2020a. "Les relations difficiles entre l'Intelligence Artificielle et les Neurosciences." *Interstices*, August. https://hal.inria.fr/hal-02925517.

———. 2020b. "Creativity Explained by Computational Cognitive Neuroscience." In . https://hal.inria.fr/hal-02891491.

———. 2021. "A Global Framework for a Systemic View of Brain Modeling." *Brain Informatics*, February. https://doi.org/10.1186/s40708-021-00126-4.

Balleine, Bernard W, and Anthony Dickinson. 1998. "Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates." *Neuropharmacology* 37 (4): 407–19. https://doi.org/10.1016/S0028-3908(98)00033-1.

Barraud, Boris. 2017. "Un algorithme capable de prédire les décisions des juges : vers une

[54] Barraud, Boris. 2017. "Un algorithme capable de prédire les décisions des juges : vers une robotisation de la justice ?" Les Cahiers de la Justice N° 1 (1): 121–39.

robotisation de la justice ?" *Les Cahiers de la Justice* N° 1 (1): 121–39.
Bindra, Dalbir. 1978. "How Adaptive Behavior Is Produced: A Perceptual-Motivational Alternative to Response Reinforcements." *Behavioral and Brain Sciences* 1 (1): 41–52. https://doi.org/10.1017/S0140525X00059380.
Boraud, Thomas, Arthur Leblois, and Nicolas P. Rougier. 2018. "A Natural History of Skills." *Progress in Neurobiology* 171 (December): 114–24. https://doi.org/10.1016/j.pneurobio.2018.08.003.
Cardinal, Rudolf N., John A. Parkinson, Jeremy Hall, and Barry J. Everitt. 2002. "Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex." *Neuroscience & Biobehavioral Reviews* 26 (3): 321–52. https://doi.org/10.1016/s0149-7634(02)00007-6.
Cisek, Paul. 2007. "Cortical Mechanisms of Action Selection: The Affordance Competition Hypothesis." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (1485): 1585–99. https://doi.org/10.1098/rstb.2007.2054.
———. 2012. "Making Decisions through a Distributed Consensus." *Current Opinion in Neurobiology* 22 (6): 927–36. https://doi.org/10.1016/j.conb.2012.05.007.
Craig, A. D. 2003. "Interoception: The Sense of the Physiological Condition of the Body." *Current Opinion in Neurobiology* 13 (4): 500–505. https://doi.org/10.1016/s0959-4388(03)00090-4.
Dietrich, Arne. 2004. "The Cognitive Neuroscience of Creativity." *Psychonomic Bulletin & Review* 11 (6): 1011–26. https://doi.org/10.3758/BF03196731.
Dolan, Ray J., and Peter Dayan. 2013. "Goals and Habits in the Brain." *Neuron* 80 (2): 312–25. https://doi.org/10.1016/j.neuron.2013.09.007.
Dowek, Gilles. 2019. *Langues et langages : Ce dont on ne peut parler, il faut l'écrire*. Le Pommier. https://www.editions-lepommier.fr/ce-dont-ne-peut-parler-il-faut-lecrire.
Duverne, Sandrine, and Etienne Koechlin. 2017. "Hierarchical Control of Behaviour in Human Prefrontal Cortex." In *The Wiley Handbook of Cognitive Control*, 207–20. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118920497.ch12.
Fuster, Joaquín M. 2001. "The Prefrontal Cortex—An Update : Time Is of the Essence." *Neuron* 30 (2): 319–33. https://doi.org/10.1016/s0896-6273(01)00285-9.
Goodale, M. A., and G. K. Humphrey. 1998. "The Objects of Action and Perception." *Cognition* 67 (1–2): 181–207.
Graziano, Michael. 2006. "The Organization of Behavioral Repertoire in Motor Cortex." *Annual Review of Neuroscience* 29 (March): 105–34. https://doi.org/10.1146/annurev.neuro.29.051605.112924.
Guilford, Joy Paul. 1962. "Creativity: Its Measurement and Development." *A Source Book for Creative Thinking*, 151–67.
Kandel, Eric. 2006. *A La Recherche de La Mémoire*. Odile Jacob.
Koechlin, E., G. Basso, P. Pietrini, S. Panzer, and J. Grafman. 1999. "The Role of the Anterior Prefrontal Cortex in Human Cognition." *Nature* 399 (6732): 148–51. https://doi.org/10.1038/20178.
Ledoux, Joseph E. 2000. "Emotion Circuits in the Brain." *Annual Review of Neuroscience* 23 (1): 155–84. https://doi.org/10.1146/annurev.neuro.23.1.155.
Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts. 1968. "What the Frog's Eye Tells the Frog's Brain." Edited by W. C. Corning and M. Balaban. *The Mind: Biological Approaches to Its Functions*, 233–58.
Mesulam, M. 2008. "Representation, Inference, and Transcendent Encoding in Neurocognitive Networks of the Human Brain." *Annals of Neurology* 64 (5): 367–78.
Mesulam, M. M. 1998. "From Sensation to Cognition." *Brain* 121 (6): 1013–52. https://doi.org/10.1093/brain/121.6.1013.
O'Reilly, Randall C. 2020. "Unraveling the Mysteries of Motivation." *Trends in Cognitive Sciences* 24 (6): 425–34. https://doi.org/10.1016/j.tics.2020.03.001.
O'Reilly, Randall C., Thomas E. Hazy, Jessica Mollick, Prescott Mackie, and Seth Herd. 2014. "Goal-Driven Cognition in the Brain: A Computational Framework." *ArXiv:1404.7591*, 2014. http://arxiv.org/abs/1404.7591.
Pezzulo, Giovanni, and Cristiano Castelfranchi. 2009. "Thinking as the Control of Imagination: A Conceptual Framework for Goal-Directed Systems." *Psychological Research PRPF* 73 (4): 559–77. https://doi.org/10.1007/s00426-009-0237-z.
Pezzulo, Giovanni, Matthijs A. A. Van der Meer, Carien S. Lansink, and Cyriel M. A. Pennartz. 2014. "Internally Generated Sequences in Learning and Executing Goal-Directed Behavior." *Trends in Cognitive Sciences* 18 (12): 647–57. https://doi.org/10.1016/j.tics.2014.06.011.

Robbins, T. W., and B. J. Everitt. 1996. “Neurobehavioural Mechanisms of Reward and Motivation.” *Current Opinion in Neurobiology* 6 (2): 228–36.
Schacter, Daniel L., Donna Rose Addis, and Randy L. Buckner. 2007. “Remembering the Past to Imagine the Future: The Prospective Brain.” *Nature Reviews Neuroscience* 8 (9): 657–61. https://doi.org/10.1038/nrn2213.
Squire, L. R. 1992. “Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory.” *Journal of Cognitive Neuroscience* 4 (3): 232–43.
Squire, L. R., and S. M. Zola. 1996. “Structure and Function of Declarative and Nondeclarative Memory Systems.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (24): 13515–22. https://doi.org/10.1073/pnas.93.24.13515.
Wise, Steven P. 2008. “Forward Frontal Fields: Phylogeny and Fundamental Function.” *Trends in Neurosciences* 31 (12): 599–608. https://doi.org/10.1016/j.tins.2008.08.008.