Laboratoire de
CYBERJUSTICE
Laboratory

Rapport sur l'épistémologie de l'intelligence artificielle (IA)

Justin-Éric BOILEAU
Ilona BOIS-DRIVET
Hannes WESTERMANN
Jie ZHU

Document de travail n°32

Juin 2022

Table des matières

CHAPTER 2
POTENTIALS A0ND LIMITATIONS OF ARTIFICIAL INTELLIGENCE (AI) METHODOLOGIES AND APPLICATIONS

SOMMAIRE EXÉCUTIF

Ce **volume 1** du rapport sur l'épistémologie de l'intelligence artificielle (IA) vise à présenter une vue panoramique de l'état de situation épistémologique sur l'intelligence artificielle ou augmentée de ses origines à nos jours et au-delà. De sa racine grecque alliant « connaissance » (*épistémê*) et « discours » (*lógos*), l'épistémologie renvoie à l'étude critique sur l'état et les limites de nos connaissances. Le présent volume est structuré en trois chapitres reproduisant le triptyque des questions épistémologiques de type gnoséologique (le « quoi »), méthodologique (le « comment ») et axiologique (le « pourquoi ») s'interrogeant sur les conditions de validité de toute connaissance.

Le **chapitre premier** retrace tout d'abord les origines multidisciplinaires de l'IA comme résultant d'un chassé-croisé de disciplines scientifiques connexes et des courants épistémologiques principaux s'intéressant aux modes de production et de partage de connaissances ainsi que leurs rapports à la vérité. Depuis que l'expression « intelligence artificielle » a été employée pour la première fois lors de la conférence de Dartmouth de 1956, les progrès de l'IA ont été marqués par l'opposition, puis une conjonction récente, de deux approches – symbolique et connexionniste – de modélisation de l'apprentissage. Dans le cadre de nos travaux et sous réserve de nouvelles percées dans l'évolution de la technologie, nous adoptons la définition suivante de l'IA, telle que proposée par le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle (GEHN IA) de la Commission européenne :

> Les systèmes d'intelligence artificielle (IA) sont des systèmes logiciels (et éventuellement matériels) conçus par des êtres humains et qui, ayant reçu un objectif complexe, agissent dans le monde réel ou numérique en percevant leur environnement par l'acquisition de données, en interprétant les données structurées ou non structurées collectées, en appliquant un raisonnement aux connaissances, ou en traitant les informations, dérivées de ces données et en décidant de la/des meilleure(s) action(s) à prendre pour atteindre l'objectif donné. Les systèmes d'IA peuvent soit utiliser des règles symboliques, soit apprendre un modèle numérique. Ils peuvent également adapter leur comportement en analysant la manière dont l'environnement est affecté par leurs actions antérieure.[1] [nos soulignements]

Quoique le Groupe d'experts emploie l'expression « intelligence artificielle », nous nous alignons avec la suggestion de l'ingénieur Luc Julia[2] pour y préférer l'expression « intelligence augmentée », comme décrivant mieux la raison d'être et l'utilité de ces systèmes qui jouent avant tout un rôle d'assistance aux humains pour optimiser (en temps, efficacité, précisions et possibilités) le traitement de l'information et l'exécution de certaines tâches.

---

1 Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle (GEHN IA), *Lignes directrices en matière d'éthique pour une IA digne de confiance*, Commission européenne, 2018 au para 143, en ligne : <ec.europa.eu/newsroom/dae/document.cfm?doc_id=60427>.

2 Luc Julia, *L'intelligence artificielle n'existe pas*, FIRST, 2018.

Le **chapitre 2** présente dans les grandes lignes la manière dont l'IA – depuis les (premiers) systèmes symboliques à l'apprentissage automatique, puis à l'apprentissage profond – acquière et génère de la connaissance exploitable par l'humain tout en automatisant, de plus en plus, le processus d'acquisition et de production de nouvelles connaissances par l'apprentissage automatique. Une description de haut niveau des principales techniques (apprentissage supervisé, apprentissage non supervisé, apprentissage par renforcement) et modèles algorithmiques utilisés est présentée, de même que les atouts et limites de chacun. Il s'ensuit quelques pistes de réflexion générales sur les défis qui se posent actuellement sur la voie vers l'« intelligence artificielle forte ». Au-delà de l'optimisation des résultats prédictifs ou de la précision d'exécution de tâches précises, il y aurait lieu de développer chez la machine, des compétences (d'apprentissage) transversales facilitant une meilleure adaptabilité et polyvalence accrue aux situations incertaines, impliquant le traitement de données incomplètes ou ambigües et faisant appel au « sens commun », sorte de conscience situationnelle diffuse. Le développement d'une métacognition « artificielle » mérite également d'être creusé pour permettre à la machine de prendre lui-même en compte, comme une variable parmi d'autres, dans l'appréciation d'une situation, et ce, afin de raffiner ses prédictions et mieux ajuster ses réactions par rapport aux attentes, actions et réactions (mentales) des humains et d'autres robots.

Le **chapitre 3** remet en perspective l'attitude enthousiaste des États face aux dernières avancées de l'IA à l'aulne de l'économie de la promesse et du cycle d'excitation de Gartner. Ailleurs comme ici, et au Québec en particulier, les politiques publiques sur l'IA et l'innovation contribuent à alimenter un climat plein d'effervescence qui tend à la fois à surestimer le bénéfice social dont on peut en tirer à long terme, qu'à sous-évaluer le montant d'investissements massifs nécessaires pour tirer parti d'une course à l'hégémonie politique que se disputent les grandes puissances, tant pour le pouvoir de coercition de l'IA (p.ex. défense nationale) que son pouvoir de convaincre (*soft power*) avec la ramification de ses implications socioéconomiques et politiques. Ce cycle d'engouements technoscientifiques pour l'IA n'en est cependant qu'un parmi d'autres avant lui, alimentés par autant de promesses suivies de la phase de découragement et de désillusion, avant d'atteindre la phase de plateau avec la maturation des nouvelles technologies.

Ce premier volume sur l'épistémologie de l'IA trace ainsi la voie pour une discussion éclairée des enjeux tant à court qu'à long terme que pose l'intelligence artificielle ou augmentée (AI) sur notre société à l'ère de l'information. Une collection d'articles portant sur les différents thèmes et explorant en profondeur la (con)jonction du droit et de l'IA compose notre **second volume**.

Mots clés : intelligence artificielle, intelligence augmentée, systèmes experts, apprentissage automatique, apprentissage profond, économie de la promesse, promesses technoscientifiques

***

INTRODUCTION GÉNÉRALE

Effet de mode ou acquis durable, l'intelligence artificielle ou augmentée (IA) est sans doute l'une de ces expressions ambivalentes qui suscitent autant d'espérance que le scepticisme. Comme promesse, elle irait jusqu'à ressusciter la légende salvatrice de Prométhée, « providence des hommes », subtilisant à notre profit un savoir – technique – au potentiel immense, jusqu'alors réservé aux dieux, inaccessible aux profanes.

En tant qu'un outil d'assistance parmi d'autres, l'IA nous surprend par une autonomie croissante dans l'exécution de différentes tâches (pas seulement répétitives). À la rescousse du journalisme, l'IA peaufine l'analyse de données financières, valide les sources d'information et participe à la création de contenu standardisé ou simplifié. Aux juristes et autres professionnels, l'IA facilite la veille réglementaire et automatise la mise en conformité. Aux écoliers, l'IA propose des leçons et exercices adaptés aux besoins, forces et difficultés de chacun. Tandis que les navettes autonomes circulent dans nos quartiers, les assistants « intelligents » nous étonnent tant par des répliques qui tombent à point que leurs fonctionnalités de plus en plus diversifiées. D'assistants virtuels aux robots partenaires, l'intelligence artificielle ne tarde pas à revendiquer une personnalité juridique … en devenir[3].

Comme technologie, l'IA se démarque par son caractère proactif dans la construction des connaissances. Des recommandations de produits à l'évaluation des risques (p.ex. de récidive pénale et criminelle, de catastrophes naturelles, d'infections), de la gestion de nos pourriels à la justice prédictive, de la reconnaissance de tumeurs à la détection de séismes, les algorithmes « intelligents » manient avec adresse l'art de la quantification, étirent notre logique binaire sur une gamme d'infinies possibilités et instillent une dose de maîtrise quantifiée dans le chaos d'un quotidien *a priori* marqué par le hasard, l'inconnu et l'incertitude. Au milieu de cette effervescence, un contexte pandémique a propulsé le potentiel de l'IA à l'avant-plan, de la détection du virus au développement des vaccins, en passant par les applications de traçage comme outils de contrôle de la propagation de la pandémie[4]. L'analyse « intelligente » des données de masse permettent également de mieux comprendre et de prévoir les effets et complications de la maladie en fonction de l'âge, du sexe, des antécédents médicaux ainsi que d'autres caractéristiques biomédicales des personnes infectées.

---

3 *Résolution du Parlement européen du 16 février 2017 contenant des recommandations à la Commission concernant des règles de droit civil sur la robotique*, 2015/2103(INL), Strasbourg, 16 février 2017, en ligne : <www.europarl.europa.eu/doceo/document/TA-8-2017-0051_FR.html>.

4 Christophe Mondin et Nathalie de Marcellis-Warin, *Recension des solutions technologiques développées dans le monde afin de limiter la propagation de la COVID-19 et typologie des applications de traçage*, Observatoire international sur les impacts sociétaux de l'IA et du numérique, octobre 2020, en ligne : <www.docdroid.com/VLokunh/recension-des-solutions-technologiques-developpees-dans-le-monde-afin-de-limiter-la-propagation-de-la-covid-19-et-typologie-des-applications-de-tracage-pdf>.

Avec le design génératif, l'IA culmine avec le raffinement d'une intuition artificielle[5] explorant rapidement l'espace de solutions possibles et repoussant tant les frontières de l'esthétique que de l'aérodynamique dans la conception de différents projets d'architecture, d'animation et de design automobile. Ou alors, interrogeant candidement le chaos de l'espace-temps et la persistance des illusions cognitives, l'artiste artificiel titille notre imaginaire avec ses compositions inédites et impressions amplifiées qui évoquent un peu Dalí; la vivacité de Le Greco; les jeux de lumière de Turne; et les perspectives multipliées de Picasso. Il y a plus, dans ces évasions artificielles propulsées par un autre système nerveux, que les « moutons électriques » de (simples) androïdes[6].

La force la plus remarquable de l'IA est sans doute cette capacité – transversale – à s'investir dans une multitude de disciplines connexes et sa vocation à s'appliquer dans des secteurs très variés. Il nous serait bien difficile en effet de nommer un domaine – depuis le mélange de saveurs culinaires inédites à la lutte contre le changement climatique[7] – qui ne soit pas teinté, de près ou de loin, par les progrès de l'IA.

Il n'est alors guère étonnant que l'intelligence artificielle (IA) mobilise à la fois le secteur privé (de la recherche et de l'industrie), la société civile et les États ainsi que des instances régionales et intergouvernementales, tous soucieux de s'allier une discipline émergeant avec d'autant d'expectatives. Certains n'hésitent pas à en comparer l'avènement aux découvertes charnières qui ont marqué l'histoire de l'humanité comme la domestication du feu[8], le moteur propulsant une quatrième révolution industrielle[9] ou encore l'avènement d'un nouvel âge de l'humanité[10].

En même temps, le rythme – effréné – de ce progrès effraie. Dans l'espace d'une vie humaine, l'informatique est passée d'une curiosité technologique à une discipline névralgique pour notre société. De l'automatisation des tâches manuelles au raffinement de (nos) capacités cognitives jusqu'à l'émergence d'une intuition artificielle, l'IA, comme discipline, nous confronte à d'autant d'incertitudes aux plans des capacités technologiques de la machine que des conséquences à

---

5 Dasong Wang, « From Self-Organizing to Self-Intuitive Intelligence : Innovative Methodology of Generative Architectural Design and Fabrication Process with Machine Learning » (2018), en ligne : <dasong-wang.com/2018/11/13/self-organizing-to-self-intuitive-intelligence-2/>.

6 *Cf.* Philip K Dick, *Do Androids Dream of Electric Sheep?*, 1968, en ligne : <files.cercomp.ufg.br/weby/up/410/o/Phillip_K._Dick_-_Do_Androids_Dream_of_Electric_Sheep_c%C3%B3pia.pdf>.

7 Programme des Nations Unies pour l'Environnement, *Définir le rôle de l'intelligence artificielle dans la prévision, l'atténuation et l'adaptation aux impacts du changement climatique*, 2019, en ligne : <fermun.org/wp-content/uploads/2019/11/UNEP_1_-FRANCAIS.pdf>.

8 Denis Guthleben, *La fabuleuse histoire des inventions; de la maîtrise du feu à l'intelligence artificielle*, Dunod, 2021.

9 MIT Technology Review Insights, « The promise of the fourth industrial revolution », *MIT Technology Review* (19 novembre 2020), en ligne : <www.technologyreview.com/2020/11/19/1012165/the-promise-of-the-fourth-industrial-revolution/>; Klaus Schwab, « The Fourth Industrial Revolution », *Encyclopedia Britannica* (23 mars 2021), en ligne : <www.britannica.com/topic/The-Fourth-Industrial-Revolution-2119734>.

10 Jason Thacker, *The Age of AI : Artificial Intelligence and the Future of Humanity*, 2020.

long terme de l'automation sur la société telle que nous la connaissons, qu'il s'agisse de l'avenir du monde du travail, de la résilience du secteur financier, de la protection de l'intégrité électorale ou encore de la lutte contre la cybercriminalité.

***

Avant de discuter des enjeux soulevés par l'utilisation de l'IA, il est temps que nous portions un regard épistémologique sur l'intelligence artificielle ou augmentée (IA) comme objet d'étude. De sa racine grecque alliant « connaissance » (*épistémê*) et « discours » (*lógos*), l'épistémologie renvoie à l'étude critique sur l'état et les limites de nos connaissances. Depuis l'intuition platonicienne sur l'existence d'un monde idéal au-delà des ombres de la caverne, l'épistémologie se démarque en tant qu'une branche de la philosophie moderne s'intéressant aux conditions de validité, ainsi qu'aux modes d'acquisition de nos connaissances. Qu'est-ce qui distingue, à un niveau fondamental, la fausseté de la connaissance vraie ? Une croyance justifiée ou partagée suffit-elle à fonder la connaissance ? Comment connaît-on ce que l'on connaît ? Quel rôle joue le sujet connaissant sur la construction de « sa » connaissance ? Quels sont les rapports entre doute, certitude, connaissance et illusion ? Voilà autant de questionnements qui, depuis le cogito cartésien (« je pense, donc je suis »), ont fasciné tant les rationalistes, exaltant la toute-puissance de la raison, que les partisans de l'empirisme, insistant sur l'apport de l'expérience sensible comme fondement de toute connaissance.

Adopter une perspective épistémologique, cela signifie s'intéresser moins à ce que l'on connaît, que ce que l'on peut connaître ainsi que les conditions de validité de sa connaissance, à la fois sur les plans de l'objet, de la méthodologie et de la portée/valeur d'une discipline. Ce triptyque de l'épistémologie sur « le statut, la méthode et la valeur de la connaissance »[11]a été énoncé par l'épistémologue constructiviste Jean-Louis Le Moigne comme comportant trois questions fondamentales :

1. **Question gnoséologique ou le « quoi »** : Qu'est-ce que la connaissance ? Elle cherche à délimiter l'objet de la discipline étudiée.

2. **Question méthodologique ou le « comment »** : Comment la connaissance s'est-elle constituée ? Elle permet de comprendre la manière dont on accède à cette connaissance.

3. **Question axiologique ou le « pourquoi »** : Comment apprécier la valeur et la finalité de cette connaissance ? Elle interroge l'attitude des sujets face à la connaissance et interpelle les disciplines dites normatives s'intéressant au « devoir-être » tels le droit, l'éthique et la politique.

Transposant ces questions à notre étude de l'IA, ce **premier volume sur l'épistémologie de l'IA** sera structuré en trois chapitres :

---

[11] Jean-Louis Le Moigne, *Les épistémologies constructivistes*, coll « Que sais-je? », Presses Universitaires de France, 2012 à la p 3.

Le **chapitre premier** retrace tout d'abord l'origine du concept de l'intelligence artificielle (IA) en mettant en exergue sa parenté avec plusieurs disciplines (scientifiques) connexes dans leur investigation croisée du fonctionnement de l'esprit humain. Les sources de l'intelligence artificielle se recoupent en effet avec plusieurs questionnements qui ont irrigué nos traditions philosophiques et scientifiques, notamment dans les domaines de la psychologie et des neurosciences, de la linguistique, des mathématiques et de l'ingénierie informatique (*section 1.1*). Depuis que l'expression « intelligence artificielle » a été présentée en 1955 lors de la fameuse conférence de Darmouth, son développement a été marqué par l'opposition entre deux courants principaux, symbolique et connexionniste, sur des frontières mouvantes qui rendent assez difficile la caractérisation précise de la discipline (*section 1.2*). D'où l'importance de clarifier la portée de certains concepts connexes, avant d'adopter une définition de l'IA qui soit assez précise sur le plan opérationnel et souple pour s'adapter aux progrès technologiques (prévisibles). Au terme de ce tour d'horizon, il convient également de nous demander si l'expression « intelligence artificielle », héritée de Darmouth, est toujours appropriée pour décrire une intelligence plutôt « augmentée » (*section 1.3*).

Les précisions liminaires posées, le **deuxième chapitre** s'intéresse aux aspects plus techniques sous-tendant les différentes approches méthodologiques de l'IA. Il cherche à répondre, de manière nuancée et critique, à la deuxième question fondamentale de l'épistémologie : « [À] quelles conditions une machine pourrait-elle connaître ? »[12] la présentation suit l'ordre chronologique du développement de l'approche symbolique (*section 2.1*), de l'apprentissage automatique (*section 2.2*), puis de l'apprentissage profond (*section 2.3*). Une perspective pragmatique, ciblant les domaines d'application ainsi que les forces et faiblesses des différentes approches, est privilégiée, puisqu'il est important de comprendre qu'aucune approche n'est préférable à d'autres dans l'abstrait; et l'approche symbolique est loin d'être périmée en raison de son apparition plus tôt dans l'histoire de la discipline. Il n'y a pas, en soi, de « bonne » ou de « mauvaise » approche; ni ne sont-elles mutuellement exclusives. Le choix des différentes méthodes dépend de l'objectif du programmeur, des tâches attendues ainsi que des possibilités technologiques disponibles. Immanquablement, la maturation des différentes approches méthodologiques (hybrides) caresse le rêve d'une « intelligence artificielle forte ». À cet égard, quoique l'atteinte (du mythe) de la singularité technologique reste à voir, plusieurs pistes de réflexion existent et sont mises en application avec succès (*section 2.4*).

Virant à un prisme politico-sociologique, le **chapitre troisième** cherche à explorer la rencontre de l'intelligence artificielle avec la (pression de) mondialisation. Parler de « choc culturel » à cet égard est sans doute un abus de langage. Mais de l'effet de foule à l'économie de la promesse, une discussion critique sera apportée au rôle des États dans l'élaboration et la mise en œuvre des politiques de soutien à l'innovation (technologique), en prenant pour étude de cas la politique de l'intelligence artificielle (IA) au Québec.

---

12 *Cf.* Serge Robert, « Réflexion épistémologique sur l'intelligence artificielle et les sciences cognitives : à quelles conditions une machine pourrait-elle connaître? » (1992) 2:2 *Horizons philosophiques* 167, doi : <doi.org/10.7202/800901ar>.

Dans une perspective épistémologique, la question axiologique interroge davantage que les attentes, de la population ou de l'État, sur le potentiel technique de l'intelligence artificielle (IA). Elle interpelle également et surtout le domaine de l'éthique ainsi que l'éventail des enjeux juridiques soulevés par l'utilisation de cette technologie. Ces questions seront abordées, par thématique et collection d'articles, dans le **second volume de notre rapport**.

CHAPTER 1
EPISTEMOLOGY OF ARTIFICIAL INTELLIGENCE

Observations préliminaires

Comme juristes, nous n'avons ni les capacités, ni l'expertise d'expliquer avec précision l'ensemble des théories et techniques associées à l'intelligence artificielle, d'autant plus que celles-ci dépassent ces quelques pages. Il s'agit ici de présenter, au meilleur de notre compréhension, les principaux concepts liés à l'intelligence artificielle et ses subtilités pour mieux saisir le flou qui l'entoure et chercher à comprendre l'objet de la discipline. Les différentes méthodes et applications de l'intelligence artificielle seront davantage discutées dans le chapitre 2 du document de travail.

Introduction

Déjà bien intégrée à notre quotidien, l'intelligence artificielle (IA) surfe sur une vague « optimiste » où tout semble possible. Sur toutes les lèvres, l'intelligence artificielle apparaît aujourd'hui comme un terme générique qui désigne un ensemble de technologies connexes, ce qui tend à créer une confusion sur ce qu'est l'intelligence artificielle et ce qu'elle peut faire. Pourtant, la discipline de l'IA ne s'est pas développée du jour au lendemain et porte des idées qui habitent notre imaginaire depuis longtemps. Ainsi, pour mieux comprendre l'intelligence artificielle et ce qu'elle est aujourd'hui, il convient d'abord d'étudier l'origine du concept de l'intelligence artificielle (IA) en mettant en exergue sa parenté avec plusieurs disciplines (scientifiques) connexes dans leur investigation croisée du fonctionnement de l'esprit humain. Les sources de l'intelligence artificielle se recoupent en effet avec plusieurs questionnements qui ont irrigué nos traditions philosophiques et scientifiques, notamment dans les domaines de la psychologie et des neurosciences, de la linguistique, des mathématiques et de l'ingénierie informatique (*section 1*). La section 2 servira ensuite à retracer l'histoire de l'intelligence artificielle au travers des deux principaux courants qui ont marqué l'évolution de la discipline : le courant connexionniste et symbolique. Finalement, avant de proposer une définition de l'IA qui soit assez précise sur le plan opérationnel et souple pour s'adapter aux progrès technologiques (prévisibles), il sera question de présenter ce que l'on entend par intelligence artificielle. Ce tour d'horizon servira ensuite à se demander si l'expression « intelligence artificielle », héritée de Darmouth, est toujours appropriée pour décrire une intelligence plutôt « augmentée » (*section 3*). L'objectif de ce chapitre est donc d'étudier les origines et la formation des idées de l'IA afin de pouvoir poser un regard critique sur les différentes approches méthodologiques de l'IA (*chapitre 2*) et comprendre notre rapport contemporain à la discipline (*chapitre 3*).

Section 1 - Sources of Artificial Intelligence

Developing artificial Intelligence (AI) is an exciting multidisciplinary enterprise. At its source, AI reaches out to the progress and insights gained from different areas of knowledge. To construct an Artificial Intelligence that closely mimics Intelligence as we know it, principles of logical

reasoning drawn from philosophy come at the forefront of our story (1). As the "queen of all sciences", philosophy lays the ground for further scientific studies of the human mind (2), of language (3) and mathematical modeling of real-world dynamics (4). Evidently the whole process is supported by advances in computer engineering (5). Meanwhile the "myth" of AI is also being entertained in popular imagination, (religious) myths of progress and up to the fantasy world of modern science-fiction (6).

1. Philosophy

If the goal of AI is to create thinking machines, then we must first understand what it means to think. This very question has plagued philosophers for thousands of years. So, in the quest to develop artificial intelligence, modern scientists have turned towards these philosophers to better understand thought and the human mind. This section will study various AI systems and the philosophical doctrines that underlie their processes.

1.1 Limits of AI

Before delving into specific AI systems and their philosophical backgrounds, it is relevant to discuss the philosophical doctrines that underly artificial intelligence in general.

1.1.1 Dualism

Substance dualism, as proposed by Descartes, is the philosophical doctrine that holds that the world is composed of two substances: physical and mental. According to this doctrine, the body (physical substance) and mind (mental substance) exist separately and independently in their respective planes[13]. Descartes' dualism evokes the limits of strong AI. This theoretical form of artificial intelligence is conscious, self-aware, and would have the ability to plan for the future. Indeed, strong AI, a theoretical concept, has intelligence equal to humans[14]. Following substance dualism, however, the subjective mental states required for strong AI are immaterial and therefore impossible to replicate in machines in the physical realm. For further reading on strong AI, refer to chapter 1, section 3.1.6.

1.1.2 Materialism

In contrast to dualism, materialism suggests that the world consists exclusively of physical substances. This doctrine thereby rejects the dualistic idea of immaterial substances. Although materialists admit that certain things seem to be outside the physical realm at first glances, such

---

13 Howard Robinson, "Dualism", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2020/entries/dualism/>.

14 Stuart J Russell & Peter Norvig, *Artificial intelligence: a modern approach, 2nd ed,* New Jersey, Pearson Education Inc., 2009 at 29.

as emotions and morality, they hold that these remain physical in nature[15]. Concerning artificial intelligence, materialism opened the door to the idea that subjective mental states could be replicated in machines considering their physical nature.

1.2 Symbolic systems

Developed in the 1950s, symbolic systems are an early class of artificial intelligence that relies on symbols. Simply put, ideas, objects, and their relationships are translated into symbols which are then encoded into computer systems. These systems are then able to reason with said symbols and arrive at conclusions. The present section will study the philosophical doctrines that act as the foundation of symbolic artificial intelligence.

1.2.1 Syllogisms and deductive reasoning

First developed in 350 BC by Aristotle in his work *Prior Analytics*, a syllogism is a form of logical argument that utilizes deductive reasoning. Syllogisms are composed of three propositions: two premises and a conclusion[16]. We will use the following example to illustrate the idea:

- All men are mortal.
- Aristotle is a man.
- Aristotle is mortal.

Propositions are composed of three terms: a subject term, a predicate term, and a middle term[17]. The Subject term is the subject of the conclusion. In the example above, the subject term is "Aristotle". The predicate term is that which modifies the subject of the conclusion. In the example above, the predicate term is "mortal". The middle term is that which links the subject and the predicate terms in both premises. In the example above, the middle term is "men/man".

Further, the proposition that contains the middle term and the predicate term is called the major premise. The proposition that contains the subject term and the middle term is called the minor premise. The proposition that contains the subject term and the predicate term is called the conclusion[18].

To ease comprehension, please refer to the annotated example below:

- All **men** are <u>mortal</u> (major premise: **middle** and <u>predicate</u> terms)

---

15 Daniel Stoljar, "Physicalism", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/sum2021/entries/physicalism/>.

16 *Internet Encyclopedia of Philosophy*, Aristotle: Logic "The Syllogism" online: <https://iep.utm.edu/aris-log/#H9>.

17 *Ibid.*

18 *Ibid.*

- *Aristotle* is a **man** (minor premise: *Subject* and **middle** terms)
- *Aristotle* is mortal (conclusion: *subject* and predicate terms)

Concerning artificial intelligence, symbolic systems can utilize Aristotle's syllogism to arrive at conclusions. For the AI system to successfully perform this task, however, the terms of the syllogism must first be translated into symbols. In the above example, "M" could represent the symbol for men/man, "X" the symbol for mortal, and "A" the symbol for Aristotle. The following example illustrates how a symbolic system would view and utilize a syllogism to reach a conclusion:

- M => X
- A => M
- A => X

There is, however, a caveat. Symbolic systems are only able to utilize syllogism for concepts that lend themselves to being translated into symbols[19].

1.2.2 Means-End analysis

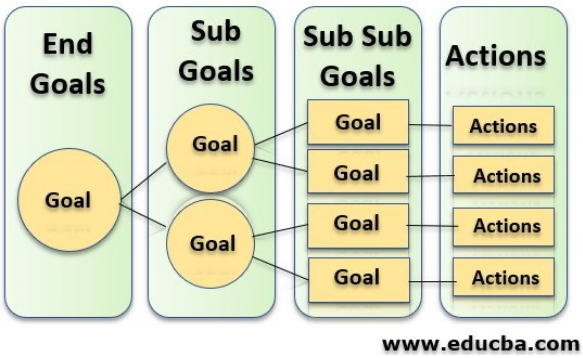
Means-end analysis is a problem-solving technique we use in our everyday lives. Following this method, a goal is divided into a series of smaller sub-goals[20]. Suppose there is no food in the pantry (current state). The absence of food in the pantry is a problem. The objective is to have a pantry filled with food (end state). This goal would then be divided into sub-goals such as getting to the grocery store, acquiring food, and returning home. These sub-goals prompt actions that need to be taken such as driving to the store, buying groceries, and driving home. Once the actions are achieved, the sub-goals will be satisfied and the current state will approach, and eventually reach, the end state[21].

---

19 Stephen F Davis & William Buskist, *21st Century Psychology: A Reference Handbook*,1st ed, SAGE Publications, (2008) at 487.

20 American Psychological Association, *APA Dictionary of Psychology*, *sub verdo* "means-end analysis", online: <https://dictionary.apa.org/means-ends-analysis>.

21 *Ibid.*



Much like us, symbolic systems utilize means-end analysis to solve problems. For instance, the General Problem Solver (GPS) utilizes means-end analysis by devising its end goal into several more manageable subgoals. For a detailed examination of the General Problem Solver, please refer to chapter 2, section 3.2.3.

For an in-depth analysis of symbolic systems, please refer to chapter 2, section 3.

1.3 Expert systems

In the 1960s and 1970s, a new form of symbolic AI was developed. As opposed to earlier symbolic systems that focused on creating general problem-solving methods, these new "expert systems" were focused on incorporating human expertise in solving specific problems[22]. We will see that expert systems "learn" in a similar fashion to human beings.

1.3.1 Empiricism

Empiricism is the philosophy that holds that knowledge can exclusively be acquired through sensory experience. Knowledge on a given subject, therefore, depends *a posteriori* on our sensory experiences. Empiricists thereby reject their philosophical antithesis, the doctrine of innateness, which suggests that human beings are born with certain innate knowledge[23].

To create an expert system, human experts, assisted by engineers, compile their knowledge into a database. The database is then translated into a language the computer can understand and uploaded to the system. After the process is complete, we are left with a machine that can solve problems based on its prior "sensory experiences" – its database containing the human expert's knowledge[24].

---

22 Bruce G Buchanan, "A (Very) Brief History of Artificial Intelligence" (2005) 26:4 *AI Magazine* 53–53 at 59.

23 Peter Markie, "Rationalism vs. Empiricism", *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism/>.

24 Bruce G Buchanan & Reid G Smith, "Fundamentals of Expert Systems" (1988) 3:1 *Annual Review of Computer Science* 23–58 at 19, 20.

1.3.2 Confirmation theory

Confirmation theory is the notion where thinkers use evidence to either confirm or reject their hypotheses[25]. Like many of the doctrines and theories we have discussed, we use confirmation theory in our everyday lives.

When a hypothesis is supported by evidence, we say that it is "confirmed"[26]. We must always keep in mind, however, that just because a hypothesis is confirmed does not make it gospel. Other hypotheses may also be compatible with the evidence on which we are leaning. For example, we can think of a doctor who mistakenly diagnosis a patient despite diligence and good faith. Evidence of a cough and fever may be evidence to confirm the theory that the patient is suffering from the flu. However, that evidence may also be compatible with a diagnosis of COVID-19.

When an expert system reaches a conclusion, they are making assumptions based on the "evidence" in their database. Moreover, broadly speaking, confirmation theory is observable in the system's database itself. Human beings rely on evidence in the pursuit of knowledge. Hence, the human expert's knowledge, which is the foundation of the database, is also shaped by confirmation theory.

1.3.3 Induction

Historically linked to confirmation theory, induction consists of ascertaining a rule from facts[27]. Another way of looking at this doctrine is that we can predict the future by looking at past and present knowledge[28]. For instance:

> The observation that mushrooms of a certain appearance have always been nourishing seems to justify that this kind of mushroom is generally nourishing and that mushrooms with a similar appearance will also be nourishing.[29]

These inferences about the unknown, based on current and passed knowledge, are called "inductive inferences"[30]. However, much like confirmation theory, the results of inductive logic

---

25 Vincenzo Crupi, "Confirmation", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/spr2021/entries/confirmation/>.

26 *Internet Encyclopedia of Philosophy*, "Confirmation and Induction" online: <https://iep.utm.edu/conf-ind/>.

27 Tristan Cazenave "Deductive Learning" In Seel N.M. (eds) *Encyclopedia of the Sciences of Learning*, Springer, Boston, MA., (2012), doi: <doi.org/10.1007/978-1-4419-1428-6_781>.

28 Internet Encyclopedia of Philosophy, "Confirmation and Induction", *supra*, note 26.

29 Leah Henderson, "The Problem of Induction", *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/spr2020/entries/induction-problem/>.

30 *Ibid.*

are not infallible. For instance, in the above example, foragers will be quick to inform you that several poisonous mushrooms closely resemble edible nutritious ones.

Expert systems utilize induction by drawing in the information contained in their database. As previously discussed, the foundation of the system's database or "intelligence" is the human expert's past and present knowledge. Hence, by utilizing the human expert's past and present knowledge, the system can draw conclusions through inductive inference. Yet, as we will see in chapter 2, keeping an expert system's database up to date is one of the great challenges surrounding these systems. Indeed, as soon as the database falls out of date, the system becomes obsolete.

1.3.4 Deduction

As we have seen in subsection 1.2.1, deduction consists of ascertaining a fact (Aristotle is mortal) from a rule (all men are mortal) and another fact (Aristotle is a man)[31]. Much like early symbolic AI, expert systems utilize deductive reasoning as they rely on facts and rules to solve problems. This is notably true for systems written in Prolog, a computer programming language devised for artificial intelligence which allows programs to be written declaratively as a set of facts and rules.

Please refer to chapter 2, section 3.3 for further reading on expert systems.

1.4 Machine learning

After discovering the limitations of expert systems, a new form of AI was born. This new method, known as machine learning, can autonomously learn from data, as opposed to the manual coding required by symbolic systems. Indeed, machine learning relies on algorithms to autonomously detect patterns and correlations in data to predict unseen outcomes. We will now study how these systems utilise empiricism and induction.

1.4.1 Empiricism

As we have seen above, empiricists suggest that knowledge is acquired exclusively from prior sensory experiences. This philosophical doctrine is not only apparent in expert systems, but in machine learning as well.

As we will discuss in chapter 2, section 4, the creation of a machine learning model requires a dataset – an immense collection of data that. Once selected (or created), the dataset will be split into two parts: a training portion and a testing portion. As the name suggests, the purpose of the

---

[31] Tristan Cazenave, *supra*, note 27.

training portion is to train the machine learning model to complete its task. Subsequently, the testing portion of the dataset will assess the model's performance in the said task[32].

Evidently, the philosophical doctrine of empiricism comes into play during the training of the machine learning model. Hence, the training portion of the dataset may be considered "prior sensory experience" which will act as the foundation of the machine learning model's reasoning.

1.4.2 Induction

We will now observe how machine learning models utilize induction to predict future outcomes based on past and present knowledge. As previously discussed, machine learning models "learn" from massive amounts of data called datasets. Typically, the more data the model has access to, the better it will perform in its task. Machine learning systems will utilize their datasets (past and current knowledge) to notice patterns and predict correlations.

Say, for instance, that the model's task is to determine the species of flower based on the size and shape of its petals. It will refer to its knowledge (dataset) of flowers and the respective size and shape of their petals to determine the species of the flower in question (inductive inference).

The present section merely studies the philosophical background of machine learning models. For an in-depth analysis of the workings of these models, please refer to section 4 of chapter 2.

1.5 Deep learning

The next step in the evolution of AI, deep learning models can learn to complete tasks without being programmed to do so[33]. Often seen as a subfield of machine learning, we will tend to deep learning separately throughout this working paper to highlight its particularities. Nevertheless, as a subfield of machine learning, the philosophical doctrines of empiricism and induction are apparent throughout the training of deep learning models as well. Hence, as the above section on machine learning holds true for deep learning, we will study two new doctrines in the present section: computationalism and connectionism.

---

32 Yufeng G, "The 7 Steps of Machine Learning", (7 September 2017), online : *Medium* <https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e>; Harini Suresh & John V Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning" (2020) arXiv:190110002 [cs, stat], online: <http://arxiv.org/abs/1901.10002> arXiv: 1901.10002; Chanin Nantasenamat, "How to Build a Machine Learning Model", (25 July 2020), online : *Medium* <https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1>.

33 Oxford Learners's dictionary, "'Machine Learning" online : <https://www.oxfordlearnersdictionaries.com /us/definition/english/machine-learning?q=machine+learning>.

1.5.1 Computationalism

Computational theory of mind, or computationalism, suggests that the human mind is a computer[34]. The foundation of computationalism is that all cognition, regardless of form, is mere computation. Indeed, according to this theory, whether made from flesh and blood or silicon chips, cognitive systems can be reduced to mathematical calculations[35].

Central to the idea of artificial intelligence is that human intelligence can be reproduced in machines. This suggests that the human brain can be understood to such an extent that we can build a machine that perfectly replicates its processes. An illustration of this idea can be found in the famous Turing Test.

The Turing Test, or imitation game, involves an interrogator, a man, and a machine. The interrogator is in one room, the man and machine in another. The interrogator's objective is to determine which of two in the other room is the man. The interrogator will therefore ask the two a series of questions. The machine will try to trick the interrogator into thinking they are human, while the man will attempt to assist the interrogator in discovering his true identity. If the machine can trick the interrogator into thinking that it is human, we will say that it is intelligent[36]. For more on Alan Turing, the Turing test, and the Turing machine, refer to chapter 1, section 2.1.1.

Although computationalism and the Turing test have been criticized in recent years, it is a nice demonstration of computationalism in the history of AI.

1.5.2 Connectionism

Rather than a philosophical doctrine, connectionism is a movement in cognitive science that studies intelligence through artificial neural networks[37]. Inspired by human biology, neural networks are composed of units and weights. Units act like artificial neurons whose role is to transmit information. As the image [38] demonstrates, there are three kinds of units. Input units that receive information, output units that present the results, and hidden units which act as an intermediary between the input and output units[39]. Collectively, these units are referred to as layers: the input units form the input layer, the hidden units the hidden layer, and the output

---

34 *Internet Encyclopedia of Philosophy*, "The Computational Theory of Mind" online: <https://iep.utm.edu/compmind/>.

35 Paul Dumouchel, "Intelligence, Artificial and Otherwise" (2019) 24:2 *Forum Phisosophicum*, at 241.

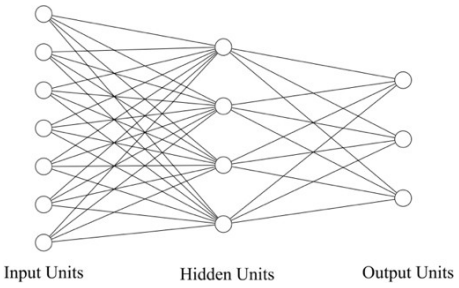
36 Graham Oppy and David Dowe, "The Turing Test", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/win2020/entries/turing-test/>.

37 Cameron Buckner and James Garson, "Connectionism", The *Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

38 *Ibid.*

39 *Ibid.*

units the output layer. These layers work together to receive, dissect, and turnout information. Weights, on the other hand, represent the level of connection between these units[40].



For example, say we have a neural network whose task is to recognize handwritten numbers 0 to 9. The units in the input layer will receive the handwritten numbers. The hidden layers will dissect these handwritten numbers into various lines and edges. Eventually, in the output layer, the neural network will predict which number, from 0 to 9, was written.

What makes deep learning different from the other forms of artificial intelligence is its level of autonomy. In comparison to other forms of artificial intelligence, deep learning systems can perform a task with significantly less human intervention[41]. Indeed, through the power of neural networks, these machines "learn" much like the human brain.

Please refer to section 4 of chapter 2 for a profound analysis of deep learning.

1.6 Conclusion

In the pursuit of developing artificial intelligence, researchers must first and foremost understand what it means to be intelligent. How can we properly replicate intelligence in machines if we do not understand it in the first place? This is why philosophy, in the context of AI, is important. It allows us to study the doctrines of philosopher's past who have dedicated their lives to answer such questions. By studying philosophy, we are afforded a better understanding of the foundation of artificial intelligence systems and the philosophical roots therein.

---

40 *Ibid.*

41 Yann LeCun, Yoshua Bengio, Geoffrey Hinton, "Deep Learning", (2015) 521 *Nature* 436-444.

Further Reading :

Buckner, Cameron, and James Garson, "Connectionism", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

Cazenave T. (2012) Deductive Learning. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_781.

Crupi, Vincenzo, "Confirmation", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/spr2021/entries/confirmation/>.

Dumouchelle, Paul, Intelligence, Artificial or Otherwise, Ritsumeikan University, Graduate School of Core Ethics and Frontier Sciences, 56-1 Kita-ku, online: <https://ajcact.openum.ca/files/sites/160/2020/01/Intelligence-artificial-and-otherwise-forumphilosophicum-1-4.pdf>.

Henderson, Leah, "The Problem of Induction", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/spr2020/entries/induction-problem/>.

Internet Encyclopedia of Philosophy, Aristotle: Logic “The Syllogism” online: <https://iep.utm.edu/aris-log/#H9>.

Markie, Peter, "Rationalism vs. Empiricism", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism/>.

Robinson, Howard, "Dualism", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/fall2020/entries/dualism/>.

Stoljar, Daniel, "Physicalism", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), online: <https://plato.stanford.edu/archives/sum2021/entries/physicalism/>.

2. Des sciences biologiques aux neurosciences

L’osmose entre l’informatique et les systèmes biologiques ne date pas d’hier. Dans notre imitation de la manière dont la nature traite, communique et organise l’information, plusieurs modèles s’offrent à nous dans la nature. Parmi les méthodes de calcul dites bio-inspirées, citons :

- les algorithmes d’optimisation[42] cherchant à reproduire une intelligence distribuée ou collective observée notamment dans la coordination des comportements complexes de groupes d’animaux, d’oiseaux et d’insectes sociaux,

42 Alberto Colorni, Marco Dorigo et Vittorio Maniezzo, « Distributed Optimization by Ant Colonies » in *Proceedings of ECAL91 – European Conference on Artificial Life*, Paris, Elsevier, 1991, 134; J Kennedy et R Eberhart, « Particle swarm optimization » (1995) 4 *Proceedings of ICNN’95 – International Conference on Neural Networks 1942*, doi : <doi.org/10.1109/ICNN.1995.488968>; Gai-Ge Wang, Suash Deb et Leandro dos S Coelho, « Elephant Herding Optimization » in IEEE, *3rd International Symposium on Computational and Business Intelligence*, 2015, 1, doi : <doi.org/10.1109/ISCBI.2015.8>.

- les algorithmes de routage inspirés du réseau mycorhizien[43]; ou encore
- les algorithmes évolutionnaires calqués sur le processus (itératif) de la sélection naturelle[44].

Le plus complexe, sophistiqué et polyvalent d'entre les différents modèles biologiques est sans doute le fonctionnement de nos systèmes nerveux (central et périphérique).

Suivant la théorie computationnelle de l'esprit [renvoi à la sous-section 1.4.1 « computationnalisme »][45], « la pensée est au cerveau ce que le logiciel informatique (*software*) est à la machine (*hardware*) »[46]. La pensée étant assimilable à un processus de calcul, il doit être possible de modéliser la pensée en reproduisant, même à base de silicone, le fonctionnement biologique sous-jacent à notre cognition. Jusqu'au début des années 1990, l'étude de nos systèmes nerveux s'était heurtée à la « boîte noire » du cerveau. À défaut de pouvoir mesurer directement l'activité du cerveau, la rétro-ingénierie de la manière dont notre cerveau fonctionne à partir de l'analyse appliquée du comportement avait ses limites. Avec l'avènement de l'imagerie par résonance magnétique (fonctionnelle) offrant la possibilité d'observer en direct l'activité des différentes zones cérébrales, la recherche sur le cerveau a été propulsée à l'avant-plan; il en résulte une compréhension plus approfondie tant de sa structure anatomique, de son fonctionnement, que du processus de traitement de l'information. Ces avancées ont donné de l'élan à l'approche connexionniste de l'esprit [renvoi à la sous-section 1.4.2 « connexionnisme »] qui postule les états mentaux comme des phénomènes émergents de la structure interconnectée et d'une disposition systémique des réseaux d'unités neuronales[47],

Notre système nerveux central, et plus particulièrement notre cortex préfrontal, excelle dans plusieurs tâches d'apprentissage et d'adaptation comportementale[48]. Si l'apprentissage « est

---

43 Voir notamment Ronaldo da Costa Bento, C. et E.C. Gomes Wille, « Bio-inspired routing algorithm for MANETs based on fungi networks » (2020) 107 *Ad Hoc Networks* 102248, doi : <doi.org/10.1016/j.adhoc.2020.102248>; Xu Hao et al, « FUNNet – A Novel Biologically-Inspired Routing Algorithm Based on Fungi » dans IEEE, *Second International Conference on Communication Theory, Reliability, and Quality of Service*, 2009, 97, doi : <doi.org/10.1109/CTRQ.2009.23>.

44 Marc Schoenauer, « Les algorithmes évolutionnaires » dans Thomas Heams, dir, *Les mondes darwiniens. L'évolution de l'évolution*, vol 2, Paris, Éditions matériologiques, 2011, 907, doi : <doi.org/10.3917/edmat.heams.2011.02.0907>.

45 Hilary Putnam, « The nature of mental states » dans WH Capitan et DD Merrill, dir, *Art, Mind, and Religion*, Pittsburgh University Press, 1967, 51.

46 Achille Weinberg, « Le modèle symbolique de l'esprit » dans Jean-François Dortier, dir., *Le cerveau et la pensée. Le nouvel âge des sciences cognitives*, Auxerre, Sciences Humaines, 2014, 35, doi : <doi.org/10.3917/sh.dorti.2014.01.0035>.

47 Stephen J Flusberg et James L McClelland, « Connectionism and the Emergence of Mind » dans Susan EF Chipman, *The Oxford Handbook of Cognitive Science*, Oxford University Press, 2014, doi : <doi.org/10.1093/oxfordhb/9780199842193.013.5>.

48 Emmanuelle Volle et Richard Levy, « Rôle du cortex préfrontal dans l'adaptation comportementale chez l'homme » (2014) 30:2 *Med Sci* 179, doi : <doi.org/10.1051/medsci/20143002016>.

parmi les compétences les plus adaptatives des espèces [vivantes] »[49], le rythme, la qualité et la modulation de cet apprentissage dépendent tant de la quantité que de l'intégration de l'information provenant de différentes sources, qu'il s'agisse de l'environnement (stimuli sensoriels), de son interaction avec l'environnement (rétroaction), de l'expérience passée (mémoire) ou encore des représentations de l'avenir (prédictions). Tant les structures multicouches de notre cerveau que la manière dont les différentes zones participent à un traitement intégré de l'information peuvent inspirer la conception de systèmes intelligents reproduisant des capacités cognitives qui se veulent similaires à celles de notre cerveau.

Dans ce processus de traitement, de (ré)organisation et de l'intégration de l'information, notre système nerveux ne fait pas que manipuler l'information, mais s'adapte également aux tâches à exécuter. Cette plasticité neuronale ou propriété d'adaptation du système nerveux lui-même aux exigences de l'environnement (plasticité neuronale), s'avère être un référent de premier plan pour la conception de systèmes d'apprentissage autonome, lesquels se démarquent par leur capacité à apprendre d'eux-mêmes à l'expérience, plutôt que de suivre mécaniquement les instructions du programmeur.

Du calcul neuromorphique[50] à la modélisation cognitive, le développement des réseaux de neurones artificiels s'inspire principalement du paradigme neurologique. Comme nous le verrons plus amplement au chapitre 2, ils sont calqués tant sur la structure des réseaux de neurones biologiques que sur leur connectivité souple (plasticité synaptique) s'ajustant au fur et à mesure des tâches à exécuter. Dans le même ordre d'idées, un réseau de neurones artificiels à la base des techniques d'apprentissage dit profond (renvoi à la section d'Hannes), repose sur une optimisation convergente des prédictions à travers plusieurs couches de neurones dont chacune donne une approximation incrémentielle de relations non linéaires. Cet apprentissage profond du monde complexe a jusque-là produit des résultats encourageants, voire surprenants dans plusieurs domaines d'application nécessitant un traitement judicieux de quantités importantes de données, dont la reconnaissance d'images, le traitement automatisé du langage, la conduite autonome ainsi que la détection des comportements et schémas (typiques) de fraudes.

---

49 Marc Philippe Lafontaine et Sarah Lippé, « Le cortex préfrontal et le processus d'apprentissage : caractérisation d'un rôle critique » (2011) 3:4 *Revue de neuropsychologie* 267, doi : <doi.org/10.3917/rne.034.0267>.

50 Carver Mead, « How we created neuromorphic engineering » (2020) 3 *Nature Electronics* 434, doi : <doi.org/10.1038/s41928-020-0448-2>.

Pour aller plus loin :

Darwish, A., « Bio-inspired computing : Algorithms review, deep analysis, and the scope of applications » (2018) 3:2 *Future Computing & Informatics Journal* 231, doi : <doi.org/10.1016/j.fcij.2018.06.001>

Fan, J. et al., « From Brain Science to Artificial Intelligence » (2020) 6:3 *Engineering* 248, doi : <doi.org/10.1016/j.eng.2019.11.012>

Hole, K.J. et S. Ahmad, « Biologically Driven Artificial Intelligence » (2019) 52 *Computer* 72, doi : <doi.org/10.1109/MC.2019.2917455>

Le, J., « Convolutional Neural Networks : The Biologically-Inspired Model », *Towards data science* (30 August 2018), online : <towardsdatascience.com/convolutional-neural-networks-the-biologically-inspired-model-f2d23a301f71>

Narcross, F., « Artificial nervous systems – A new paradigm for artificial intelligence » (2021) 2:6 *Patterns* 100265, doi : <doi.org/10.1016/j.patter.2021.100265>

Peer, P., C.M. Travieso-González, V.K. Asari et M.K. Dutta, « IEEE Access Special Section Editorial : Trends and Advances in Bio-Inspired Image-Based Deep Learning Methodologies and Applications » in *IEEE Access*, vol 9, 2021, 86657, doi : <doi.org/10.1109/ACCESS.2021.3088621>

3. Linguistique

Sur le plan de l'évolution, la faculté du langage est un des éléments constitutifs de nos capacités cognitives et marque, selon certains, l'avènement de notre toute première ère de l'information[51]. Un des mystères de l'évolution[52], l'utilisation du langage articulé – au-delà de sons ou de gestuels répétitifs, automatiques ou purement en réaction aux stimuli émotionnels – aurait permis à l'humain de construire ses pensées, de structurer sa connaissance et de la communiquer à d'autres, à telles enseignes que philosophes et linguistes n'hésitent pas à postuler l'indissociabilité du lien entre la pensée et le langage[53]. Dans une perspective évolutionniste, le langage aurait joué un rôle essentiel en tant qu'un outil de survie qui a facilité et accéléré l'adaptation de l'espèce à son environnement par le biais d'un médium culturel plus ciblé que le caractère aléatoire des mutations génétiques[54].

Alors que l'invention de l'écriture (langage écrit) avait marqué une étape charnière dans le développement des civilisations en facilitant la diffusion à grande échelle et intergénérationnelle

---

51 Daniel L Everett, *Language: The Cultural Tool*, Knopf Doubleday Publishing Group, 2012.

52 Pour une revue de la littérature multidisciplinaire sur l'origine du langage, voir Marc D Hauser et al, « The Mystery of Language Evolution » (2014) 5 Front Psychol 401, doi : <doi.org/10.3389/fpsyg.2014.00401>.

53 Jerry A Fodor, *The Language of Thought*, Harvard University Press, 1975, Pour une revue de littérature sur ce sujet, voir Yanik Simard, *L'indissociabilité de la pensée et du langage*, mémoire de maîtrise, Faculté de philosophie, Université Laval, 1997, en ligne : <www.collectionscanada.gc.ca/obj/s4/f2/dsk3/ftp05/mq25736.pdf>.

54 Mark Pagel, « Q&A : What is human language, when did it evolve and why should we care? » (2017) 15 BMC Biology 64, doi : <doi.org/10.1186/s12915-017-0405-3>; Daniel L Everett, *supra* note 51.

des connaissances acquises[55], le passage de l'écriture manuscrite, entièrement dépendante de l'humain, à l'écriture automatisée du processeur informatique a propulsé de façon exponentielle les possibilités de réalisations technologiques se répercutant sur tous les plans de la société et de notre culture.

En effet, au-delà du langage vocal ou de ce « phonocentrisme » de la linguistique traditionnelle, le langage, au sens large, est avant tout un système – structuré – de communication et de partage de connaissances[56], qu'il s'agisse de symboles graphiques, d'un code binaire, du non-verbal ou d'un code gestuel (langue des signes). Cette diversité des langues, entendue au sens large, partage certaines caractéristiques communes qui transcendent leur origine biologique. Ce sont autant de systèmes consistant notamment en la manipulation de symboles établis par convention et qui se démarquent par une cohérence structurelle (p.ex. règles de syntaxe et de grammaire) les rendant applicables dans une multitude de contextes.

Dans le même ordre d'idées, la linguistique informatique permet d'articuler la connaissance computationnelle et surtout, de faire le pont entre la cognition artificielle et l'intelligence – naturelle – qui est la nôtre. L'impératif est réciproque : il s'avère tout aussi important pour la machine de comprendre l'humain que l'humain, la manière dont la machine fonctionne.

Du langage humain au langage informatique, la prise en compte du contexte situationnel de la communication complique la tâche. La manipulation seule de symboles, même de façon structurée et cohérente, ne donne pas tout son sens à l'acte communicationnel. La signification complète d'un « système » de langage dépend aussi, à des degrés variables, d'éléments « hors système » que sont les différents contextes (p.ex. culturels, circonstanciels, interactionnels, sociaux, perceptifs-corporels) dans lesquels s'inscrit un acte communicationnel[57]. En remettant en question la suffisance du test de Turing pour démontrer la présence d'un esprit dans la machine, le philosophe John Searle avait cherché à démontrer, par son expérience de pensée de la chambre chinoise[58], que la simple manipulation de symboles dans le respect des règles programmées n'équivaut pas à une réelle compréhension d'un langage donné. Or, programmer une manipulation « sans faute » de symboles qui tiendrait compte de tous les éléments « hors système » que constitue le contexte général dans lequel des échanges ont lieu, n'est pas simple. La prise en compte de ce contexte général appellerait une synthèse « connexionniste » [renvoi à la sous-section 1.4.2 « connexionnisme »] des relations que l'interlocuteur tisse avec son

---

55 Voir par exemple Amalia E Gnanadesikan, *The Writing Revolution : Cuneiform to the Internet*, Wiley-Blackwell, 2008.

56 Ferdinand de Saussure, *Cours de linguistique générale*, Lausanne, Payot, 1916; Noam Chomsky, *Syntactic Structures*, La Haye, Mouton, 1957; Noam Chomsky, *Current Issues in Linguistic Theory*, La Haye, Mouton, 1964.

57 Voir notamment Charles Goodwin et Alessandro Duranti, « Rethinking Context : An Introduction », dans Alessandro Duranti et Charles Goodwin, *Rethinking Context : Language as an interactive Phenomenon*, Cambridge University Press, 1992, 1; MAK Halliday, « The Context of Linguistics (1975) » dans *On Language and Linguistics*, 2014, 74, doi : <doi.org/10.5040/9781474211932>.

58 John R Searle, « Minds, brains, and programs » (1980) 3:3 *Behavioral & Brain Sciences* 417.

environnement, depuis ses perceptions sensorielles à son expérience tant corporelle, sociale, relationnelle que linguistique[59]. En effet, ce non-dit, l'implicite ou ambiguïté situationnelle inhérente aux langages humains, s'avère être la principale difficulté[60] que l'on rencontre à ce jour pour automatiser la traduction, l'analyse du langage et le traitement du langage naturel, non seulement pour ces tâches en elles-mêmes, mais aussi au regard de leurs applications notamment dans le domaine de l'intelligence artificielle (IA) explicable [renvoi à la section 2.4]. Cela étant, c'est en apprenant à la machine à comprendre, à interpréter et à « parler », pour ainsi dire, le langage humain qu'il sera possible de faire le pont entre nos acquis et les connaissances acquises par la machine, de rendre celles-ci compréhensibles à l'humain et par ailleurs, de nous faire bénéficier de ce nouvel éclairage qu'apporte la machine à notre linguistique[61].

Pour aller plus loin :

Church, K. et M. Liberman, « The Future of Computational Linguistics : On Beyond Alchemy » (2021) Front Artif Intell, doi : <doi.org/10.3389/frai.2021.625341>

McShane, M. et S. Nirenburg, *Linguistics for the Age of AI*, MIT Press, 2021

Pace-Sigge, M., Spreading Activation, Lexical Priming and the Semantic Web, Palgrave Pivot, 2018

Wilson, H.J. et P.R. Daugherty, « The Next Big Breakthrough in AI Will Be Around Language », *Harvard Business Review* (23 septembre 2020), online : <hbr.org/2020/09/the-next-big-breakthrough-in-ai-will-be-around-language>

4. Mathématique

4.1 Du calcul formel aux modélisations mathématiques

Au-delà de l'interprétation du langage humain, tout système informatique fonctionne à la base à l'aide des règles du calcul formel ou symbolique, enregistrant et traitant l'information en code binaire. Si le langage mathématique partage quelques propriétés avec le langage humain – dont la manipulation de symboles, l'existence d'une syntaxe (algorithme [renvoi à la sous-section 3.1.3 du chapitre 1 « Intelligence artificielle et algorithme »]) ainsi que d'une structure, les mathématiques excellent dans la formalisation de structures, régularités et relations qui sont, pour ainsi dire, « abstraites » de la complexité du réel ou du chaos apparent, à l'aide de

59 Nick C Ellis, « Emergentism, Connectionism and Language Learning » (1998) 48:4 *Language Learning* 631.

60 W. Knight, « AI's Language Problem », *MIT Technology Review* (9 août 2016), en ligne : <www.technologyreview.com/2016/08/09/158125/ais-language-problem/>.

61 James W Goodwin et Uwe Hein, « Artificial intelligence and the study of language » (1982) 6:3-4 J *Pragmatics* 241, doi : <doi.org/10.1016/0378-2166(82)90003-0>.

représentations simplifiées, schématiques qui auraient l'avantage d'être facilement manipulables, itératives et généralisables[62].

Que l'on épouse le point de vue platonicien – selon lequel les mathématiques seraient le code source de l'univers[63] – ou une perspective constructiviste – considérant les mathématiques comme un produit de l'esprit humain[64], la modélisation mathématique s'avère être d'une utilité indéniable au confluent des disciplines scientifiques et de la technologie[65]. Dans la simulation d'un cerveau artificiel, elle aide à reconstituer les divers états mentaux qu'il est possible de traduire en représentations, dont certaines de nos fonctions cognitives comme le raisonnement, la mémoire de travail et la planification.

4.2 Probabilités et statistique : Optimisation et prévisions

Appliquées au domaine de l'intelligence artificielle (IA), la théorie des probabilités et la statistique trouvent un terreau fertile dans l'étude de phénomènes complexes qui se démarquent *a priori* par un haut degré d'incertitude. Quelle que soit l'approche probabiliste retenue (fréquentiste ou bayésienne), la modélisation statistique vise ultimement à approximer, dans la mesure du possible, la réalité non seulement dans ce qu'elle a été, mais surtout dans ce qu'elle sera très probablement. À l'instar de toute théorie scientifique, une modélisation statistique n'est vraiment valide que dans la mesure où elle permet d'anticiper l'avenir, de faire des prédictions les plus justes possible en minimisant les taux d'erreur qui résultent tant de sur-diagnostics (faux positifs) que des cas sous-diagnostiqués (faux négatifs).

Des indicateurs aux prévisions, il en faut de peu pour que les algorithmes soient mobilisés en tant qu'aides à la prise de décision dans une multitude de domaines d'application, depuis le marketing prédictif à l'analyse de grandes tendances économiques et financières, voire géo-politiques. Des algorithmes intelligents viennent en améliorer la performance en modélisant efficacement des situations complexes, dites non linéaires, qui impliquent des quantités importantes de données et une diversité des (hyper)paramètres modulant leur interaction.

---

62 Michael C Mitchelmore et Paul White, « Abstraction in Mathematics Learning » dans NM Seel, dir, *Encyclopedia of the Sciences of Learning*, Boston, Springer, 2012, doi : <doi.org/10.1007/978-1-4419-1428-6_516>.

63 Parmi les tenants du platonisme mathématique en tant que découverte (plutôt qu'une invention de l'esprit humain), citons notamment Roger Penrose, *The Emperor's New Mind : Concerning Computers, Minds, and the Laws of Physics*, New York, Oxford University Press, 1989; John D Barrow, *PI in the Sky : Counting, Thinking, and Being*, Back Bay, 1993.

64 Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics*, Oxford, Basil Blackwell, 1956; Philip Kitcher, *The Nature of Mathematical Knowledge*, Oxford University Press, 1983, doi : <doi.org/10.1093/0195035410.001.0001>; Paul Ernest, *Social Constructivism as a Philosophy of Mathematics*, SUNY Press, 1997.

65 Sven Ove Hansson, « Technology and Mathematics » (2020) 33 Philosophy & Technology 117, doi : <doi.org/10.1007/s13347-019-00348-9>.

4.3 Théorie des jeux : Optimisation et compétition

Dans le but de coordonner les prises de décisions entre plusieurs agents (rationnels), la modélisation mathématique est également susceptible d'intervenir pour aider les algorithmes (intelligents) à traiter des situations avec information incomplète et en présence de compétition. La théorie des jeux, formalisée au courant des années 1940[66], offre un cadre d'analyse intéressant pour aider la machine à prendre des décisions en présence d'asymétries informationnelles où l'optimisation du résultat global (*cf.* dilemme du prisonnier) nécessite non seulement de prendre en compte des données incomplètes sur une base unilatérale, mais aussi d'anticiper la manière dont d'autres agents – aux intérêts divergents mais tout aussi rationnels et intelligents – décideront à la lumière des mêmes données incomplètes. Les stratégies d'optimisation inspirées de la théorie des jeux permettent aux algorithmes intelligents de prédire certains comportements humains dans des domaines ludiques (p.ex. jeux de société ou de stratégie[67]) ou encore de se renforcer mutuellement leur apprentissage au moyen notamment de réseaux antagonistes ou adverses génératifs (renvoi à la sous-section 5.3.3 « Generative Adversarial Networks » du chapitre 2)[68].

Pour aller plus loin :

Ellacott, S.W., J.C. Mason et I.J. Anderson, *Mathematics of Neural Networks*, Springer, 1997

Nowé A., P. Vrancs et Y.M. De Hauwere, « Game Theory and Multi-agent Reinforcement Learning » in M. Wiering et M. van Otterlo, dir., *Reinforcement Learning. Adaptation, Learning, and Optimization*, vol 12, Berlin / Heidelberg, Springer, 2012, 441, doi :<doi.org/10.1007/978-3-642-27645-3_14>

Widdows, D., K. Kitto et T. Cohen, « Quantum Mathematics in Artificial Intelligence » (2021), online : <arxiv.org/abs/2101.04255>

Yuan, Y. et al., « Deep learning from a statistical perspective » (2020) 9:1 Stat e294, doi : <doi.org/10.1002/sta4.294>

Zadeh, L.A., « Is Probability Theory Sufficient for Dealing with Uncertainty in AI : A Negative View » in L.N. Kanal et J. F. Lemmer, dir., *Machine Intelligence and Pattern Recognition*, vol 4, 1986, 103, doi : <doi.org/10.1016/B978-0-444-70058-2.50012-7>

5. Computer engineering

Computer engineering is found at the intersection of electrical engineering and computer science. The job of a computer engineer is to research and develop computer hardware or software. Hardware involves the physical components of the computer such as the monitor, microprocessors, memory chips, etc. Software, on the other hand, refers to intangibles such as

66 Oskar Morgenstern et John von Neumann, *Theory of Games and Economic Behavior*, PUP, 1944.

67 Voir notamment Gabe Stechschulte, "Game Theory Concepts Within AlphaGo", *Towards Data Science* (9 May 2020), online : <towardsdatascience.com/game-theory-concepts-within-alphago-2443bbca36e0>.

68 Voir notamment Barbara Franci and Sergio Grammatico, "A Game-Theoretic Approach for Generative Adversarial Networks" (2020), online : <arxiv.org/abs/2003.13637>.

the operating system and applications amongst other things[69]. This section will study the history of the computer and how artificial intelligence owes its existence to this field.

The history of modern computers begins in the 1940s. Alan Turing, an English mathematician created a computational device in 1940 whose sole purpose was to decipher Nazi messages during the second world war[70]. During this same period, other researchers were making breakthroughs as well. In 1941, Konrad Zuse, a German engineer created the first programable computer dubbed the Z3[71]. Around the same time at Iowa University, Professor John Atanasoff and his graduate student Clifford Berry created the first electronic computer, the ABC[72]. Later in 1943-1944, John Mauchly and J. Presper Eckert from the University of Pennsylvania built the ENIAC as part of a secret military project[73]. The ENIAC was the first general-purpose, electronic, digital computer[74] that filled a 20-foot by 40-foot room[75]. In 1947, William Shockley, John Bardeen, and Walter Brattain invented the transistor[76], a foundational device of modern electronics which is used to either amplify or switch electronic signals and power[77].

The second half of the 20th century saw a shift. Computers, who until then were largely reserved to governments and businesses, were now making their way to the general public. Jack Kilby and Robert Royce independently invented the integrated circuit in 1958 and 1959, respectively[78]. The advent of the integrated circuit, also known as the computer chip, paved the way for the invention of modern, more portable, electronic devices such as the mobile phone and the laptop computer[79]. In 1964, Douglas Engelbart, an American engineer, unveiled a prototype of the computer as we know it with a graphical user interface and fitted with a mouse[80]. In the early 1970s, further leaps were made with the invention of the first dynamic access memory chip in

---

69 Jim Lucas, "What is Computer Engineering ?", *Live Science,* (17 October 2014), online: <https://www.livescience.com/48326-computer-engineering.html>.

70 Stuart J Russell & Peter Norvig, *supra* note 14*,* at 14.

71 *Ibid.*

72 *Ibid*.

73 *Ibid*.

74 *Ibid*.

75 Timothy Williamson, "History of Computers: A Brief Timeline", *Live Science,* (1 December 2021), online: <https://www.livescience.com/20718-computer-history.html>.

76 *Ibid.*

77 Chris Woodford, "Transitors", *Explain that Stuff!,* (8 December 2021), online: <https://www.explainthatstuff.com/howtransistorswork.html>.

78 PBS, "Integrated Circuits", (1999), online: <https://www.pbs.org/transistor/background1/events/icinv.html>.

79 Anysilicon, "The History of the Integrated Circuit", online: <https://anysilicon.com/history-integrated-circuit/>.

80 Timothy Williamson, *supra* note 75*.*

1970 and the floppy disk in 1971[81]. By the mid-1970s, personal computers advertised to consumers hit the market[82]. This increase in accessibility to computers stimulated further study in the field and has led to constant innovation.

Artificial intelligence systems are nothing more than computer programs. They, therefore, owe their existence to the software side of computer engineering. Indeed, computer engineering provides the programming languages required to create artificial intelligence systems[83]. Further, as the field of computer engineering progresses and computers become more powerful, so do the capabilities of artificial intelligence.

Further Reading :

Artificial Intelligence, a modern approach
https://www.livescience.com/20718-computer-history.html

6. AI in the media

Since the birth of AI in the 1940s, members of the field were quick to sing its praises. Plagued by over-enthusiasm, promises were made that could not be kept. We were led to believe that sapient machines were around the corner. Yet here we are. Eighty years after the birth of the field, and there is a notable absence of droids among us. This section will study the false perception of artificial intelligence in the media and how AI's founding fathers could have had a role to play in this fallacy.

6.1 The myth of progress

As we will discuss in chapter 1, section 2.1.2, pioneers overestimated the early potential of AI. In 1957, Herbert Simon, a multidisciplinary American researcher, announced that there were currently machines that could think, learn, and create. Moreover, Simon stated that soon these machines would be able to solve problems to the same degree as the human brain[84].

In 1958, a New York Times article reported that the perceptron, created by Frank Rosenblatt the previous year, was the first piece of technology to think like the human brain. It further conveyed the possibility of a perceptron that possessed consciousness. Moreover, the article noted that the perceptron was so advanced that the United States government struggled to call it a "machine"[85].

---

81 *Ibid.*

82 *Ibid.*

83 Stuart J Russell & Peter Norvig, *supra* note 14, at 14.

84 Stuart J Russell & Peter Norvig, *supra* note 14, at 20.

85 « Electronic "Brain" Teaches Itself », *New York Times* 116 (1958), online: <http://timesmachine.nytimes.com/timesmachine/1958/07/13/91396361.html>.

The above statements were not only premature but possibly out of touch. Hence, with such promises floating around the media during the early days of AI, it is no surprise that the public's perception of artificial intelligence quickly surpassed its technological reality.

6.2 Literature and film

Servants, all-knowing entities, overlords, killing machines – AI has been depicted in countless ways on the page and silver screen. Interestingly, these fictitious intelligent machines predate the birth of actual artificial intelligence by hundreds of years. Perhaps the first appearance of AI is the Golem, a man-made entity in the Jewish tradition. Although the term Golem is present in the bible, its reference to the creation of artificial intelligence dates to the year 500 in the Babylonian Talmud[86]. A more recent narrative tells the story of a Golem created by a rabbi to protect the powerless members of his community. Tragically, this protective figure becomes too powerful and difficult to control thereby recking havoc[87]. Since, similar stories have been told in Marry Shelly's Frankenstein, The Terminator, The Matrix, I Robot, and innumerable other titles which have shaped the cultural perception of AI. For some, as a result, the term "artificial intelligence" is evocative of a strong emotion: fear.

6.3 God

The discomfort some feel when confronted with artificial intelligence may be traced back to Judaeo-Christian ideas. As human beings create autonomous machines capable of thought, we approach the status of the creator. We, therefore, fear that intelligent machines will escape our control much like we escaped God's[88]. Like many things in life, fear and judgment stem from misunderstanding.

6.4 Conclusion

While Arnold Schwarzenegger can certainly attest to the lucrative nature of depicting AI as murderous, autonomous machines, it is not doing our field of study any favors. The media has projected a false image of AI. As a result, the public is misinformed as to its true nature. Truth is, AI lives within the parameters of its programming. Unlike genuine intelligence, artificial

---

86 Artificial Intelligence Avant La Lettre: The Golem of Jewish Mysticism, Legend and Art' by Emily D. Bilski, Barbican Center, online: <https://artsandculture.google.com/exhibit/meet-the-golem-the-first-artificial-intelligence-barbican-centre/0QLiTNxULrWYKg?hl=en>.

87 *Ibid.*

88 France, Sénat (par MM. André Gattolin, Claude Kern, Cyril Pellevat et Pierre Ouzoulias), Intelligence artificielle : l'urgence d'une ambition européenne, rapport d'information fait au nom de la Commission des affaires européennes sur la stratégie européenne pour l'intelligence artificielle, session ordinaire de 2018-2019, n° 279, 31 janvier 2019, en ligne : <www.senat.fr/rap/r18-279/r18-2791.pdf>.

intelligence is not preoccupied with morality or the consequences of its actions[89]. It is a tool, nothing more, nothing less.

Further Reading :

Dumouchelle, Paul, *Intelligence, Artificial or Otherwise*, Ritsumeikan University, Graduate School of Core Ethics and Frontier Sciences, 56-1 Kita-ku, online: <https://ajcact.openum.ca/files/sites/160/2020/01/Intelligence-artificial-and-otherwise-forumphilosophicum-1-4.pdf>.

Stuart J Russell & Peter Norvig, *Artificial intelligence: a modern approach, 2e éd.,* New Jersey, Pearson Education Inc., 2009 at 29.

Transhumanism on artificial intelligence portrayed in selected science fiction movies and TV series, online: <http://www.medicjournalcampus.it/fileadmin/MEDICS/archivio/vol_1-2_2017/numero_1_giu_2017/10_Ezpeleta_Segarra.pdf)>.

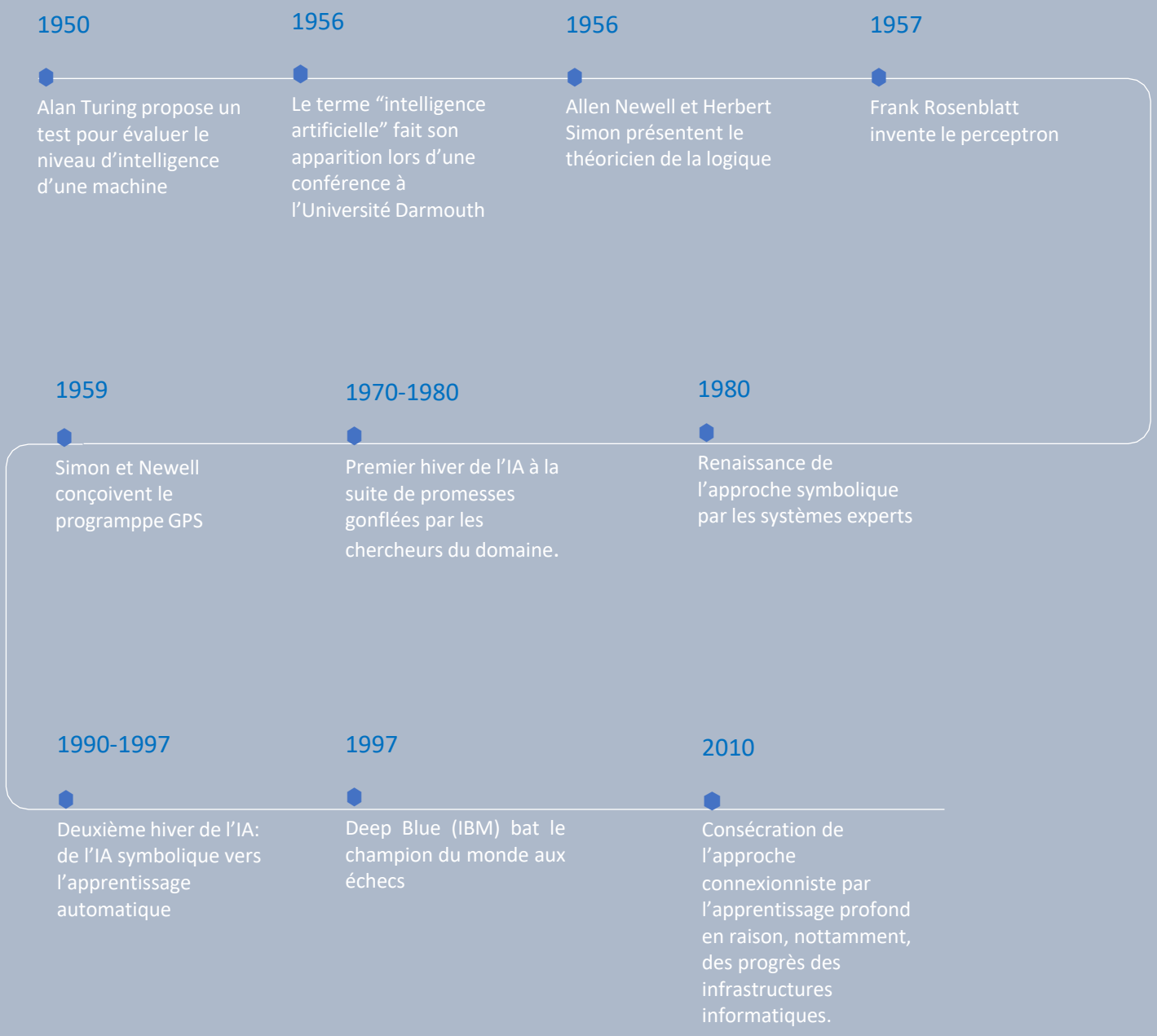
89 Paul Dumouchel*, supra* note 35*,* at 253, 254.

Conclusion

AI thus is making great strides through the emerging field of cognitive sciences. The term encompasses many subfields of studies gravitating around the studies of information processing, knowledge sharing and decision-making process, including philosophy of mind, neurosciences, linguistics, maths, and computer engineering. Parallel to its scientific sources, AI also fuelled popular expectations and fear as well as media enthusiasm on the advent of “superhuman” creatures, “transhuman” future and “godless” world. From this broad survey of AI’s multidisciplinary origins, we will delve further into the history of AI (section 2) and the deceptive polysemy of "artificial intelligence” (section 3).

Section 2 - Les développements en intelligence artificielle : une histoire marquée par l'opposition

Aperçu des moments marquants dans le développement de l'intelligence artificielle.



Dans cette section sur l'histoire de l'intelligence artificielle (IA), il sera question :

- de présenter en quoi l'opposition entre le courant symbolique et connexionniste a servi à façonner et structurer le domaine de la recherche en IA;
- de discuter des incidences des promesses gonflées par les chercheurs sur le développement de la discipline;
- de présenter les critiques soulevées pendant les différents hivers de l'IA.

1. Une histoire marquée par l'opposition

L'histoire de l'intelligence artificielle (IA) est marquée par une opposition entre le courant symbolique et connexionniste. Si l'apprentissage profond est aujourd'hui considéré comme l'approche la plus prometteuse dans le domaine de l'IA, le courant connexionniste, qui en est à la base, a longtemps été mis à l'écart. En effet, le terme « intelligence artificielle » a été cité pour la première fois dans le but de lancer un projet de recherche dans le domaine de l'IA symbolique et de se dissocier de la première cybernétique[90] à l'origine de l'approche connexionniste[91]. Le courant symbolique est donc le cadre de référence initial en matière d'intelligence artificielle[92]. Quant au courant connexionniste, longtemps marginalisé, il a évolué parallèlement à celui du domaine de l'intelligence artificielle pour ensuite en rejoindre ses rangs.

L'origine de cette opposition s'explique principalement par une différence dans la représentation de l'intelligence. L'approche symbolique repose sur l'idée qu'il est possible de simuler le raisonnement humain en récréant notre capacité à se représenter des concepts abstraits (symboles)[93]. Notamment, des objets, des plans, des décisions, des émotions, etc. En utilisant un certain nombre de règles logiques, il est possible de construire des systèmes capables d'agir sur ces symboles, par exemple pour atteindre certains objectifs[94]. À l'opposé, l'approche connexionniste emprunte des raisonnements propres à la neuropsychologie. Pour les connexionnistes, le raisonnement n'est pas suffisant pour rendre une machine intelligente, car trop restrictif, il écarte nécessairement des dimensions importantes telles que la notion d'apprentissage par expérience, perception et intuition. C'est en entraînant la machine à

---

90 Voir *infra*, Sous-section 1.1 « Les balbutiements de l'intelligence artificielle » de la Section 2 « Les développements en intelligence artificielle : une histoire marquée par l'opposition » du Chapitre 1.

91 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, « La revanche des neurones: l'invention des machines inductives et la controverse de l'intelligence artificielle » (2018) 5:211 à la p 6; John McCarthy et al, *A Proposal for the Darmouth Summer Research Project On Artificial Intelligence*, Proposition de recherche, Darmouth, 1995.

92 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 6.

93 John Haugeland, *Artificial Intelligence: The Very Idea*, Cambridge: MIT Press, 1985 aux pp 112, 113.

94 Voir *infra*, Chapitre 2 « Symbolic systems », « Fundamental concepts ».

apprendre par elle-même grâce à des réseaux de neurones artificiels, inspirés du fonctionnement du cerveau, qu'elle sera en mesure d'agir de manière intelligente et plus nuancée[95].

Cette opposition, qui servira de fil conducteur pour retracer l'histoire de l'intelligence artificielle, servira à souligner trois éléments au cœur de la discipline :

1. l'intelligence artificielle, comme on la connait aujourd'hui, repose sur des idées issues de ces deux courants qui sont à la base de la discipline depuis plusieurs décennies;

2. peu importe l'approche mise de l'avant, la compréhension et l'étude de l'esprit humain sont centrales;

3. le travail en parallèle et la tension entre ces deux courants a participé à faire en sorte que les progrès projetés en matière d'IA ont souvent été exagérés[96].

1.1 Les balbutiements de l'intelligence artificielle

L'opportunité d'appliquer une forme d'intelligence biologique à des machines se concrétise par les développements technologiques des années 40 dont la Deuxième Guerre mondiale a été un accélérateur[97]. En 1943, Warren McCulloch et Walter Pitts présentent ce qui est considéré aujourd'hui comme un des premiers modèles d'intelligence artificielle à l'origine du courant connexionniste : le neurone formel. Inspiré par le fonctionnement des neurones dans le cerveau, le neurone formel est un modèle mathématique et informatique du neurone biologique[98]. Puis, en 1950, Alan Turing, mathématicien et créateur de l'ordinateur moderne, lance sans la nommer l'idée d'une possible intelligence « artificielle ». Dans son célèbre article *Computing Machinery and Intelligence*, Turing propose le « jeu de l'imitation »[99]. Selon lui, si une machine est capable de simuler l'intelligence humaine de manière à laisser croire à une personne qu'elle interagit avec un humain, cette machine peut être considérée comme intelligente[100]. Pour Turing, si une machine est capable de tromper un humain, c'est qu'elle a des compétences cognitives assez élaborées pour être capable de raisonner et se représenter des connaissances qui lui permettent de prendre de bonnes décisions dans une grande variété de situations[101]. Contesté aujourd'hui, ce « jeu de l'imitation » a longtemps été considéré comme la principale référence pour

---

95 Yann LeCun, *Quand la machine apprend*, Odile Jacob éd, Paris, 2019, « Connexionniste et Perceptron ».

96 Cet élément sera discuté plus amplement au troisième chapitre du document de travail.

97 « Histoire de l'intelligence artificielle », en ligne : Conseil de l'Europe <https://www.coe.int/fr/web/artificial-intelligence/history-of-ai>.

98 Stuart J Russell & Peter Norvig, *Artificial intelligence: a modern approach*, Prentice Hall series in artificial intelligence, Englewood Cliffs, N.J, Prentice Hall, 1995 à la p 16; Conseil de l'Europe, *supra* note 97.

99 Alan M Turing, « Computing Machinery and Intelligence », (1950) LIX:236 Mind 433, doi : <doi.org/10.1093/mind/LIX.236.433>.

100 *Ibid*. Conseil de l'Europe, *supra* note 97.

101 Alan Turing, *supra* note 99.

déterminer si une machine avait atteint un niveau d'intelligence supérieur ou égal à celui de l'humain[102].

Cependant, le terme « intelligence artificielle » fait son apparition plus tard, soit en 1956, lors d'une conférence à l'Université Darmouth. Le terme est proposé par John McCarthy et se cristallise par l'adoption du projet de recherche : *A Proposal for the Darmouth Research Project on Artificial Intelligence* où plusieurs chercheurs s'engagent à étudier l'intelligence artificielle symbolique pendant une période de deux mois[103]. Ces chercheurs avancent qu'« il est possible de décrire avec précision chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence afin qu'une machine puisse réussir à simuler des comportements intelligents »[104].

1.2 Évolution fulgurante de l'intelligence artificielle

Lors de la conférence à l'Université de Darmouth, Allen Newell, John C. Shaw et Herbert A. Simon présentent le *Logic Theorist*, (le théoricien de la logique) un ordinateur capable de reproduire la logique d'un humain dans sa résolution de problèmes[105].

S'ensuit, en 1958, la création du programme LISP (*List Processing*) par John McCarthy. À l'époque, McCarthy est parmi les premiers à créer un langage de programmation simple et interactif et celui-ci devient rapidement le langage de programmation standard[106]. LISP est encore utilisé aujourd'hui et est à l'origine des langages de programmation modernes comme JavaScript et Python[107].

L'intelligence artificielle symbolique est en plein essor et plusieurs chercheurs promettent des avancées extraordinaires dans les années à venir. Parmi ceux-ci, Hebert Simon énonce :

> Mon but n'est pas de vous surprendre ou de vous choquer, mais la façon la plus simple de résumer l'état de la recherche en intelligence artificielle est de dire qu'il y a maintenant, dans le monde, **des machines qui pensent, qui apprennent et qui créent**. En

---

102 Robert M French, « The Turing Test: The First Fifty Years » (2000) 4 *Trends in Cognitive Sciences* 115; Patrick Hayes & Kenneth Ford, *Turing Test Considered Harmful*, (1995) Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 1:972-77.

103 John McCarthy et al, *supra* note 91.

104 *Ibid*. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

105 Yann LeCun, sous-section « La logique avant toute chose » du chapitre 2 « Brève histoire de l'IA … et de ma carrière », dans *supra* note 95; voir *infra*, Chapitre 2, Sous-section 1.2.1 « Syllogisms and deductive reasoning » : le théoricien logique était chargé de partir d'un ensemble d'axiomes et de générer un certain nombre d'étapes intermédiaires pour arriver à un théorème, et ainsi le prouver. Le théoricien de la logique partait du problème et travaillait à rebours pour trouver un lien entre les axiomes et le théorème.

106 *Ibid*. Stuart J Russell & Peter Norvig, *supra* note 98.

107 Cade Metz, « John McCarthy -Father of AI and Lisp - Dies at 84 » *Wired*, en ligne : <https://www.wired.com/2011/10/john-mccarthy-father-of-ai-and-lisp-dies-at-84/>.

> outre, leur capacité à faire ces choses va rapidement s'accroître jusqu'à ce que, **dans un avenir visible**, la gamme des problèmes qu'elles peuvent traiter **soit égale à celle des capacités de l'esprit humain**.[108] [Nos soulignements]

Aidé par les médias qui rapportent l'arrivée imminente d'une intelligence artificielle capable d'apprendre par elle-même, le courant connexionniste lui non plus n'est en pas en reste.

En 1957, Frank Rosenblatt crée le Perceptron, un réseau de neurone artificiel simple capable d'effectuer des calculs qui lui permettent de détecter des tendances et classifier les données d'entrées qui lui sont soumises[109]. Dans un article de 1958 du *New York Times* sur le Perceptron *«* Electronic Brain Teaches Itself *»*[110] , il est rapporté que le programme sera le premier appareil électronique à penser comme le cerveau humain. L'article évoque aussi la possibilité de construire une version du Perceptron capable de se reproduire sur une chaîne de montage et d'être consciente de son existence. Finalement, l'article raconte que le gouvernement américain est tellement emballé par les premiers essais concluants du Perceptron, qu'il a hésité à l'appeler une machine tellement le programme est « semblable à un humain, mais sans vie »[111].

1.3 Puis une descente fulgurante

Les chercheurs sont enthousiastes et des développements rapides sont prévus. Pourtant, en 1959, Simon et Newell conçoivent le programme GPS (*General Problem Solver*) qui dérive du programme *Logic Therorist* et celui-ci ne produit pas les résultats espérés[112].

Fondé sur la logique symbolique, Simon et Newell ont observé les étudiants et leur processus de résolution de problèmes pour formaliser cette procédure dans un système informatique[113]. Le GPS prend les problèmes sous la forme d'objets et d'opérateurs. Les objets sont considérés comme étant similaires à des symboles et peuvent représenter n'importe quoi, d'expressions

---

108 Stuart J Russell & Peter Norvig, *supra* note 98 à la p 20: "It is not my aim to surprise or shock you—but the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which human mind has been applied."

109 Margot, P., « Perceptron: qu'est-ce que c'est et à quoi ça sert? », *DataScientest* (2021), en ligne : <https://datascientest.com/perceptron#:~:text=Un%20Perceptron%20est%20un%20neurone%20artificiel%2C%20et%20donc%20une%20unit%C3%A9,apprentissage%20supervis%C3%A9%20de%20classificateurs%20binaires>.

110 F Rosenblatt, « The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain » (1958) 65:6 *Psychological Review* 386, doi : <doi.org/10.1037/h0042519>.

111 « Electronic "Brain" Teaches Itself », *New York Times* 116 (1958), en ligne : <http://timesmachine.nytimes.com/timesmachine/1958/07/13/91396361.html>.

112 Stuart J Russell & Peter Norvig, *supra* note 98 à la p 17.

113 « Report on a General Problem Solving Program », en ligne : <http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf> à la p 2.

mathématiques aux pièces d'échec[114]. Les opérateurs transforment ces objets en d'autres objets. À titre d'exemple, il peut s'agir de mouvements d'échecs qui déplacent les pièces vers d'autres endroits de l'échiquier. L'utilisateur du système doit alors définir un objectif, comme celui de transformer un type d'objet en un autre. Par exemple, changer les positions des pièces sur un échiquier de façon à ce que l'adversaire soit mis en échec[115]. Le système tente ensuite d'atteindre l'objectif en concevant des sous-objectifs plus faciles que l'objectif global[116]. Ce faisant, il tente d'imiter la façon dont un humain résoudrait un problème pour atteindre un objectif[117].

Cependant, le programme est trop ambitieux pour l'époque, car celui-ci a une capacité d'action limitée et ne peut résoudre que des problèmes simples. En effet, il est difficile de décrire certains types de tâches en termes d'objets et d'opérateurs - comment décrire une tâche telle que faire du vélo ou reconnaître une image ? De plus, alors que la méthode devrait en théorie être capable de résoudre de nombreux problèmes complexes (comme jouer aux échecs), en pratique, le nombre énorme de possibilités rend cette approche irréalisable pour ce type de problèmes[118]. Bref, au-delà de certains types de problèmes simples, le GPS ne réussit pas à se représenter avec nuance, comme un humain, les problèmes qu'on lui pose[119].

Puis, le Conseil national de la recherche aux États-Unis coupe le financement pour les projets de traduction automatique après qu'un projet de traduction de textes russes n'ait pas eu les résultats espérés. Le modèle d'IA proposé est programmé à partir de manipulations syntaxiques très simples. Le programme de traduction ne connaît que très peu de choses sur son sujet et n'est pas en mesure d'interpréter le sens des mots dans leur contexte, ce qui donne lieu à des traductions erronées[120].

---

114 *Ibid.*, aux pp 3, 4.

115 *Ibid.,* à la p 6.

116 *Ibid.*, aux pp 8-24.

117 Stuart J Russell & Peter Norvig, *supra* note 98 à la p 18.

118 Voir *infra* chapitre 2, sous-section « 1.4.5 Hypothèse ontologique ».

119 Larousse, « Intelligence artificielle », *Encyclopédie Larousse en ligne,* en ligne : <https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257>.

120 Stuart J Russell & Peter Norvig, *supra* note 98 à la p 21; Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 190.

Pour aller plus loin :

Cardon, D., J-P, Cointet et A. Mazières, « La revanche des neurones: l'invention des machines inductives et la controverse de l'intelligence artificielle » (2018) 5:21

LeCun,Y., *Quand la machine apprend*, Odile Jacob éd, Paris, 2019

McCarthy J. et al., A Proposal for the Darmouth Summer Research Project on Artificial Intelligence, Darmouth, 1995.

Russell, Stuart J. Rus et P. Norvig, *Artificial intelligence: a modern approach*, Englewood Cliffs (N.J), Prentice Hall, 1995.

Turing, Alan, « Computing Machinery and Intelligence » (1950) LIX:236, 433.

1.4 Des promesses exagérées

Ces échecs marquent le début d'une série de coupures dans la recherche qui seront marquées par la publication de rapports visant à souligner que les avancées projetées en IA sont en fait largement exagérées.

1.4.1 L'alchimie et l'IA : une critique de l'approche symbolique

En 1965, le philosophe Hubert Dreyfus, à la demande de la Rand[121], publie un rapport sur les limites de l'IA intitulé *L'alchimie et l'IA*. Dreyfus y critique fortement l'approche symbolique et ses fondements philosophiques. Selon lui, en avançant que l'esprit humain puisse être réduit à un ensemble de règles, les chercheurs en IA se sont montrés trop optimistes. Dreyfus rejette l'hypothèse que l'humain puisse fonctionner comme une machine, à partir d'un ensemble de règles, car cela écarte l'idée selon laquelle l'humain dépend de processus inconscients pour prendre des décisions[122]. À la suite de son rapport, Dreyfus publie un ouvrage : *What Computer Can't Do*, *A Critique of Artificial Reason*[123] (ce que les ordinateurs ne peuvent pas faire) où il discute davantage des limites nommées dans son rapport.

Dans *Alchimie et IA* et *Ce que les ordinateurs ne peuvent pas faire*, Dreyfus identifie quatre hypothèses philosophiques à l'origine des premiers modèles d'intelligence artificielle symboliques : l'hypothèse biologique, psychologique, épistémologique et ontologique. À partir de ces hypothèses, il critique le fait que :

> Dans chaque cas, l'hypothèse est prise par les travailleurs en IA comme un axiome, **garantissant les résultats**, considérant qu'il ne s'agit en fait que d'une **hypothèse possible** parmi d'autres, à savoir **testée par le succès de ces travaux**.[124]

---

121 « History and Mission », en ligne : *RAND Corporation* <https://www.rand.org/about/history.html>.

122 Stuart J Russell & Peter Norvig, *supra* note 98 à la p 827; Hubert Dreyfus, *What Computers Can't Do, A Critique of Artificial Reason*, Harper & Row, 1972 à la p 68.

123 Hubert Dreyfus, *supra* note 122.

124 *Ibid.,* à la p 69.

Dreyfus s'appuie sur l'analyse de ces différentes hypothèses pour en souligner les difficultés conceptuelles, pour mieux comprendre l'optimisme persistant des chercheurs en IA et pour recadrer les promesses avancées par les chercheurs[125]. Dans son argumentaire, Dreyfus rejette les quatre hypothèses philosophiques de l'IA symbolique.

1.4.2 L'hypothèse biologique

L'hypothèse biologique avance que le cerveau traite l'information au moyen d'un équivalent biologique de la fonction marche/arrêt des interrupteurs[126]. Appuyée par Walter Pitts et Warren McCulloch, cette hypothèse fait valoir que les neurones fonctionnent comme un ordinateur et qu'il suffit d'alimenter la machine de symboles binaires de zéro et de un pour que celle-ci fonctionne comme le cerveau[127]. Dreyfus réfute cette hypothèse en s'appuyant sur des recherches en neurologie qui suggèrent que l'interaction entre les neurones du cerveau a des composantes analogiques, c'est-à-dire que l'information envoyée au cerveau est traitée en continu entre l'information initiale et sa représentation[128].

1.4.3 L'hypothèse psychologique

L'hypothèse psychologique s'appuie sur l'idée que l'esprit fonctionne selon des règles strictes. Au contraire, Dreyfus avance que les êtres humains en jouant, en résolvant des problèmes complexes ou en faisant des associations, interprètent les situations en s'appuyant sur le contexte dans lequel ils sont, leur apprentissage, leur interprétation d'une situation et le sens qu'il leur donne[129].

1.4.4 L'hypothèse épistémologique

L'hypothèse épistémologique s'appuie sur la croyance que toutes les connaissances peuvent être formalisées[130]. Pour Dreyfus, nous ne pouvons pas comprendre le comportement humain en s'appuyant exclusivement sur des règles. Pour appuyer son argument, Dreyfus reprend la critique Wittgenstienne de la règle qui soutient, comme Platon et Aristote, qu'il doit y avoir une place laissée à l'interprétation. En effet, pour être capable de réduire notre comportement en règles, il faudrait que celles-ci permettent à une machine de reconnaître le contexte dans lequel elles doivent être appliquées. Donc, il faudrait aussi établir des règles pour comprendre la situation,

125 *Ibid.*, à la p 68.

126 *Ibid*.

127 *Ibid.*, à la p 71.

128 *Ibid.*, aux pp 72, 73.

129 *Ibid.,* à la p 198.

130 *Ibid.*, à la p 115.

les intentions des intervenants, etc. Les règles seraient alors infinies, ce qui est impossible à accomplir[131].

1.4.5 Hypothèse ontologique

L'hypothèse ontologique suggère que le monde puisse être représenté de manière exhaustive par des symboles, soit la logique, le langage et les mathématiques. Selon Dreyfus, pour arriver à se représenter le monde de cette façon, il faudrait être en mesure d'alimenter la machine de données explicites et déterminées, où la pertinence et la signification d'une situation sont déjà données[132]. Pour Dreyfus, ces données n'existent pas, car les connaissances ne peuvent pas être classées en catégories[133]. Pour que cela soit possible, il faudrait être en mesure de se représenter une erreur, une collision ou n'importe quelle situation donnée à partir de données claires[134]. De plus, il faudrait que le programmeur informatique soit en mesure d'établir une hiérarchie de contextes et de règles générales pour guider la machine. Cela demande au programmeur de faire fi de ce qu'il a internalisé, de ses propres expériences et de son apprentissage pour poser un regard extérieur et neutre sur ce qu'il considère normalement comme allant de soi[135]. En bref, l'existence humaine ne peut pas être décrite par la logique, car l'interprétation d'une situation donnée ne peut être fixée. Celle-ci est influencée par l'acculturation et les changements d'auto-interprétation de l'être humain[136].

Pour finir, Dreyfus émet une hypothèse qui pourrait se rapprocher du courant connexionniste : il évoque la possibilité de programmer un ordinateur pour qu'il se comporte comme un enfant capable d'apprendre de son expérience pour devenir intelligent[137]. Bref, dans ces deux ouvrages, la position de Dreyfus est claire : les théories à l'origine de l'approche symbolique doivent être complètement rejetées. De plus, Dreyfus va plus loin encore en remettant en question les motifs sous-jacents qui poussent les chercheurs à s'intéresser à la discipline, soit le désir de vouloir créer des machines intelligentes semblables aux humains :

> Les ordinateurs ne peuvent traiter que des faits, mais l'homme, la source des faits, n'est pas un fait ou un ensemble de faits, mais plutôt un **être qui se crée lui-même** à partir du monde dans lequel il évolue. Ce monde humain avec ses objets reconnaissables est **organisé par les êtres humains** qui utilisent leurs capacités incarnées pour satisfaire leurs besoins incarnés. **Il n'y a aucune raison de supposer qu'un monde organisé en fonction**

131 *Ibid.* aux pp 115-117.

132 *Ibid.,* à la p 220.

133 *Ibid.,* à la p 118.

134 *Ibid.,* à la p 122.

135 *Ibid.,* à la p 201.

136 *Ibid*.

137 *Ibid.,* à la p 203.

> **de ses capacités humaines fondamentales devrait être accessible à tout autre**.[138] [nos soulignements]

C'est donc sans surprise que la sortie de ces deux ouvrages aura pour effet de mettre un frein à la recherche dans le domaine de l'IA symbolique pour les années à venir. Encore aujourd'hui, les critiques de Dreyfus demeurent d'actualité aujourd'hui et ont servi d'argumentaire au courant connexionniste pour remettre en question les fondements théoriques du courant symbolique.

Perceptrons : An Introduction to Computational Geometry (1969)[139] : une critique du courant connexionniste

À la même époque, le courant connexionniste surfe sur une vague de succès notamment grâce aux promesses rapportées par les médias sur le Perceptron[140]. En 1969, Marvin Minsky et Samuel Papert publient un ouvrage consacré à démontrer l'inefficacité des réseaux de neurones[141]. Les auteurs de l'ouvrage s'appuient sur une version simplifiée et à une couche du Perceptron pour démontrer que même si ce modèle de réseaux de neurones peut apprendre tout ce qu'il est capable de se représenter, il n'arrive à se représenter que très peu de choses[142]. Par cette critique majeure du fonctionnement des réseaux de neurones, les deux chercheurs tentent avant tout d'exclure l'approche connexionniste du domaine de la recherche en intelligence artificielle[143]. Cependant, la publication de l'ouvrage aura un effet encore plus large : soit de couper le financement et d'arrêter la recherche sur les réseaux neuronaux[144]. L'objectif des deux auteurs sera atteint, car pour les années à venir, les travaux liés à l'approche connexionniste se mèneront à l'écart du champ de l'intelligence artificielle[145]. Les efforts des chercheurs connexionnistes seront concentrés vers des projets plus réalistes comme la création des modems ou des filtres adaptatifs[146].

---

138 *Ibid*. Computers can only deal with facts, but man the source of facts is not a fact or set of facts, but a being who creates himself and the world of facts in the process of living in the world. This human world with its recognizable objects is organized by human beings using their embodied capacities to satisfy their embodied needs. There is no reason to suppose that a world organized in terms of these fundamental human capacities should be accessible by any other.

139 Marvin Minsky & Seymour Papert, *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge, Mass, 1969 258.

140 « Electronic "Brain" Teaches Itself », *supra* note 111.

141 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 11.

142 Stuart J Russell & Peter Norvig, *supra* note 98 à la p 22.

143 *Ibid*.

144 *Ibid*; Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 11.

145 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 11.

146 Yann LeCun, « L'hiver général », *supra* note 95.

Le rapport Lighthill (1972) : une critique de la discipline de l'IA

Au Royaume-Uni, le Conseil de la recherche scientifique commande un rapport indépendant et charge le mathématicien James Lighthill d'évaluer l'état de la recherche en IA afin que celui-ci puisse apporter un regard critique et extérieur au domaine de l'IA[147].

Dans son rapport *Artificial Intelligence : A General Survey*[148], Lighthill propose de diviser les domaines de l'IA en trois catégories de recherche qu'il appelle *The ABC of the subject*.

La lettre A signifie *Advanced Automation*. Cette catégorie de recherche étudie comment remplacer les humains par des machines pour des fins spécifiques. Ces fins peuvent être industrielles, militaires, mathématiques ou scientifiques. Par exemple, la reconnaissance vocale, la traduction, les missiles automatisés, la création et le stockage de banques de données, etc. Cette catégorie dépend particulièrement de la capacité des ordinateurs à stocker l'information et à résoudre des problèmes. L'objectif de cette catégorie est d'avoir des programmes capables de prendre des décisions et d'apprendre à partir de leur expérience[149].

La catégorie C fait référence au *Comptuer-based CNS Research*, soit le système nerveux central informatisé. Par exemple, les réseaux neuronaux. Cette catégorie s'inspire de la neurobiologie et psychologie pour créer des programmes capables d'exécuter des tâches comme la reconnaissance visuelle de formes, la mémoire et l'acquisition des connaissances[150].

Puis, il y a la catégorie B qui signifie non seulement *Bridge Activity*, mais aussi *Building Robots*. La catégorie B consiste à construire des dispositifs automatiques capables de simuler des comportements intelligents égaux à celui de l'humain dans le but d'alimenter les recherches dans les catégories A et C[151].

Comme Dreyfus, Lighthill critique l'optimisme des chercheurs en déclarant que « dans aucune partie du domaine, les découvertes faites jusqu'à présent n'ont produit l'impact majeur qui était alors promis »[152]. Lighthill admet qu'il y a eu des progrès dans les catégories A et C, mais moindres que ce qui avait été espéré et attendu. Par contre, c'est dans la catégorie B que les progrès sont considérés comme les plus décourageants. Lighthill considère que la promesse d'un programme d'intelligence artificielle doté de sens commun et capable de coordination a été largement

147 James Lighthill, « Artificial Intelligence : A General Survey », (1972), en ligne : <http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm>.

148 *Ibid,* « The ABC of the Subject».

149 *Ibid*.

150 *Ibid*.

151 *Ibid*.

152 *Ibid*., « Past Disappointments » : "In no part of the field have the discoveries made so far produced the major impact that was then promised."

gonflée[153]. Il admet qu'il soit possible de programmer des robots capables d'accomplir des tâches hautement spécialisées, mais considère que ces progrès sont insuffisants. En effet, ce que Lighthill critique le plus fortement, c'est l'incapacité des chercheurs à reconnaître les implications de l'explosion combinatoire dans le domaine de l'IA. Principe mathématique, l'explosion combinatoire est :

> L'obstacle général à la construction d'un système auto-organisé sur une large base de connaissances qui résulte de la croissance explosive de toute expression combinatoire, représentant un certain nombre de façons possibles de regrouper des éléments de la base de connaissances selon des règles particulières, à mesure de que la taille de la base augmente.[154]

Autrement dit, ce que Lighthill critique est le fait que bien qu'un programme d'intelligence artificielle puisse réussir à accomplir des tâches simples et précises, les chercheurs n'ont pas prévu la possibilité pour l'IA de résoudre des problèmes plus complexes devant tenir compte des contraintes, du contexte et des limites du problème donné. De plus, cette approche mise de l'avant par les chercheurs écarte complètement la notion d'intelligence émotionnelle qui est essentielle à notre survie. Un robot ne peut évoluer aux côtés d'humains et être considéré comme intelligent s'il n'est pas capable de nuance, de créer des liens ou de ressentir quoi que ce soit. Et donc, en mettant cet aspect de la recherche de côté, la catégorie B échoue à faire le pont entre la catégorie A et C. Sans ce lien entre les trois catégories, l'intelligence artificielle ne peut être considérée comme un domaine de recherche[155].

Finalement, Lighthill porte un regard intéressant sur la source de ces promesses exagérées et l'impact de celles-ci sur le domaine de la recherche. Selon lui, les chercheurs, souvent financés par des organismes publics ayant une mission particulière, vont préférer montrer une certaine unité afin d'aller chercher le maximum de financement. Un débat public sur la science, bien qu'il participe à la réflexion, peut aussi rendre les financiers plus réticents à financer les recherches sur lesquelles les chercheurs ne sont pas d'accord et qui pourraient ne pas produire de résultats concrets. Ainsi, plutôt que d'afficher leur désaccord et risquer de perdre des subventions, certains chercheurs évitent de remettre en question ces résultats projetés. En rendant ces projections publiques, les chercheurs participent à alimenter les attentes et les rendre plus élevées. Par contre, cela a aussi pour effet de rendre les échecs encore plus graves et embarrassants et participe à saper la confiance du public et des financiers dans la recherche en IA[156].

---

153 *Ibid*. « Past Disappointment: Category B ».

154 *Ibid*. "This is a general obstacle to the construction of a self-organising system on a large knowledge base which results from the explosive growth of any combinatorial expression, representing numbers of possible ways of grouping elements of the knowledge base according to particular rules, as the base's size increases."

155 *Ibid*.

156 *Ibid*.

Au final, invité à se projeter sur l'évolution de la recherche en intelligence artificielle, Lighthill prévoit une fission entre les trois catégories. Les catégories A et C continueront à évoluer parallèlement et réaliseront des progrès importants, tandis que les objectifs grandioses de la catégorie B ne seront pas réalisés. Au fil du temps, l'unité déjà fragile de ces trois catégories sera rompue. Sans lien entre ces catégories, Lighthill considère que le concept général de recherche en intelligence artificielle ne peut exister et que la discipline tend à s'essouffler[157].

Ces différents ouvrages et rapports, particulièrement critiques, visent à replacer les réels accomplissements et possibilités de la recherche en intelligence artificielle. Au-delà de cette évaluation, ces rapports viennent aussi remettre en question la légitimité et la nécessité d'un domaine de recherche en IA. Ces remises en question auront pour conséquence de couper les budgets de recherche et de freiner les différents développements de la discipline. Le domaine de la recherche en intelligence artificielle connaît donc ce que la communauté scientifique appelle son premier « hiver », soit une période creuse de 10 ans où il y aura très peu d'avancées dans le domaine.

Pour aller plus loin :

Cardon, D., J-P, Cointet et A. Mazières, « La revanche des neurones: l'invention des machines inductives et la controverse de l'intelligence artificielle » (2018) 5:21

Dreyfus, H., What Computers Can't Do, A Critique of Artificial Reason, Harper & Row, 1972

Lighthill, J. « Artificial Intelligence: A General Survey », (1972), en ligne : <http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm>.

Lussier-Lejeune, F., *Le développement sociohistorique de l'intelligence artificielle sous l'angle des économies de la promesse*, 9 décembre 2020, en ligne : <https://www.cyberjustice.ca/programme-virtuel/epistemologie-de-lia/>.

2. Une renaissance de l'approche symbolique : les systèmes experts

En 1980, l'intérêt renaît pour les programmes d'IA symboliques. La mémoire des ordinateurs, leur puissance de calcul et l'avènement des microprocesseurs permettent maintenant aux ordinateurs de mieux recevoir une grande quantité d'informations[158]. Ces progrès technologiques permettent entre autres d'améliorer l'architecture des systèmes symboliques[159]. Les premiers modèles d'IA symboliques, à l'architecture simple, rigide et axée sur des méthodes de résolution de problèmes, sont donc remplacés par les systèmes experts.

Comme leur nom l'indique, les systèmes experts incorporent des connaissances spécifiques à un domaine pour créer des systèmes capables de résoudre des problèmes typiques du monde réel,

157 *Ibid.* « The Next Twenty-Five years ».

158 Conseil de l'Europe, *supra* note 97.

159 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 14.

comme aider à poser des diagnostics ou à prendre des décisions[160]. Les systèmes experts cherchent donc à reproduire la structure et la logique d'un expert dans un domaine donné[161]. Pour formaliser ces connaissances, les programmeurs agissent comme « ingénieurs de la connaissance »[162]. Ceux-ci détaillent le raisonnement d'un expert pour pouvoir le traduire en règle et ainsi alimenter le programme d'intelligence artificielle. Une fois les règles encodées, le système doit trouver un moyen de raisonner et d'interpréter les règles afin d'arriver à une conclusion qui pourrait être utile à l'utilisateur. Cette partie peut être désignée comme un moteur d'inférence. Le moteur doit tenir compte d'un certain contexte (tel qu'un lecteur de capteur ou des réponses à des questions posées précédemment par le système), puis utiliser les règles pour trouver une réponse à un problème[163]. Une fois la réponse trouvée, le système peut interagir avec l'utilisateur et lui permettre d'explorer les étapes de raisonnement qui l'ont mené à une certaine réponse[164]. Bref, les systèmes experts sont capables de déterminer quand et comment appliquer des règles en fonction des faits qu'ils reçoivent et expliquer leur processus de raisonnement[165].

À titre d'exemple, on peut citer le programme MYCIN. Développé au début des années 70 à l'Université Stanford, MYCIN devait servir d'outil aux médecins afin de les aider à identifier des infections du sang et proposer un traitement[166]. Dans leur première demande de subvention en octobre 1973, les chercheurs présentent le programme MYCIN comme un programme informatique interactif destiné à :

> Fournir aux médecins des **conseils consultatifs** concernant **un choix approprié** de la thérapie antimicrobienne tel que déterminé **à partir des données disponibles dans les laboratoires de microbiologie** et de chimie clinique et **des observations cliniques directes saisies par le médecin en réponse à des questions générées par ordinateur**.[167] [nos soulignements]

---

160 Voir *infra*, Chapitre 2, « Expert System, Technological Explanation ».

161 *Ibid*.

162 Yann LeCun, *supra* note 95, « GOFAI ».

163 Bruce G Buchanan & Reid G Smith, *supra* note 24, aux pp 17–20; Voir i*nfra*, chapitre 2, « The inference engine ».

164 *Ibid.*, aux pp 20, 21; Voir i*nfra*, Chapitre 2, « User interface ».

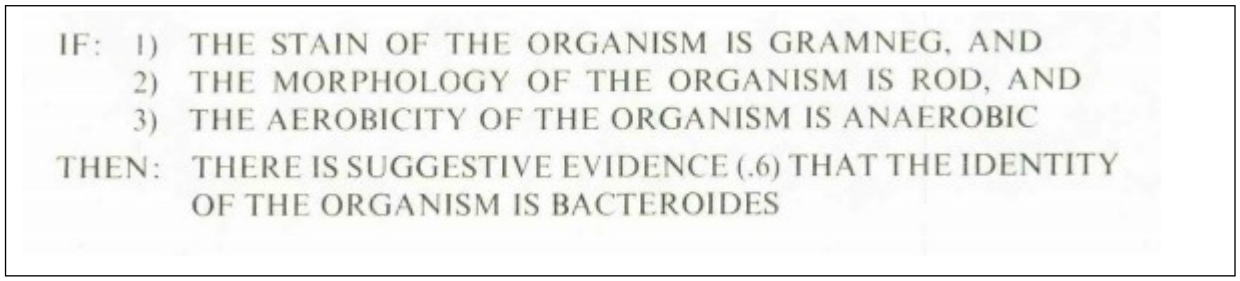
165 Mark Stefik et al, « The Organization of Expert Systems, a Tutorial » (1982) 18:2 Artificial Intelligence 135-173.

166 Bruce G Buchanan & Edward Hance Shortliffe, dir., *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, The Addison-Wesley series in artificial intelligence, (MA: Addison-Wesley, 1984) "The Context of the MYCIN Experiment".

167 *Ibid* à la p 10: central component of the MYCIN system is an interactive computer program to provide physicians with consultative advice regarding an appropriate choice of antimicrobial therapy as determined from data available from the microbiology and clinical chemistry laboratories and from direct clinical observations entered by the physician in response to computer-generated questions.

Le système MYCIN a donc été entraîné et nourrit à partir d'une série de règles modifiables et conditionnelles[168]. **Si (une prémisse fixée)** tels symptômes sont présents **et** que l'organisme présente telles anomalies **alors** MYCIN établit un pourcentage de probabilité qu'une personne soit atteinte de telle infection[169] (voir figure 1). Ensuite, MYCIN émet plusieurs hypothèses de diagnostic. Plus il y aura d'informations sur les symptômes, plus les hypothèses qu'il a en banque seront réduites et influencées par les réponses qui lui sont données. Au final, MYCIN sera capable de poser un diagnostic avec un indice de confiance (probabilité)[170]. Aussi, pour permettre au médecin d'explorer comment et pourquoi une règle a été choisie, MYCIN offrait un système d'explication[171]. Il était également possible pour les médecins d'ajouter de manière interactive de nouvelles règles au système[172].

Figure 1
Un exemple de règle de décision du système MYCIN[173]


IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) THE AEROBICITY OF THE ORGANISM IS ANAEROBIC
THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE IDENTITY OF THE ORGANISM IS BACTEROIDES

Cette nouvelle architecture des programmes symboliques permet d'offrir des solutions aux lacunes soulevées lors du premier hiver de l'IA. D'abord, en retraçant le processus intellectuel d'un expert, le contexte est intégré aux règles. Le programme est alimenté de données (connaissances) tirées du monde réel (l'expert) où les erreurs, le contexte et les différents choix à prendre en compte sont considérés[174]. Ensuite, le résultat probabiliste des systèmes experts permet d'obtenir des résultats moins rigides. Les règles constituent des règles d'expertise, de bonnes pratiques qui n'ont plus à donner un résultat rigide : oui/non, vrai/faux. Le système, en s'appuyant sur différentes hypothèses, propose la solution la plus probable. Ainsi, le résultat est

168 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 aux pp 15, 16; Bruce G Buchanan & Edward Hance Shortliffe, *supra* note 166 à la p 4.

169 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 15; Bruce G Buchanan & Edward Hance Shortliffe, *supra* note 166.

170 Yann LeCun, *supra* note 95, « La logique avant toute chose ».

171 Bruce G Buchanan & Edward Hance Shortliffe, *supra* note 166 aux pp 315-318; Voir *infra* Chapitre 2 « MYCIN – an expert system for medical diagnosis ».

172 *Ibid*.

173 *Ibid.*, à la p 305

174 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 17.

conditionnel. Il dépend des nuances du monde réel que le système ne peut percevoir. Le résultat demeure une possibilité, car il est possible qu'il ne puisse pas s'appliquer à la situation donnée[175].

L'engouement pour l'intelligence artificielle symbolique reprend à toute vitesse. Pourtant, l'histoire tend à se répéter et les systèmes experts ne tiennent pas leurs promesses. Pour reprendre l'exemple donné plus haut du programme MYCIN, celui-ci ne dépasse même pas le stade expérimental[176]. Le programme s'intègre difficilement au quotidien. La saisie de données sur le programme est complexe, peu fiable et peut prendre jusqu'à 1 heure par consultation[177]. Au final, l'expert arrive à faire plus vite en faisant le cheminement par soi-même. De plus, les données (connaissances) qui alimentent le système deviennent rapidement obsolètes. Malgré l'approche contextualisée, les connaissances évoluent à grande vitesse et le système n'arrive pas à suivre ces mutations[178]. Cela demande un soutien constant d'un professionnel. Dans ce contexte, il devient irréaliste de recenser manuellement l'ensemble des connaissances d'un expert pour les réduire en règles. Le même problème demeure : l'IA symbolique peine à intégrer les subtilités du monde réel[179].

Plus largement, la programmation de systèmes experts est très complexe et demande en moyenne la mise en place de 200 à 600 règles[180]. À force d'alimenter le système d'une multitude de règles, il n'est plus possible de retracer le raisonnement du programme[181]. Cependant, comprendre le raisonnement est essentiel pour ce type de système, car il permet à son utilisateur de comprendre le choix d'hypothèse mis de l'avant par la machine[182]. Finalement, dans son ouvrage The *Applied Side of Artificial Intelligence*, l'informaticien Edward Feigenbaum, pionner des systèmes experts, avance que l'approche mise de l'avant dans la conception des systèmes experts, si elle est n'est pas davantage automatisée, apparaît de plus en plus comme une tâche impossible et inefficace :

> Ces connaissances sont actuellement acquises d'une manière très minutieuse qui rappelle les industries artisanales, dans lesquelles **les informaticiens individuels travaillent avec des experts individuels dans des disciplines minutieuses** pour expliquer les heuristiques. Si l'intelligence artificielle appliquée doit être importante dans les décennies à venir, nous

---

175 *Ibid.,* à la p 18.

176 Bruce G Buchanan & Edward Hance Shortliffe, *supra* note 166 à la p 673 « Experimental results ».

177 *Ibid.,* à la p 600.

178 *Ibid.*, « Human Engineering of Medical Expert Systems ».

179 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 18.

180 Conseil de l'Europe, *supra* note 97; Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 17.

181 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91; Conseil de l'Europe, *supra* note 97.

182 Edward A Feigenbaum, « Knowledge Engineering.: The Applied Side of Artificial Intelligence » (1984) 426:1 *Computer Cult Ann NY Acad Sci* 91-107.

> devons **disposer de moyens plus automatiques** pour remplacer ce qui est actuellement une procédure très fastidieuse, longue et coûteuse.[183] [nos soulignements]

À partir de la fin des années 1980, le domaine de la recherche en intelligence artificielle symbolique entame son deuxième hiver et tombe en désuétude pour les années à venir[184]. Seul espoir : en 1997, Deep Blue, un ordinateur d'IBM entraîné par un système expert, réussi à battre le champion d'échec Gary Kasparov[185]. Capable d'analyser et prévoir plus de 200 millions de coups possibles en échec, cette victoire est une des premières à confirmer qu'il est possible pour un système d'IA de rivaliser avec l'intelligence humaine[186]. Cette victoire demeure toutefois symbolique, car les capacités de Deep Blue se limitent à analyser les règles du jeu d'échec sans pour autant être capable de comprendre le jeu ou le contexte dans lequel celui-ci était appelé à jouer[187]. C'est pourquoi, cette victoire n'est pas suffisante à susciter un regain d'intérêt pour le courant symbolique et les développements en IA demeurent limités.

Bref, marqué par ces deux hivers, le domaine de la recherche s'en retrouve durement touché et le terme d'intelligence artificielle n'est presque plus utilisé[188]. L'utilisation de termes plus prudents comme « informatique avancée » est préférée[189]. Par contre, comme l'a projeté Feigaibaum, l'automatisation des connaissances aura finalement lieu, mais pas par les méthodes envisagées à l'époque. Il sera plutôt question de méthodes statistiques basées sur l'apprentissage automatique. Toutefois, on recense encore aujourd'hui certaines applications de systèmes experts. Notamment, dans le domaine des finances[190] et des affaires[191].

---

183 *Ibid.,* à la p 2. This knowledge is currently acquired in a very painstaking way that reminds one of cottage industries, in which individual computer scientists work with individual experts in disciplines painstakingly to explicate heuristics. If applied Artificial Intelligence is to be important in the decades to come, we must have more automatic means for replacing what is currently a very tedious, time-consuming and expensive procedure.

184 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 18.

185 Conseil de l'Europe, *supra* note 97.

186 Luke Stark, Zenon W. Pylyshyn, « Intelligence artificielle (IA) au Canada », (6 février 2066) *L'encyclopédie canadienne*, en ligne : <https://www.thecanadianencyclopedia.ca/fr/article/intelligence-artificielle>.

187 *Ibid.*

188 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 18; Conseil de l'Europe, *supra* note 97.

189 Conseil de l'Europe, *supra* note 97.

190 Voir Noura Metawa, Mohamed Elhoseny, & Aboul Ella Hassanien, « Expert Systems in Finance: Smart Financial Applications in Big Data Environments » *Routledge & CRC Press*, en ligne : <https://www.routledge.com/Expert-Systems-in-Finance-Smart-Financial-Applications-in-Big-Data-Environments/Metawa-Elhoseny-Hassanien-Hassan/p/book/9780367729011>.

191 Voir K Metaxiotis & John Psarras, « Expert systems in business: applications and future directions for the operations researcher » (2003) 103:5 *Industrial Management & Data Systems* 361-368.

Pour aller plus loin :

Buchanan, B. & E. Hance Shortliffe, *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, The Addison-Wesley series in artificial intelligence, (MA: Addison-Wesley, 1984)

Cardon, D., J-P, Cointet et A. Mazières, « La revanche des neurones: l'invention des machines inductives et la controverse de l'intelligence artificielle » (2018) 5:21

LeCun,Y., *Quand la machine apprend*, Odile Jacob éd, Paris, 2019

3. Une fenêtre d'opportunité pour l'apprentissage automatique

Les problèmes récurrents rencontrés par l'IA symbolique ouvrent la porte à un changement de paradigme dans le domaine de la recherche en intelligence artificielle :

> Les systèmes experts ont mis en évidence la nécessité pour les systèmes de pouvoir s'appuyer sur **des détails d'expérience, de détecter des cas similaires, de prendre des décisions sur les cas pertinents et d'appliquer les connaissances existantes**. Ces systèmes devraient être capables d'identifier les similitudes entre des problèmes difficiles récurrents afin de créer de nouveaux cas et scénarios, ainsi que de **mettre à jour leurs ensembles de règles existants**. Ces systèmes devaient permettre aux machines de **« tirer des leçons » de leurs expériences**. Cette prise de conscience a donné naissance au terme désormais largement utilisé d'« **apprentissage automatique** », qui regroupe les statistiques, la logique floue, l'acquisition de connaissances, l'intelligence artificielle, les bases de données, l'exploration de données, l'informatique et les neurosciences en 1987 et 1989.[192] [nos soulignements]

Le raisonnement logique ne suffit pas à simuler l'intelligence. Pour apprendre, une machine doit être capable de perception, d'intuition et de se fier à ses expériences[193]. Plutôt que de coder des règles, une nouvelle approche inductive est préférée, soit celle d'entraîner la machine à apprendre par elle-même. À partir de la fin des années 80, l'apprentissage automatique (*machine learning*) devient la méthode dominante dans le domaine de l'IA[194].

---

192 Amy Shi-Nash & David R Hardoon, *Data Analytics and Predictive Analytics in the Era of Big Data*, internet of things and data analytics handbooks éd., John Wiley & Sons, 2017 à la p 330: "The difficulties faced by expert systems highlighted the need for systems to be able to fall back on details of experience, detect similar cases, make decisions on which case was relevant, and apply existing knowledge. These systems would need to be able to identify similarities between recurring tough problems to create new cases and scenarios, as well as update their existing sets of rules. These systems needed to enable machines to « learn » from their experiences. This realization gave birth to the now widely used term of « machine learning »; a coming together of statistics, fuzzy logic, knowledge acquisition, artificial intelligence, databases, data mining, computer science, and neuroscience back in 1987 and 1989".

193 Yann LeCun, « Machine Learning », *supra* note 95.

194 Conseil de l'Europe, *supra* note 97.

Parmi les techniques d'apprentissage automatique, c'est l'apprentissage profond qui est aujourd'hui préconisé. Inspirée des travaux connexionnistes de la fin des années 1950, la consécration de l'apprentissage profond en 2010 marque un renouveau du courant connexionniste après un long hiver et une histoire marquée par des développements interrompus.

Consécration de l'approche connexionniste par l'apprentissage profond

Le courant connexionniste connaît une recrudescence au milieu des années 80, grâce à la publication d'un article par David Rumelhart, Geoffrey Hinton et Ronald Williams sur la rétropropagation de gradient[195]. La rétropropagation offre une solution aux critiques exposées par Minsky et Papert[196] en intégrant l'hypothèse de réseaux de neurones multicouches et jette les bases de l'apprentissage profond (*deep learning*)[197]. Comme l'explique Yann LeCun dans son ouvrage *Quand la machine apprend*[198], la rétropropagation :

> Permet l'entraînement de réseaux de neurones multicouches, constitués de milliers de neurones organisés en couches, avec des centaines de milliers de connexions. Chaque couche de neurones combine, traite et transforme les informations de la couche précédente, et transmet le résultat à la couche suivante, jusqu'à produire une réponse sur la couche finale.[199]

Parallèlement à cela, le physicien Terry Sejnowski présente *NetTalk*, un réseau multicouche entraîné par rétropropagation capable de transformer un texte en phrases vocalisées[200]. Les résultats impressionnants du programme attirent l'attention des chercheurs sur la question des réseaux de neurones. Par contre, cet intérêt n'est pas suffisant. Certains chercheurs continuent de lui opposer d'autres systèmes tout aussi efficaces pour les jeux de données qu'il y a traités, notamment les modèles par analogie[201]. Les développements dans le domaine de l'apprentissage profond demeurent en marge, mais cela n'empêche pas certaines avancées importantes. Par exemple, en 1995, Yann LeCun invente les réseaux convolutifs[202], un algorithme capable, entre autres, de reconnaître les chiffres sur les codes postaux et les chèques[203]. Malgré ces progrès intéressants, 1995 marque aussi le début d'années noires pour les réseaux de neurones au sein

---

195 David E Rumelhart, Geoffrey E Hinton et Ronald J Williams, « Learning representations by back-propagating errors » (1986) 323 Nature 533, doi : <doi.org/10.1038/323533a0>.

196 Voir supra Chapitre 2, sous-section « 2.2.5 Discussion ».

197 Yann LeCun, « Machine Learning », *supra* note 95.

198 *Ibid*.

199 *Ibid.* « De l'usage de la rétropropagation de gradient ».

200 *Ibid*. Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 19.

201 *Ibid*.

202 Voir infra Chapitre 2, « Convolutional Neural Networks ».

203 Yann LeCun, « Les années Bell Labs », *supra* note 95; Cardon, Cointet & Mazières, *supra* note 91 à la p 20.

des recherches concernant l'apprentissage automatique[204]. Les résultats demeurent limités, les publications sur la question sont majoritairement refusées et les chercheurs connexionnistes obtiennent peu de soutien institutionnel[205]. Pour expliquer cette réticence, Yann Lecun explique d'abord que les réseaux de neurones sont considérés comme compliqués :

> Les réseaux convolutifs sont très gourmands en calculs. Or à l'époque les ordinateurs sont lents et coûteux ; les jeux de données sont trop petits – nous sommes avant l'explosion d'Internet. Il faut donc les collecter soi-même, ce qui a un prix et limite les applications ; enfin, les logiciels pour les réseaux de neurones, comme SN, doivent être écrits à la main de A à Z par les chercheurs eux-mêmes.[206]

Ensuite, LeCun souligne qu'à l'époque, les entreprises pratiquent le « chacun-pour-soi » et que la plupart des simulateurs de réseau de neurones ne sont pas disponibles en *open source*, ce qui rend la technologie peu accessible et difficile à adopter[207]. Entre 1995 et 2010, d'autres méthodes appartenant au domaine de l'apprentissage automatique, comme les machines à vecteurs de support (SVM)[208] et les méthodes à noyau[209] sont préférées. Les réseaux de neurones entament un hiver d'une quinzaine d'années[210].

En 2004, l'Institut canadien de recherches avancées (l'ICRA) lance un programme qui s'échelonne sur cinq ans : *Calcul neuronal et perception adaptative* dirigé par Geoffrey Hinton[211]. Le programme, dont font partie Yann Lecun et Yoshua Bengio, permet d'élargir le cercle de la recherche dans le domaine de l'apprentissage profond et de créer un noyau de chercheurs intéressés à faire circuler et alimenter l'état de la recherche sur la question des réseaux de neurones[212]. Par contre, entre 2004 et 2006, les articles sur la question continuent de se faire refuser dans les congrès d'apprentissage automatique[213]. C'est finalement en 2007, lors du congrès NIPS qui rallie les chercheurs en apprentissage automatique, que le terme *deep learning* (apprentissage profond) est adopté dans la littérature spécialisée de l'IA[214]. En effet, lors de ce congrès, Bengio, LeCun et Hinton organisent une session pirate où ils présentent en quoi les réseaux de neurones sont plus efficaces et performants que les SVM. L'atelier a un succès

---

204 Yann LeCun, « Un Tabou ? », *supra* note 95.

205 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 21.

206 Yann LeCun, « Un Tabou ? », *supra* note 95.

207 *Ibid*.

208 Voir *infra*, Chapitre 2, « Supervised technologies ».

209 *Ibid*.

210 Yann LeCun, « Un Tabou ? », *supra* note 95.

211 *Ibid.* « La conspiration du Deep Learning »; « Notre histoire », en ligne : *CIFAR* <https://cifar.ca/fr/notre-histoire/>.

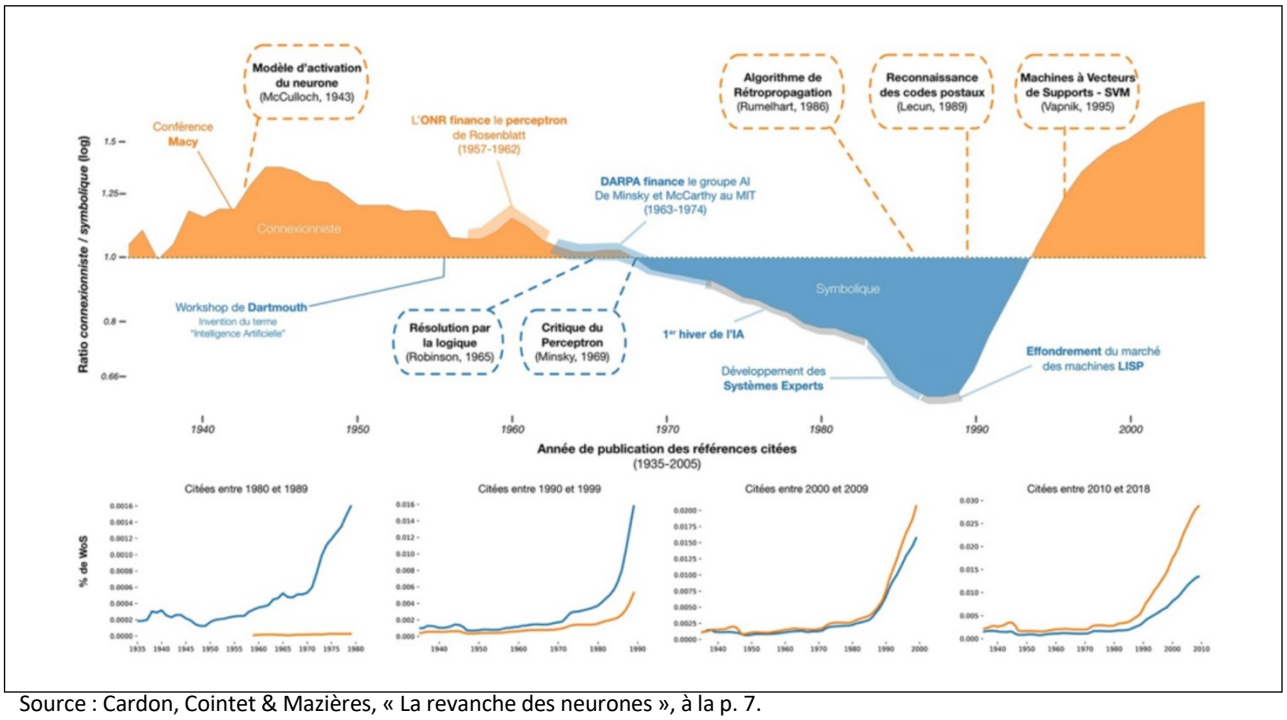
212 Yann LeCun, « La conspiration du Deep Learning », *supra* note 95.

213 *Ibid*.

214 *Ibid*.

inespéré et marque un tournant dans l'adoption de l'apprentissage profond comme méthode d'apprentissage[215].

Figure 2
Évolution de l'influence académique des approches connexionniste et symbolique



Source : Cardon, Cointet & Mazières, « La revanche des neurones », à la p. 7.

La consécration de l'approche connexionniste, des réseaux de neurones et de son application au *deep learning* a finalement lieu aux alentours de 2010 et s'explique par l'avènement d'Internet[216]. D'abord, la croissance exponentielle de la puissance des calculateurs sur les ordinateurs permet un entraînement plus rapide, à moindre coût, des réseaux de neurones[217]. Ensuite, l'accès à un plus grand volume de données facilite l'échantillonnage nécessaire à l'entraînement de l'algorithme[218]. De plus, l'accès à ces données massives participe à améliorer considérablement la performance des algorithmes dans le traitement de la reconnaissance vocale et de la reconnaissance d'image[219]. En effet, lors du concours de reconnaissance d'images *ImageNet* de 2012, un réseau de neurones développé par Geoffrey Hinton surpasse tous ses concurrents. Le taux d'erreur de la machine est de 17 % comparé au taux d'erreur moyen de 25 % en matière de

215 *Ibid*. Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 22.

216 Conseil de l'Europe, *supra* note 97.

217 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 31; Conseil de l'Europe, *supra* note 97.

218 *Ibid*.

219 Conseil de l'Europe, *supra* note 97.

traitement d'image par d'autres programmes[220]. L'efficacité des réseaux de neurones artificiels et leur plus-value par rapport aux méthodes conventionnelles sont progressivement prouvées dans plusieurs domaines d'application de l'IA. Notamment, dans le domaine de l'analyse boursière[221], la géographie urbaine[222] et la reconnaissance des formes[223]. Ces résultats sont en grande partie responsables de la montée fulgurante de l'intérêt en recherche pour l'apprentissage profond. Soudainement, les publications concernant l'apprentissage profond ne sont plus ignorées (voir *supra* figure 2). Par exemple, l'article *Gradient-Based Learning Applied to Document Recognition*[224] publié en 1998 par Yann LeCun, Léon Bottou, Yoshua Bengio et Patrick Haffner au sujet du fonctionnement des réseaux convolutifs passe de quelques dizaines de citations par an à 5 400 citations en 2018. En 2019, il devient un des articles les plus cités dans le domaine avec plus de 20 000 citations. Il est aujourd'hui considéré comme l'article fondateur des réseaux convolutifs[225].

En 2016, AlphaGo, le système d'IA de Google spécialisé dans le jeu de Go, bat le champion du monde Fan Hui. Entraîné à partir d'une méthode d'apprentissage profond[226], AlphaGo a joué plus d'un million de parties contre lui-même afin de découvrir les règles du jeu, apprendre de ses erreurs et devenir meilleur[227]. Cette victoire est une grande surprise pour plusieurs, car le Go était alors considéré comme le jeu le plus difficile à maîtriser pour les ordinateurs, en raison du nombre massif de mouvements possibles et des différents états de jeu.[228]. La méthode d'apprentissage automatique utilisée pour entraîner AlphaGo se montre donc plus efficace que la méthode par système expert utilisée pour concevoir DeepBlue[229].

Finalement, en 2018[230], Yoshua Bengio, Yann LeCun et Geoffrey Hinton remportent le prix Turing de *l'Association for Computing Machinery* (ACM) de New York. Qualifiés de « pères de la

---

220 Dominique Cardon, Jean-Philippe Cointet & Antoine Mazières, *supra* note 91 à la p 2; voir *infra* Chapitre 2, « Using an existing dataset ».

221 Voir M Firdaus et al, *Literature review on Artificial Neural Networks Techniques Application for Stock Market Prediction and as Decision Support Tools* (2018).

222 Voir George Grekousis, « Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis » (2019) 74 *Computers, Environment and Urban Systems* 244.

223 Voir O I Abiodun et al, « Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition » (2019) 7 *IEEE Access* 158820.

224 Voir Yann LeCun et al, « Gradient-Based Learning Applied to Document Recognition » (1998) *PROCOF THE IEEE*, en ligne : <http://yann.lecun.com/exdb/publis/pdf/lecun-98.pdf>.

225 Yann LeCun, « Un Tabou ? », *supra* note 95.

226 Voir *infra*, Chapitre 2 « Deep Reinforcement Learning ».

227 DeepMind, « Alpha go », en ligne : <https://deepmind.com/research/case-studies/alphago-the-story-so-far>.

228 Voir *infra*, Chapitre 2 « Deep Reinforcement Learning ».

229 Voir supra Chapitre 1, Section 2, Sous-section « 2. Une renaissance de l'approche symbolique : les systèmes experts ».

230 Voir « A.M. Turing Award Winners by Year », en ligne : <https://amturing.acm.org/byyear.cfm>.

révolution de l'apprentissage profond », l'ACM souligne que « la croissance de l'intelligence artificielle et l'intérêt qu'elle suscite viennent, en grande partie, des récentes percées en apprentissage profond dont Yoshua Bengio, Geoffrey Hinton et Yann LeCun ont posé les fondements »[231]. Cette nomination a donc une portée très symbolique : le courant connexionniste, longtemps mis à part, est maintenant au-devant de la scène du champ de l'IA. Cela signifie que les connexionnistes ont réussi à intégrer le domaine de l'intelligence artificielle en mettant de l'avant une approche qui au départ devait être exclue de la discipline.

Malgré la consécration de l'approche connexionniste, l'apport du courant symbolique n'est pas à minimiser. L'IA symbolique a servi à jeter les bases théoriques de la discipline et est souvent intégrée aux applications d'apprentissage automatique. Par exemple, une voiture autonome équipée d'un système de reconnaissance visuelle capable de reconnaître les objets sur la route est entraînée par des réseaux neuronaux convolutifs. Cependant, une fois ces objets détectés, ce sont des systèmes classiques de trajectoires, à base de règles, qui sont entraînés à prendre une décision[232]. De plus, les systèmes d'IA symbolique ont l'avantage d'être faciles à expliquer. Il est plus commode de documenter la prise de décision d'un algorithme quand celui-ci s'appuie sur des bases de faits et de règles. Au contraire, les algorithmes d'apprentissage profond sont critiqués pour leur effet « boîte noire ». Les systèmes sont tellement complexes qu'à partir du moment où la machine apprend par elle-même, son fonctionnement devient opaque et il n'est pas possible de retracer son processus de décision[233]. Pour répondre à cet enjeu, l'idée d'intégrer les deux approches, la symbolique par la connexionniste, est envisagée. Il s'agit d'utiliser un système d'IA symbolique pour enseigner un ensemble de règles à une autre machine d'apprentissage profond. En plus d'offrir une solution à l'effet « boîte noire », ce transfert de connaissances permet à la machine d'apprentissage profond de raisonner à un niveau abstrait et d'atteindre de meilleures fonctions cognitives comme l'apprentissage par transfert, le raisonnement par analogie et le raisonnement basé sur des hypothèses[234].

---

231 UdeMNouvelles, « Le prix Nobel de l'informatique » (2019), en ligne : <https://nouvelles.umontreal.ca/article/2019/03/27/le-prix-nobel-de-l-informatique/>.

232 Yann LeCun, *supra* note 95, « Cocktail d'ancien et de moderne ». Voir aussi Jiayuan Mao et al, « The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision » (2019), en ligne : <http://arxiv.org/abs/1904.12584>: le *Neuro-Symbolic Concept Learner*, apprend les concepts visuels, les mots et l'analyse sémantique des phrases grâce à un module de raisonnement neuro-symbolique.

233 Marta Garnelo, Kai Arulkumaran & Murray Shanahan, « Towards Deep Symbolic Reinforcement Learning » (2016), en ligne : <http://arxiv.org/abs/1609.05518>.

234 *Ibid*. Voir aussi Lina, Lim Tong Ming & Leow Soo Kar, *A Hybrid Connectionist-Symbolic Approach for Real-Valued Pattern Classification*, Boston, MA, Springer US, 2005; Vasant Honavar & Leonard Uhr, « Symbolic Artificial Intelligence, Connectionist Networks & Beyond. » (1994) 76 *Computer Science Technical Reports* 45.

Pour aller plus loin :

Cardon, D., J-P, Cointet et A. Mazières, « La revanche des neurones: l'invention des machines inductives et la controverse de l'intelligence artificielle » (2018) 5:21

Conseil de l'Europe, « Histoire de l'intelligence artificielle », en ligne : *Conseil de l'Europe* <https://www.coe.int/fr/web/artificial-intelligence/history-of-ai>.

LeCun, Y., *Quand la machine apprend*, Odile Jacob éd, Paris, 2019

Conclusion

En conclusion, l'histoire de l'intelligence artificielle nous apprend que l'intelligence artificielle est une discipline aux frontières mouvantes. Ce qui était considéré comme appartenant au domaine de l'intelligence artificielle dans les débuts de la discipline ne l'est plus nécessairement aujourd'hui et tend à évoluer à la vitesse des innovations technologiques. Bien que les modèles d'intelligence artificielle soient maintenant capables d'accomplir des tâches spécifiques, parfois même mieux que les humains, un programme d'IA est encore bien loin d'avoir les capacités cognitives d'un humain. Dans ce contexte, le terme intelligence artificielle est-il approprié pour désigner la discipline ? Qu'est-ce qui caractérise l'intelligence artificielle ? C'est ce qu'il convient d'évaluer dans la prochaine section de ce chapitre.

Section 3 - Définir l'(intelligence) artificielle : une discipline aux frontières floues

Dans cette section sur la définition d'intelligence artificielle (IA), il sera question de présenter :

- en quoi l'intelligence artificielle est différente des concepts connexes auxquels elle est souvent associée;
- une taxonomie de l'IA développée par la Commission européenne en 2019;
- en quoi comparer l'intelligence artificielle à l'intelligence humaine peut être trompeur et génère des discussions autour d'un concept abstrait et mal défini;
- en quoi remplacer le terme intelligence artificielle par intelligence augmentée est une proposition qui doit être adoptée plus largement dans nos travaux, car c'est une terminologie plus nuancée et précise pour se référer à la discipline;
- une revue des définitions de l'intelligence artificielle à travers trois perspectives des différents acteurs du milieu de l'IA : la perspective politique, de recherche et de l'industrie;
- proposer une définition complète et précise de l'intelligence artificielle.

Avant de se pencher sur les définitions possibles de l'intelligence artificielle (IA), il est intéressant de présenter ce que l'on entend généralement par « intelligence artificielle ». L'intelligence artificielle est une discipline aux frontières floues dont les développements présents et projetés font couler beaucoup d'encre. Par la recherche, l'enseignement supérieur, l'industrie, les états et les organisations internationales, le développement et le déploiement de l'IA semblent concerner tout et tout le monde sans qu'il soit encore possible d'en proposer une définition commune. Sans définition claire de l'intelligence artificielle, il est difficile d'en comprendre les progrès et les limites. Ce manque de précision participe à alimenter l'imaginaire populaire où tout progrès semble être l'annonce de l'imminence d'une intelligence artificielle superpuissante capable de détruire et/ou de remplacer l'espèce humaine. Pourtant, ces scénarios de science-fiction sont très peu plausibles. C'est pourquoi, il sera d'abord question d'analyser en quoi l'IA est différente des concepts connexes auxquels elle est souvent associée pour ensuite présenter une taxonomie de l'IA développée par la Commission européenne en 2019[235]. Enfin, à partir de ces observations et d'une revue des définitions de l'IA, nous pourrons proposer une définition de l'intelligence artificielle.

[235] European Commission, *Ai Watch: Defining Artificial Intelligence: Towards an Operational Definition and Taxonomy of Artificial Intelligence*, Publications Office of the European Union, 2020.

1. Sémantique : Qu'est-ce que l'on entend par intelligence artificielle ?

1.1 Intelligence artificielle et intelligence artificielle symbolique

Comme il en a été question dans la section précédente, l'intelligence artificielle symbolique est, pour certains, tombée en désuétude et ne fait plus partie du domaine de l'intelligence artificielle. En effet, toutes définitions qui considèrent l'intelligence artificielle comme des systèmes capables d'apprendre par expérience excluent nécessairement l'IA symbolique. Par exemple, Andréas Kaplan[236], dans un article qui s'intéresse aux potentiels et risques de l'IA[237], définit l'intelligence artificielle comme :

> La capacité d'un système à **interpréter** correctement des données externes, à **tirer des enseignements** de ces données et à **utiliser ces enseignements** pour atteindre des objectifs et des tâches spécifiques grâce à une **adaptation souple.** [Nos soulignements]

Si on se fie à cette définition, probablement que les systèmes experts[238], sous-discipline de l'IA symbolique, ne pourraient être considérés comme de l'intelligence artificielle, car ils ne peuvent s'adapter de la même façon que le font les modèles d'apprentissage automatique. La grande majorité des systèmes experts apprennent par des règles logiques structurées à partir de connaissances d'un expert et spécifiques à un domaine[239]. Ainsi, nous pensons que l'intelligence artificielle doit être vue plus largement et ne doit pas seulement inclure les méthodes d'apprentissage automatique[240]. Rappelons que l'intelligence artificielle symbolique a servi à jeter les bases théoriques de la discipline et que les notions de logique, rationalité et prise de décisions à l'origine de cette branche de recherche sont encore au cœur de la recherche en IA[241]. De plus, les systèmes experts sont souvent intégrés aux applications d'apprentissage automatique[242]. Finalement, la recherche combinant des méthodes d'apprentissage profond et symboliques a récemment gagné en popularité[243]. C'est pourquoi, aux fins du présent document de travail, l'intelligence artificielle symbolique sera considérée comme un moyen d'application de l'intelligence artificielle.

---

236 Andreas Kaplan & Michael Haenlein, « Siri, Siri, in my hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence » (2019) 62:1 15. Andréas Kaplan est enseignant-chercheur à *l'ESCP Business School,* il est spécialisé dans les réseaux sociaux, le marketing viral, le big data et l'intelligence artificielle.

237 *Ibid.,* à la p 15.

238 Voir infra Chapitre 2, Section 1, Sous-section « 2.3 Expert Systems ».

239 European Commission, *supra* note 235. Voir infra Chapitre 2, Section 1, Sous-section « 2.3 Expert Systems ».

240 *Ibid*.

241 *A definition of AI: Main Capabilities and Disciplines, High-Level Expert Group on Artificial Intelligence*, 2019 à la p. 1.

242 Jiayuan Mao et al, *supra* note 232.

243 Marta Garnelo, Kai Arulkumaran & Murray Shanahan, *supra* note 233; Vasant Honavar & Leonard Uhr, *supra* note 234; Lina, Lim Tong Ming & Leow Soo Kar, *supra* note 234.

1.2 Intelligence artificielle, apprentissage automatique et profond

La section précédente sur l'histoire de l'intelligence artificielle nous apprend que l'apprentissage automatique (*machine learning*) est aujourd'hui l'un des principaux moyens d'application de l'intelligence artificielle. Cette méthode d'IA regroupe l'ensemble de méthodes statistiques permettant aux machines et ordinateurs d'apprendre par eux-mêmes et d'optimiser leur performance sans en avoir été explicitement programmé[244].

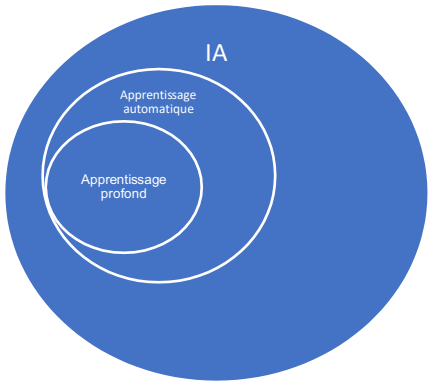
L'apprentissage profond (*deep learning*) est une branche de l'apprentissage automatique. Issu du courant connexionniste, l'apprentissage profond reproduit le fonctionnement du cerveau humain par « un réseau de neurones artificiels composé de plusieurs couches de neurones hiérarchisées selon le degré de complexité des concepts, et qui, en interagissant entre elles, permettent à un agent d'apprendre progressivement et efficacement à partir de mégadonnées »[245]. Particulièrement efficace pour la reconnaissance d'objets et d'images, l'apprentissage profond est au centre des avancées récentes en apprentissage automatique[246]. Par contre, comme nous le verrons au chapitre 2 du document de travail, l'apprentissage profond n'est pas la seule méthode d'apprentissage automatique. Ces deux termes ne sont donc pas synonymes. Même si l'apprentissage profond est parmi les méthodes les plus prometteuses, ce n'est qu'une application parmi d'autres de l'IA. Donc, l'apprentissage automatique est une sous-discipline de l'intelligence artificielle dont fait partie l'apprentissage profond. Ces trois concepts s'emboîtent, mais sont distincts [voir figure 3].

244 Jie Zhu, *Glossaire en intelligence artificielle*, 2020.

245 Office québécois de la langue française, « Apprentissage profond », en ligne : *Fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26532876>.

246 Voir supra Section 2, Sous-section « 3. Une fenêtre d'opportunité pour l'apprentissage automatique ».

Figure 3
À noter que l'apprentissage automatique inclut d'autres techniques, celles-ci seront présentées au chapitre 2 du document de travail.



1.3 Intelligence artificielle et algorithme

Les algorithmes soulèvent plusieurs questions éthiques et philosophiques[247]. Leur usage est donc au centre des conversations concernant l'encadrement de l'intelligence artificielle par le droit. L'algorithme est un concept mathématique qui a été développé bien avant l'avènement informatique. L'origine du mot « algorithme » remonte au 12e siècle et est issue de la transcription latine du nom du mathématicien arabe al-Kharezmi (783-850) et du grec arithmos, qui veut dire « nombre »[248]. Un algorithme désigne une suite finie et précise d'instructions permettant d'arriver à un résultat déterminé[249]. Pris au sens large, une recette de cuisine, un protocole expérimental et des consignes de sécurité peuvent être des algorithmes[250].

En programmation, un algorithme désigne une méthode de résolution d'un problème transcrite d'une manière non ambigüe et susceptible d'être codée sur un ordinateur et exécutée par un programme[251]. Pour reprendre les explications de la Commission nationale informatique et libertés (CNIL) :

> Pour qu'un algorithme puisse être mis en œuvre par un ordinateur, il faut qu'il soit exprimé dans un **langage** informatique, **transcrit en un programme** (une sorte de texte

247 *Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, Synthèse du débat publique animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, by CNIL, Synthèse du débat publique animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, 2017.

248 Rob Kitchin, « Thinking critically about and researching algorithms » (2016) 20 Information, Communication & Society 1-16 à la p 3; Jie Zhu, *supra* note 244.

249 *Ibid*. *CNIL, supra* notre 247, à la p 15; Jie Zhu, *supra* note 244.

250 CNIL, *supra* note 247; Jie Zhu, *supra* note 244.

251 *Ibid*.

> composé de commandes écrites, également appelé « code source »). Ce programme peut alors être **exécuté dans un logiciel** ou compilé sous la forme d'une application.[252] [Nos soulignements]

Autrement dit, un algorithme est une recette à suivre, étape par étape, rédigée par un programmeur pour permettre à un logiciel d'accomplir une tâche prédéfinie[253]. En intelligence artificielle, la « recette » est inspirée d'une méthode d'intelligence artificielle. Prenons par exemple les systèmes de détection de fraudes. Imaginons qu'un programmeur veut faire appel à des algorithmes d'apprentissage automatique afin de créer un logiciel capable de déceler des transactions et des mouvements de fonds irréguliers. Le programmeur a soumis des données à son logiciel de fraudes lui donnant des exemples de fraudes et de mouvements de fonds irréguliers, mais aucun programme n'indique au logiciel quoi faire de ces données. Pour donner des instructions à son logiciel, le programmeur doit créer un algorithme. Ainsi, à partir d'un langage de programmation, le programmeur fournit, étape par étape, toutes les instructions que le logiciel de fraude doit suivre pour savoir quoi faire avec ces données. Ces instructions sont l'algorithme. L'algorithme est transcrit dans un programme et celui-ci est exécuté dans le logiciel de fraude. À partir de l'algorithme, le logiciel de fraude peut apprendre des données afin d'éventuellement reconnaître les fraudes par lui-même. Un logiciel peut avoir recours à plus d'un algorithme[254]. Dans l'exemple que nous venons de donner, la saisie des données, soit les données donnant des exemples de fraudes soumises au logiciel, est faite à partir d'un algorithme.

Généralement, les algorithmes d'IA peuvent être séparés en deux grandes catégories : les algorithmes classiques et les algorithmes d'apprentissage automatique. Les algorithmes classiques, associés à l'intelligence artificielle symbolique sont déterministes, c'est-à-dire que « leurs critères de fonctionnement sont explicitement définis par ceux qui les mettent en œuvre »[255]. Les algorithmes associés aux techniques d'apprentissage automatique, comme l'exemple cité plus haut, sont autoapprenants et probabilistes. Ils apprennent en fonction des données qui leur sont fournies et évoluent au fur et à mesure de leur utilisation afin d'être en mesure d'automatiser certaines tâches[256]. Comme les méthodes d'apprentissage automatique sont maintenant les plus répandues dans le domaine de l'IA, les algorithmes d'apprentissage automatique et la discipline de l'intelligence artificielle sont souvent confondus. Cependant, le terme « intelligence artificielle » inclut plusieurs techniques algorithmiques, dont les algorithmiques classiques et d'apprentissage automatique. De plus, un algorithme peut être utilisé pour agir sur différents secteurs (l'éducation, les ressources humaines, la justice, la santé, etc.) et son élaboration peut servir plusieurs fonctions. Par exemple, générer des connaissances,

---

252 CNIL, *supra* note 247 à la p 15.

253 The Privacy Expert's Guide to Artificial Intelligence and Machine Learning, by Future of Privacy Forum, 2018 à la p 4.

254 CNIL, *supra* note 247 à la p 15. Cet exemple couvre seulement les méthodes d'apprentissage automatique, un programmeur pourrait également créer un système expert pour agir sur l'information.

255 *Ibid.,* à la p 18.

256 *Ibid*.

prédire des comportements ou évaluer un niveau de risques, faire des liens, recommander et aider dans la prise de décision[257].

Bref, il ne fait aucun doute que l'algorithme est au cœur de l'intelligence artificielle, car c'est ce qui permet de donner une forme aux méthodes d'IA utilisées. Connaître, même de façon générale, leur mise en œuvre est essentielle pour mieux saisir les différents enjeux que leur développement suscite.

1.4 Intelligence artificielle, algorithmes d'apprentissage automatique et mégadonnées

Intelligence artificielle et algorithmes d'apprentissage automatique renvoient nécessairement à la notion de données et plus largement à celle de mégadonnées (*big data).* Si les techniques d'apprentissage automatique semblent si prometteuses ces dernières années, c'est principalement attribuable à la puissance de calcul et de mémoire des ordinateurs, à la disponibilité d'une grande quantité de données et au développement d'algorithmes plus performants[258]. Plus il y a de données disponibles, plus un algorithme d'apprentissage automatique peut « apprendre ». Cependant, ces données sont si volumineuses qu'il devient difficile de les traiter et les stocker à l'aide d'outils d'analyse classiques[259]. Ainsi, l'appellation *Big data* fait référence à :

> Un ensemble d'une très grande quantité de données, structurées ou non, se présentant sous différents formats et en provenance de sources multiples, qui sont collectées, stockées, traitées et analysées dans de courts délais, et qui sont impossibles à gérer avec des outils classiques de gestion de bases de données ou de gestion de l'information.[260]

Le concept de mégadonnées est aussi défini par la formule des quatre V : Volume, Variété, Vélocité et Véracité[261].

---

257 *Ibid.,* à la p 22. **Voir** le tableau de la CNIL à ce propos.

258 Conseil de l'Europe, *supra* note 97.

259 J Zhu, *supra* note 244.

260 Office québécois de la langue française, « Mégadonnées », en ligne : *fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/Resultat.aspx>.

261 Doug Laney, *3D Data Management : Controlling Data Volume, Velocity and Variety* , Meta Group Inc. (2001); Bastien L, « Les quatre V du Big Data expliqués par IBM », (2016), en ligne : *LeBigData.fr* <https://www.lebigdata.fr/infographie-quatre-v-big-data-expliques-ibm>; Jie Zhu, *supra* note 244; In Lee, « Big data: Dimensions, evolution, impacts, and challenges » (2017) 60:3 Business Horizons 293-303.

Tableau 1
Définition de mégadonnées par la formule des 4V

Volume	Variété	Vélocité	Véracité
Quantité de données qu'une organisation ou un individu recueille/ génère.	Type de données collectées. Cela fait référence à la variété de formats : par exemple, du texte, des images, vidéos, etc. Mais aussi de domaines : par exemple, la santé, les activités sur les réseaux sociaux, les habitudes de consommations, etc.	Vitesse à laquelle les données sont générées et traitées.	Le traitement de ces données est conditionnel à une vérification de leur crédibilité et validité. En effet, les données ont un caractère subjectif, que les techniques statistiques peuvent ne pas reconnaître.

Certaines entreprises ajoutent à ces 4V les dimensions de variabilité (les flux de données sont imprévisibles et dépendent d'évènements ou de modes difficiles à suivre et à gérer sans les équipements informatiques adéquats)[262] et de valeur[263] (la capacité de transformer ces données en quelque chose qui a de la valeur pour l'industrie).

À la lumière de ces définitions, l'on comprend que le terme « mégadonnées » réfère non seulement à un très grand volume de données, mais aussi à l'ensemble des techniques et technologies qui permettent de les traiter, les trier et les analyser afin d'en faire ressortir leur valeur et les faire « parler »[264]. Donc, ces concepts sont interdépendants car la qualité des données, la façon dont elles sont introduites dans un logiciel et la façon dont le logiciel est « formé » à l'analyse des données aura un impact direct sur la validité, l'exactitude et l'utilité des informations générées par un algorithme[265].

1.5 Intelligence artificielle, robotique

Bien qu'elle y contribue en partie, la robotique ou la science des robots est une discipline qui est souvent confondue avec l'intelligence artificielle. Décrite comme « l'IA en action dans le monde physique »[266], la robotique est une application possible de l'intelligence artificielle. En effet, des algorithmes d'IA peuvent être intégrés à une machine (le robot) afin de la rendre capable d'exécuter de manière autonome une ou plusieurs tâches dans des environnements

---

262 SAS, « Big Data: What it is and why it matters », en ligne : <https://www.sas.com/en_ca/insights/big-data/what-is-big-data.html>.

263 In Lee, « Big data », *supra* note 261 à la p 294.

264 CNIL, *supra* note 247 à la p 18.

265 Pedro Domingos, « A few useful things to know about machine learning » (2012) 55:10 *Commun ACM* 78-87.

266 A definition of AI: Main Capabilities and Disciplines, *supra* note 241 à la p 4.

spécifiques[267]. Par exemple, les voitures autonomes, les drones, les aspirateurs et tondeuses à gazon, les robots chirurgicaux, etc.[268]. La robotique sert d'enveloppe physique à certains systèmes d'intelligence artificielle. Par contre, d'autres disciplines jouent un rôle dans la conception des robots. La conception et la mise au point des robots abordent un ensemble de problèmes mécaniques, électroniques et techniques qui ne relèvent pas des intérêts de l'intelligence artificielle[269]. Bref, un robot peut être programmé à partir de techniques d'intelligence artificielle et la robotique est influencée par les développements en IA, mais la robotique est aussi une discipline dont certaines techniques sont à l'extérieur du domaine de l'intelligence artificielle. Ainsi, l'intelligence artificielle dépasse l'image du robot à l'apparence humaine. En fait, la plupart des logiciels d'IA sont intégrés à notre quotidien de manière beaucoup plus subtile, par exemple, par des applications (Uber, Find my Friend, etc.) ou par des systèmes de recommandations (Netflix, Spotify, etc.).

1.6 Intelligence artificielle étroite, générale et super intelligente

Malgré les progrès technologiques des dernières années et les avancées considérables dans plusieurs domaines d'IA, l'intelligence artificielle ne peut se comparer à notre intelligence. L'intelligence artificielle, comme on la connaît aujourd'hui, est étroite (*narrow AI*) et consiste à accomplir, avec précision, des tâches spécifiques et prédéfinies, ce qui veut dire que pour pouvoir évoluer et être appliquée dans un nouveau contexte, une IA doit être programmée par des êtres humains[270]. Autrement dit, les systèmes d'IA sont performants que pour accomplir les tâches pour lesquelles ils ont été programmés. Par exemple, filtrer des pourriels, reconnaître votre visage pour ouvrir votre téléphone, vous recommander un film à écouter, vous donner l'état du trafic, assister un médecin lors d'une chirurgie, etc. Ceux-ci ne peuvent s'émanciper de cette tâche, car ils manquent de sens commun[271]. Même les méthodes les plus performantes d'IA (apprentissage automatique et profond) ne dépassent pas ce niveau étroit[272]. En conséquence, l'intelligence artificielle générale qui comprend et fonctionne comme un être humain et évolue par elle-même, n'est encore qu'une possibilité théorique[273]. La capacité de généraliser des connaissances ou des compétences, de les appliquer dans un autre contexte ou encore d'évoluer

---

267 Office québecois de la langue française, « Robotique », en ligne : *Fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=2078479>.

268 *A definition of AI: Main Capabilities and Disciplines*, *supra* note 241.

269 Jie Zhu, *supra* note 244.

270 Future of Privacy Forum, *supra,* note 253 à la p 6.

271 Naveen Joshi, « 7 Types Of Artificial Intelligence » *Forbes* (19 juin 2019), en ligne : <https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/>; Future of Privacy Forum, *supra* note 253; Luc Julia, *supra* note 2, à la p 29.

272 *Ibid*.

273 Future of Privacy Forum, *supra* note 253 à la p 5. Les chercheurs ne savent pas encore s'il est possible d'arriver à un stade d'IA général à partir des techniques d'IA étroite. Certains pensent que l'IA générale sera possible à partir de techniques de programmation plus sophistiquées qui n'ont pas encore été découvertes. Voir Chapitre 2, Section 2 « Vers une intelligence artificielle (IA) forte ? ».

par soi-même sont des capacités qui ne peuvent être réalisées que par des humains[274]. Donc, la superintelligence artificielle qui suppose qu'un système d'IA pourrait reproduire les multiples facettes de l'intelligence humaine, mais aussi les dépasser en ayant une meilleure mémoire, une plus grande vitesse d'analyse et une meilleure prise de décision et poser une menace à notre survie est encore loin d'arriver et n'existera peut-être jamais[275]. Cela ne veut pas dire que l'intelligence artificielle n'est pas utile. Au contraire, ces modèles du monde simplifiés servent à rendre notre vie meilleure. Par contre, il est faux de penser que ces systèmes puissent, dans un futur proche, nous remplacer ou encore se retourner contre nous[276].

Bref, ce premier aperçu de la discipline permet de préciser ce que l'on entend par intelligence artificielle. En effet, l'IA ne correspond pas à un ensemble de concepts rigides et homogènes. Ceux-ci sont plutôt changeants et évoluent au gré des progrès technologiques. Ces progrès, à leur tour, influencent les performances des méthodes d'IA. En revanche, pour mieux comprendre la discipline et les différents enjeux qui y sont liés, il est important de cerner en quoi ces différents concepts sont liés, mais pas synonymes. De plus, il faut rappeler que malgré ce qui est parfois partagé dans les médias ou par des personnalités publiques[277], l'IA est encore loin de pouvoir nous égaler et doit plutôt être vue comme un outil servant à rendre notre quotidien plus simple et efficace. Finalement, afin de compléter cette vue d'ensemble de l'IA, la prochaine sous-section aura pour objectif de présenter une taxonomie de l'IA proposée par la Commission européenne[278] et nous permettra d'identifier les principales fonctionnalités de l'IA, ainsi que ces sous-domaines.

---

274 *Ibid.*

275 Luc Julia, *supra* note 2. La singularité technologique sera discutée à la fin du Chapitre 2, Section 2, Sous-section « 7. De la singularité technologique ».

276 Luc Julia & Ondine Khayat, *L'intelligence artificielle n'existe pas*, first forum éd, 2019; Naveen Joshi, *supra* note 271.

277 Rory Cellan-Jones, « Stephen Hawking warns artificial intelligence could end mankind », *BBC News* (2 décembre 2014), en ligne : <https://www.bbc.com/news/technology-30290540>; Sam Shead, « Elon Musk says DeepMind is his "top concern" when it comes to A.I. », (29 juillet 2020), en ligne : *CNBC* <https://www.cnbc.com/2020/07/29/elon-musk-deepmind-ai.html>.

278 European Commission, *supra* note 235.

Pour aller plus loin :

CNIL, *Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, 2017, en ligne : <https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf>

Future of Privacy Forum, *The Privacy Expert's Guide To Artificial Intelligence and Machine Learning*, 2018, en ligne : <https://iapp.org/resources/article/the-privacy-experts-guide-to-ai-and-machine-learning/>

« Grand lexique français », en ligne : *Data franca* <https://datafranca.org/wiki/index.php?title=Cat%C3%A9gorie:GRAND_LEXIQUE_FRAN%C3%87AIS&pagefrom=I>.

Jie Zhu, *Glossaire en intelligence artificielle*, 2020, en ligne : <https://cyberjustice.openum.ca/glossaire-ia-laboratoire-de-cyberjustice-2020/>

Naveen Joshi, « 7 Types Of Artificial Intelligence » *Forbes* (19 juin 2019), en ligne : <https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/>.

« Vocabulaire de l'intelligence artificielle (liste de termes, expressions et définitions adoptés) », en ligne : *Légifrance* <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037783813?init=true&page=1&query=Intelligence+artificielle&searchField=ALL&tab_selection=all>.

2. Une taxonomie de l'IA

En 2020, *AI Watch*, le service de connaissances de la Commission européenne chargé de surveiller le développement, l'adoption et l'impact de l'intelligence artificielle en Europe, a publié un document de recherche ayant pour objectif « d'établir une définition opérationnelle de l'IA formée d'une taxonomie concise et d'un ensemble de mots-clés qui caractérisent les domaines fondamentaux et transversaux de l'IA »[279]. Dans le cadre du présent document de travail, cette taxonomie nous permettra de couvrir et classer le paysage de la discipline afin de proposer une définition d'intelligence artificielle.

La taxonomie présentée par *AI Watch* s'appuie sur 55 définitions développées entre 1955 et 2019 représentant trois perspectives cibles : le secteur politique et institutionnel, le domaine de la recherche et celui de l'industrie. À partir de ces définitions, il a été possible de déterminer un ensemble de mots-clés qui caractérisent l'ensemble de domaines et sous-domaines qui représentent l'intelligence artificielle. Les domaines de l'IA ont été divisés en deux grandes catégories : les domaines au cœur de la discipline (1) et les compétences transversales (2). Ces domaines sont ensuite associés à des sous-domaines (voir Tableau 2) Finalement, même s'ils sont décrits de manière distincte, il est à noter que ces domaines, en pratique, sont souvent liés et peuvent s'entrecouper.

279 *Ibid.,* à la p 6.

Tableau 2
Taxonomie de l'IA présentée par *AI Watch*[280]

Table 1. AI domains and subdomains constituting one part of the operational definition of AI

	AI taxonomy	
	AI domain	**AI subdomain**
Core	**Reasoning**	Knowledge representation
		Automated reasoning
		Common sense reasoning
	Planning	Planning and Scheduling
		Searching
		Optimisation
	Learning	Machine learning
	Communication	Natural language processing
	Perception	Computer vision
		Audio processing
Transversal	**Integration and Interaction**	Multi-agent systems
		Robotics and Automation
		Connected and Automated vehicles
	Services	AI Services
	Ethics and Philosophy	AI Ethics
		Philosophy of AI

Source: Authors' elaboration

2.1 Les domaines au cœur de la discipline de l'IA

Le raisonnement correspond à la manière dont les systèmes d'IA sont en mesure de transformer les données en connaissances (représentation des connaissances) pour en déduire des faits et raisonner à partir de règles symboliques[281].

La planification automatisée concerne la conception et l'exécution de stratégies afin qu'un système d'IA puisse comprendre l'espace et s'y déplacer. Par exemple, ce domaine fait référence aux voitures autonomes ou aux robots aspirateurs qui doivent être conçu afin d'être en mesure d'optimiser leurs réactions et agir adéquatement dans un espace multidimensionnel[282].

L'apprentissage réfère aux techniques d'apprentissage automatiques qui mettent en place des systèmes ayant la capacité d'apprendre, prédire, prendre des décisions, s'adapter et s'améliorer par expérience[283].

La communication est la capacité pour un système d'IA, par le traitement du langage naturel (NLP), d'identifier, comprendre et traiter les informations écrites ou vocales provenant

280 *Ibid.,* à la p 11.

281 *Ibid.,* à la p 12.

282 *Ibid*.

283 *Ibid*.

d'individus. Par exemple, la traduction automatique de textes, la génération de textes pour répondre à nos courriels, etc.[284].

La perception s'intéresse à la capacité des systèmes d'IA à prendre conscience de leur environnement à travers les sens. Actuellement, ce sont les domaines de l'audition et de la vision qui sont les plus développés. Par exemple, la vision par ordinateur qui inclut les techniques de reconnaissance faciale et d'objets[285].

2.2 Les compétences transversales

Les compétences transversales ne sont pas spécifiquement liées à une discipline de recherche, mais représentent des considérations dont il faut tenir compte lorsqu'on s'intéresse aux domaines mentionnés plus haut[286].

L'intégration et l'interaction sont la combinaison entre la perception, le raisonnement, l'action, l'apprentissage, l'interaction avec l'environnement et d'autres notions comme la coordination, l'autonomie et la coopération afin de créer des outils qui servent à assister ou se substituer aux humains. Cette compétence peut donc inclure d'autres disciplines comme la robotique. Par exemple, ces outils peuvent être des drones, des robots médicaux, un système d'aide à la conduite, etc.[287].

Les services d'IA sont les plateformes ou infrastructures logicielles qui offrent un service ou une application qui permettent de rendre plus efficace la gestion d'infrastructures complexes. Par exemple, les services infonuagiques (*cloud*), l'accès à des services de stockages, des assistants virtuels, etc.[288].

L'éthique et la philosophie sont de plus en plus au cœur des initiatives politiques concernant l'encadrement de l'intelligence artificielle. En effet, déjà bien intégrée à notre quotidien, notre confiance en ces systèmes dépend de plus en plus du respect de principes et valeurs éthiques dans leur conception et leur utilisation[289].

À partir de cette présentation générale de l'intelligence artificielle, il est maintenant plus aisé de proposer une définition de la discipline. Cette sous-section donne un aperçu des principales branches théoriques de l'IA et des domaines de recherche qui y touchent et qui doivent être considérés en parallèle de la discipline. Comme il est difficile de trouver une définition unifiée de l'intelligence artificielle, cette taxonomie de l'IA sert à décrire la discipline dans son ensemble et

---

284 *Ibid*.

285 *Ibid.,* à la p 13.

286 *Ibid.,* à la p 11.

287 *Ibid.,* à la p 13.

288 *Ibid*.

289 *Ibid*.

permet de cibler les éléments importants à être intégrés dans une définition de l'intelligence artificielle.

La prochaine sous-section aura donc pour objectif de proposer une définition de l'intelligence artificielle. Pour ce faire, il sera d'abord question de présenter en quoi définir l'IA en l'opposant à l'intelligence humaine est une approche trompeuse. Il s'agira ensuite de faire une revue de définitions de l'intelligence artificielle afin de présenter quelles caractéristiques, en cohérence avec les concepts et sous-disciplines présentés ci-dessus, servent à définir l'intelligence artificielle.

> Pour aller plus loin :
>
> Commission Européenne, AI watch: defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence, 2020, en ligne : <https://publications.jrc.ec.europa.eu/repository/handle/JRC118163>

3. Définir l'IA en l'opposant à l'intelligence naturelle : une approche trompeuse

La définition classique de l'intelligence artificielle ou, plutôt, du « problème de l'intelligence artificielle », est attribuée à John McCarthy, Marvin L. Minsky, Nathaniel Rochester et Claude E. Shannon. Dans leur désormais célèbre article, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*[290], les auteurs décrivent l'intelligence artificielle comme suit : « making a machine behave in ways that would be called intelligent if a human were so behaving »[291]. Cette définition, laquelle renvoie à un test conçu par Alan Turing en 1950[292] pour établir si les machines peuvent penser, semble avoir modelé les définitions actuelles qui présentent l'intelligence artificielle comme une « une branche de l'informatique traitant de la **simulation** du comportement intelligent dans les ordinateurs » [nos soulignements][293] ou « la capacité d'une machine à **imiter** le comportement humain intelligent » [nos soulignements][294]. Pourtant, ces définitions restent peu concluantes, car elles ont tendance à présenter l'IA comme un concept vague. Après tout, selon la logique proposée, un outil relativement simple tel qu'une calculatrice pourrait être considéré comme une IA. De plus, ces définitions semblent proposer une cible idéale à atteindre : réussir à reproduire l'intelligence humaine, plutôt qu'un concept de recherche mesurable[295].

---

290 John McCarthy et al, *supra* note 103.

291 Des parties de cet article ont depuis été réimprimées dans (2006) AI Magazine 27(4) : 12-14.

292 Alan Turing, *supra,* note 99.

293 Merriam-Webster Dictionary, « Definition of ARTIFICIAL INTELLIGENCE », en ligne : <https://www.merriam-webster.com/dictionary/artificial+intelligence>.

294 *Ibid.*

295 European Commission, *supra* note 235 à la p 7.

D'ailleurs, comme il en a été question dans la précédente section sur l'histoire de l'intelligence artificielle[296], l'hypothèse de départ du projet de recherche sur l'IA proposée par McCarthy et ses pairs, où il était question de pouvoir, « décrire avec précision chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence afin qu'une machine puisse réussir à simuler des comportements intelligents »[297] a rapidement été réfutée et cela a donné lieu au premier hiver de l'IA. Il est donc important de rappeler que l'intelligence artificielle s'est développée au gré de périodes d'optimisme et de désillusions où les promesses ambitieuses concernant le développement de la discipline ont la plupart du temps donné lieu à des coupures dans la recherche plutôt qu'à de réelles avancées. Ainsi, il est étonnant que l'on parle toujours d' « intelligence » artificielle alors que l'histoire nous apprend que l'intelligence naturelle, appliquée sur des machines, ne peut être réduite à une vision computationnelle où toute cognition peut se traduire par des règles ou des calculs[298]. Pourtant, l'intelligence artificielle — qu'elle soit abordée à travers une perspective politique, de recherche ou de l'industrie — est encore largement définie à partir de cette première définition de l'IA. La prochaine sous-section aura donc pour objectif de présenter en quoi cette comparaison peut être trompeuse et génère des discussions autour d'un concept abstrait et mal défini.

3.1 Les mystères de l'intelligence naturelle et sa simplification excessive

L'intelligence naturelle demeure peu comprise et est difficile à mesurer[299]. Comme l'explique Paul Dumouchel, « l'intelligence semble entrer dans la catégorie des choses dont nous disons : je peux la reconnaître quand je la vois, mais je ne pourrais pas la définir »[300]. Ce caractère abstrait de l'intelligence fait en sorte que certains attributs de notre intelligence (comme le raisonnement ou la mémoire) sont souvent évalués en corrélation avec ce que nous jugeons important pour une vie réussie (comme la réussite économique ou scolaire)[301]. Par exemple, évaluer notre intelligence à travers un test de QI censé nous donner, à travers un chiffre, une garantie (ou pas) de notre intelligence et notre capacité à évoluer en société. Cependant, selon Paul Dumouchel, cette idée de pouvoir évaluer l'intelligence par des tests universels donne faussement l'idée que l'intelligence est un concept simple et unifié qui peut s'évaluer indépendamment des individus[302]. Ainsi, il n'est pas étonnant que l'intelligence artificielle, portée depuis ses débuts par l'idée qu'il soit possible de « simuler » ou « imiter » l'intelligence humaine, soit considérée comme quelque

---

296 Voir Section 2 « Les développements en intelligence artificielle : une histoire marquée par l'opposition », Sous-section « 1.4 Des promesses exagérées ».

297 John McCarthy et al, *supra* note 103 à la p 2.

298 Luc Julia, *supra,* note 2 à la p 9.

299 Grands Dossiers, « Intelligence : de quoi parle-t-on ? » *Sciences Humaines* (2007), en ligne : <https://www.scienceshumaines.com/intelligence-de-quoi-parle-t-on_fr_21032.html>.

300 Paul Dumouchel, *supra* note 35.

301 *Ibid*.

302 *Ibid.,* à la p 242; Howard Gardner, *Frames of mind: the theory of multiple intelligences*, Basic Books éd., New York, 1983.

chose qui peut exister en soi et parvenir à égaler l'espèce humaine. Pourtant, cette comparaison est traître, car elle suggère que l'intelligence (un concept subjectif et abstrait) peut s'évaluer de la même façon sur chaque individu et sur deux porteurs différents : les êtres humains et les systèmes artificiels[303]. Cette comparaison est malavisée, car l'intelligence humaine ne peut être définie ou mesurée dans sa totalité et est plutôt constituée d'un certain nombre de compétences connexes. L'IA doit plutôt être considérée comme un outil, qui peut dépasser les capacités humaines dans certains cas, mais pas dans d'autres[304].

En effet, les systèmes d'intelligence artificielle sont créés pour exécuter certaines de nos habiletés mieux que nous et réussissent la majorité du temps à bien le faire.[305] Cependant, ces systèmes ne peuvent s'émanciper de leurs tâches, car ils manquent de sens commun et ne sont pas capables de se représenter leur savoir[306]. Au contraire, un individu est impliqué dans les connaissances qu'il produit. Pour prendre une décision, même dans sa forme la plus élémentaire comme choisir comment s'habiller, un individu s'appuie sur ses connaissances, sa volonté, son interprétation du contexte, son humeur, etc., alors qu'un système d'IA ne fait qu'exécuter sans avoir la capacité de comprendre ses choix[307].

En effet, selon Paul Dumouchel, la notion d'incarnation (*embodiement*) est, en grande partie, ce qui distingue l'intelligence humaine de l'IA[308]. L'intelligence humaine est extrinsèquement liée à un individu, manifestée dans un corps. La façon dont nous raisonnons sur les objets est donc également conditionnée par cette manifestation physique[309]. Une poignée de porte, par exemple, ne peut être considérée sans notre capacité à l'ouvrir avec notre main. De plus, une poignée de porte prend une autre signification lorsque nous voulons passer la porte. Tout cela influence la façon dont nous reconnaissons, percevons et choisissons d'agir sur les objets du monde[310]. Les systèmes artificiels, quant à eux, ne sont pas incarnés mais plutôt des systèmes mathématiques résidant dans un ordinateur. Ils n'ont pas de point de vue individuel relatif à un corps, ni les mêmes outils ou désirs que nous[311]. Par exemple, en 2016, Microsoft a développé un agent conversationnel baptisé « Tay », lequel avait pour fonction de converser avec des êtres humains via Twitter. À l'intérieur d'une journée, les Tweets de l'agent étaient devenus racistes parce que des trolls d'Internet[312] avaient bombardé Tay de données offensives et erronées (des

---

303 Paul Dumouchel, *supra* note 35 à la p 242.

304 Voir *infra*, Chapitre 2 « Common Sense, Causality and Embodiment ».

305 CNIL*, supra* note 247 à la p 28.

306 Paul Dumouchel, *supra,* note 35 à la p 250.

307 *Ibid*.

308 *Ibid*.

309 *Ibid*., voir *infra* Chapitre 2, « Common sense, causality, embodiement ».

310 *Ibid.,* aux pp 241-258.

311 *Ibid*.

312 Office québecois de la langue française, « Troll d'Internet », en ligne : *Fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26522696>.

gazouillis incendiaires) afin d'influencer ses propres gazouillis[313]. Notons ici que Tay ne souffrait d'aucun défaut au sens commun du terme. L'algorithme a fonctionné parfaitement. Par contre, Tay n'avait pas les connaissances, ni le sens commun, ni la compréhension du contexte pour saisir que ce qu'il partageait était inapproprié, il ne faisait qu'accomplir ce pourquoi il avait été programmé en s'appuyant sur les données (les gazouillis) qui lui étaient soumises. Ainsi, les systèmes d'IA ne peuvent être considérés comme des individus à part entière ni comme ayant une personnalité qui leur soit propre, car il est impossible pour ces systèmes de comprendre et d'interagir avec le monde comme nous le faisons[314].

Bref, les systèmes d'IA sont des agents immatériels, incarnés dans quelque chose (un robot ou un système) pour interagir avec le monde matériel. Ils n'existent que de façon limitée, à l'aide d'un ensemble de technologies qui leur permettent de comprendre **certains** aspects du monde matériel[315]..

Finalement, même les tâches que peuvent accomplir les systèmes d'IA ne peuvent être considérées comme de l'intelligence naturelle. Par exemple, un système d'IA entraîné à reconnaître une image de chat aura besoin d'environ 100 000 images de chat pour pouvoir atteindre un score de 98 %[316]. Un humain n'aura en général besoin que de deux ou trois photos pour faire le même exercice[317]. Bref, l'intelligence naturelle et l'intelligence artificielle n'ont pas grand-chose à voir. Par cette affirmation, nous ne nions pas que l'intelligence artificielle s'inspire de certaines capacités cognitives humaines pour les appliquer à des machines. Cependant, ces comparaisons sont locales et partielles et portent sur des aspects spécifiques[318]. Il n'existe donc pas d'échelle universelle pour comparer l'intelligence artificielle à l'intelligence naturelle[319].

3.2 Une approche trompeuse

Adopter une définition qui compare l'intelligence artificielle à l'intelligence naturelle peut être trompeur, car comme nous venons de l'expliquer, ces deux idées peuvent, dans les faits, difficilement être comparées. De plus, cette comparaison réfère implicitement à la question de savoir si l'IA pourra un jour nous dépasser. Pourtant, se concentrer sur cette question alors que c'est encore une hypothèse lointaine nous détourne des vraies questions concernant l'intelligence artificielle, soit la place et le rôle que les différentes technologies d'IA devraient

313 Dave Lee, « Tay: Microsoft issues apology over racist chatbot fiasco », *BBC News* (25 mars 2016), en ligne : <https://www.bbc.com/news/technology-35902104>.

314 Paul Dumouchel, *supra* note 35 à la p 249.

315 *Ibid.,* à la p 250.

316 Luc Julia, *supra* note 276 à la p 9.

317 *Ibid*.

318 Paul Dumouchel, *supra* note 35.

319 *Ibid.,* à la p 244.

avoir dans notre société[320]. Se représenter l'intelligence artificielle comme une discipline qui pourra servir à nous remplacer au travail ou représenter une menace pour l'humanité par ses capacités super intelligentes, crée non seulement de faux problèmes, mais génère des attentes qui tiennent plus de promesses abstraites que de théories réalistes. Ces promesses servent ensuite à justifier l'élaboration de politiques publiques sur des notions encore lointaines et justifier des investissements à partir d'hypothèses plus ou moins vérifiées[321]. Abandonner le fantasme de pouvoir mesurer dans son ensemble la « force » ou la « puissance » de l'intelligence artificielle par rapport à l'intelligence humaine permet de ramener le débat aux applications réelles de l'intelligence artificielle[322]. D'un point de vue juridique, cela a une influence sur notre façon d'encadrer l'IA et ses usages. En effet, en s'éloignant de cette comparaison, on s'éloigne aussi de l'idée que l'IA est une discipline qui nous glisse tranquillement entre les doigts. Ces systèmes sont mis en place pour des raisons particulières (générer des sous-titres dans une vidéo, analyser des images, reconnaître des objets, recommander de la musique, etc.) et sont le reflet des choix (et des biais) qui ont été faits lors de leur conception. Ce sont ces choix qui influencent le pouvoir que peuvent avoir ces systèmes sur les individus et non l'IA en elle-même[323]. Nous avons le contrôle de décider de ce qui doit et devrait être fait avec l'IA, ce qui veut dire que nous sommes aussi responsables de ses mauvais usages[324]. Bref, définir l'intelligence artificielle pour ce qu'elle est et non pour ce qu'elle pourrait faire est essentiel et a un impact direct sur notre façon de l'encadrer et de comprendre son rôle dans la société.

3.3 Intelligence augmentée plutôt qu'artificielle

Cette confusion autour de la comparaison entre intelligence naturelle et artificielle est pour certains attribuable au terme lui-même d'« intelligence artificielle ». Comme nous venons de le présenter, utiliser ce terme a une connotation particulière qui participe à alimenter le mythe du robot superpuissant qui se retourne contre son créateur. Cette image nous éloigne de ce que les systèmes d'IA sont réellement capables d'accomplir et alimente le flou autour de la discipline. C'est pourquoi Luc Julia propose que l'expression intelligence augmentée soit préférée[325]. C'est

---

320 *Ibid*.

321 Yves Gingras, « L'intelligence artificielle n'existe pas », (3 juin 2018), en ligne : *Radio-Canada* <http://ici.radio-canada.ca/premiere/emissions/les-annees-lumiere/segments/chronique/74731/science-critique-gingras-intelligence-artificielle-n-existe-pas>; Yves Gingras & Jonathan Roberge, *Ateliers SociologIA : conférences disponibles en ligne*, 2019, en ligne : <https://www.cirst.uqam.ca/nouvelles/2019/ateliers-sociologia-conferences-disponibles-en-ligne/>; Julia, *supra* note 2. À ce sujet, voir chapitre 3, sous-section « 3.2. Regard critique sur les politiques de l'intelligence artificielle (IA) ».

322 *Ibid*. Paul Dumouchel, *supra* note 35 à la p 246.

323 Paul Dumouchel, *supra* note 35 à la p 250.

324 Luc Julia, *supra* note 2.

325 Luc Julia & Ondine Khayat, *supra* note 276; Daniel Faggella, « What is Artificial Intelligence? An Informed Definition », en ligne : *Emerj* <https://emerj.com/ai-glossary-terms/what-is-artificial-intelligence-an-informed-definition/>. Emerj définit l'IA comme un continuum: l'intelligence assistée, augmentée et autonome.

encore de l'intelligence artificielle, mais présentée autrement. Il s'agit ici de présenter l'IA pour ce qu'elle est capable de faire : augmenter notre propre intelligence et nous aider à mieux faire ce que nous faisons[326]. Que ce soit un programme autopilote qui assiste notre conduite ou une simple recherche sur un moteur de recherche, ces systèmes servent d'outils. Ils nous assistent au quotidien, nous rendent plus efficaces, mais nécessitent notre intervention, car ils ne peuvent agir de manière complètement autonome. D'ailleurs, cette proposition a été adoptée dans un document de travail du *Joint Technology Committee*[327] (JTC) sur l'intelligence artificielle où le terme intelligence « augmentée » a été préféré. Considéré comme plus précis et acceptable pour décrire la discipline, le JTC a indiqué que « pour les tribunaux, l'IA est de l'intelligence augmentée, car c'est l'utilisation de la technologie pour faire ce que les humains font, mais plus vite et mieux »[328].

Bref, l'utilisation du terme « intelligence artificielle » a une connotation très forte qui suggère que quelque chose d'important et de plus grand que nature est possible, alors que ce n'est pas encore le cas et que ça ne le sera probablement jamais. Interpréter l'intelligence artificielle comme de l'intelligence augmentée est une proposition qui doit être adoptée plus largement dans nos travaux, car c'est une terminologie plus nuancée et précise pour se référer à la discipline.

Pour aller plus loin :

Dumouchel, P. « Intelligence, Artificial and Otherwise » (2019) 24:2 Forum Phisosophicum 241

Gingras Y., L'intelligence sociologique confrontée à l' « intelligence artificielle » *Ateliers SociologIA*, 2019, En ligne : https://www.cirst.uqam.ca/nouvelles/2019/ateliers-sociologia-conferences-disponibles-en-ligne/

Julia, L. et K. Ondine, *L'intelligence artificielle n'existe pas*, First Forum éd, 2019.

4. L'apport des institutions, de la recherche et de l'industrie sur les définitions de l'IA

Pour parvenir à proposer une définition de l'intelligence artificielle (ou augmentée), nous avons recensé des définitions de l'IA provenant de trois grands secteurs qui touchent la discipline[329] :

---

326 Luc Julia & Ondine Khayat, *supra* note 276.

327 Voir NCSC, « Joint Technology Committee », (20 mai 2020), en ligne : <https://www.ncsc.org/about-us/committees/joint-technology-committee>.

328 *Introduction to AI for Courts*, by Joint Technology Committee (JTC), JTC Resource Bulletin version 1.0, 2020 à la p 1.

329 Voir Maxime Colleret et Yves Gingras, *L'intelligence artificielle au Québec : un réseau tricoté serré*, note de recherche 2020-07, Montréal, Centre interuniversitaire de recherche sur la science et la technologie (CIRST), 2020, en ligne : <cirst2.openum.ca/files/sites/179/2020/12/Note_2020-07_IA.pdf>. Les frontières entre ces secteurs sont parfois très minces. Au Québec, plusieurs acteurs du domaine de l'IA évoluent dans plusieurs secteurs à la fois. Voir aussi Chapitre 3, Sous-section « 3.1. Perspective locale : la politique de l'intelligence artificielle (IA) au Québec) ».

1. Secteur politique et institutionnel
2. Les secteurs académiques et de la recherche
3. Secteur de l'industrie

4.1 Observations préliminaires

Il est important de préciser qu'en raison des points soulevés précédemment, nous avons fait le choix conscient de présenter des définitions qui faisaient le moins possible de comparaisons entre l'intelligence artificielle et l'intelligence naturelle. De plus, les définitions sélectionnées ne donnent qu'un bref aperçu des nombreuses définitions adoptées par les différents documents de travail/ de recherche qui concernent l'intelligence artificielle.

4.2 Secteur politique et institutionnel

Le secteur politique et institutionnel fait référence aux organisations internationales et aux États chargés de mettre en place des stratégies nationales sur les usages de l'intelligence artificielle ou son développement[330]. Dans ce contexte, l'IA est souvent définie sous deux angles : soit comme un instrument de croissance et de développement dans lequel on doit investir ou comme un phénomène technologique qui doit être régulé à partir de principes éthiques[331] ou philosophiques.

Tableau 3
Revue des définitions du secteur politique et institutionnel

Nom de l'institution	Source	Définition
Commission européenne	HLEG, 2019[332]	Les systèmes d'intelligence artificielle (IA) sont des **systèmes logiciels** (et **éventuellement matériels**) conçus **par des êtres humains** et qui, ayant **reçu un objectif** complexe, agissent dans le **monde réel ou numérique** en **percevant leur environnement** par **l'acquisition de données**, en interprétant les données structurées ou non structurées collectées, en appliquant un **raisonnement aux connaissances**, ou en **traitant les informations**, dérivées de ces données et en décidant de **la/des meilleure(s) action(s) à prendre pour atteindre l'objectif donné**. Les systèmes d'IA peuvent soit utiliser des **règles symboliques, soit apprendre un modèle numérique**. Ils peuvent également **adapter leur comportement** en analysant la manière dont l'environnement est affecté par leurs actions antérieures. [Nos soulignements]

330 European Commission Joint Research Centre, AI Watch, Historical Evolution of Artificial Intelligence: Analysis of the Three Main Paradigm Shifts in AI., LU, Publications Office, 2020 à la p 7.

331 Ad Hoc Expert Group for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence, *Document final: avant-projet de Recommandation sur l'éthique de l'intelligence artificielle*, 2020.

332 A definition of AI: Main Capabilities and Disciplines, *supra* note 241 à la p 8.

Nom de l'institution	Source	Définition
OCDE	Recommandation du Conseil sur l'intelligence artificielle[333]	Un système d'intelligence artificielle (ou système d'IA) est un système automatisé qui, pour un ensemble donné **d'objectifs définis par l'homme**, est en mesure d'établir des **prévisions, de formuler des recommandations, ou de prendre des décisions influant sur des environnements réels ou virtuels**. Les systèmes d'IA sont conçus pour fonctionner à des **degrés d'autonomie divers**. [Nos soulignements]
UNESCO	Avant-projet de Recommandation sur l'éthique de l'intelligence artificielle[334]	La présente recommandation ne cherche pas à donner de définition unique de l'IA, celle-ci étant appelée à évoluer en fonction des progrès technologiques. Son objectif est plutôt de **traiter des caractéristiques des systèmes d'IA** qui revêtent une importance majeure sur le **plan éthique** et font l'objet d'un vaste consensus international. En conséquence, la présente Recommandation envisage les systèmes d'IA comme des systèmes technologiques capables de traiter **l'information par** un processus s'apparentant à un comportement **intelligent**, et comportant généralement des fonctions de raisonnement**, d'apprentissage, de perception, d'anticipation, de planification ou de contrôle**. [Nos soulignements]
DARPA	Site internet de la DARPA[335]	Artificial intelligence is a programmed ability to **process information**. This is evaluated according to the system ability to **perceive rich, complex and subtle information**, **learn within an environment**, abstract to **create new meaning, reason to plan and decide**. [Nos soulignements]

4.3 Les secteurs académiques et de la recherche

Afin de décrire les définitions provenant du secteur de la recherche, il semble pertinent de reprendre ce qui a été présenté par Peter Norvig et Stuart Russell dans leur livre phare *Artificial Intelligence. A Modern Approach*[336]*.* Les chercheurs y présentent huit définitions autour de quatre approches de l'IA et celles-ci offrent un portrait des différentes approches et méthodes mises de l'avant pour définir l'IA comme discipline de recherche.

---

333 OCDE, *Recommandation of the Council on Artificial Intelligence*, 2019, en ligne : <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

334 Ad Hoc Expert Group for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence, *supra* note 331 à la p 6.

335 John Launchbury, A DARPA Perspective on Artificial Intelligence, 2020.

336 Stuart J Russell & Peter Norvig, *supra,* note 98

Tableau 4
Présentation de huit définitions autour de quatre approches de l'IA par Stuart Russell et Peter Norvig[337]

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Figure 1.1 Some definitions of artificial intelligence, organized into four categories.

> Les définitions du haut concernent les processus de pensée et le raisonnement, tandis que celles du bas concernent le comportement. Les définitions de gauche mesurent le succès en termes de fidélité à la performance humaine, tandis que celles de droite mesurent par rapport à une mesure de performance idéale, appelée rationalité. Un système est rationnel s'il fait « ce qu'il faut », compte tenu de ce qu'il sait. Cela nous donne quatre objectifs possibles à poursuivre en intelligence artificielle.[338]

Plus récemment, les auteurs Darrell M. West et John R. Allen de *Brookings Institution* ont défini l'intelligence artificielle autour de trois qualités de l'IA : l'intentionnalité, l'intelligence et l'adaptabilité :

- **L'intentionnalité** : Les systèmes d'IA sont conçus pour prendre des décisions à partir de données. Ce sont des systèmes capables de fournir des réponses mécaniques et prédéterminées. Les progrès technologiques des dernières années les rendent particulièrement performants pour l'analyse et la prise de décisions.

- **L'intelligence** : Pour être performant, un système d'IA dépend surtout des données qui lui sont soumises, particulièrement pour les méthodes d'apprentissage automatique.

- **L'adaptabilité** : L'IA peut apprendre à partir de leur « expérience » et s'adapter[339].

---

337 *Ibid.,* à la p 5.

338 *Ibid*.

339 Darrell M West and John R Allen, « How artificial intelligence is transforming the world », (24 avril 2018), en ligne : <https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/>.

Ici, l'accent est mis sur les méthodes d'apprentissage automatique qui domine le domaine de la recherche par leurs progrès récents en matière d'analyse de données et de prise de décision rapide.

Dans ce secteur, l'IA est donc souvent définie comme un concept de recherche qui tend vers un objectif précis (simuler l'intelligence naturelle, raisonner correctement, etc.) et propose des définitions plus larges que les définitions qui proviennent du secteur politique et institutionnel.

4.4 Le secteur de l'industrie

Le secteur de l'industrie en matière d'IA est florissant et offre toute sorte de services (outils de gestion de documents, services infonuagique, serveurs, relecture de contrats, etc.) Dans ce secteur, l'intelligence artificielle est présentée sous l'angle du développement, de sa valeur économique et des perspectives de marchés futurs[340]. De plus, les définitions proposées par les acteurs de l'industrie proviennent souvent ou sont présentées sous forme de blogue[341] et mettent généralement l'accent sur les domaines les plus prometteurs de la discipline.

Tableau 5
Revue des définitions du secteur de l'industrie

Industrie	Source	Définition
SAS : programme qui permet l'analyse de données dans plusieurs secteurs de l'industrie, dont la santé, les banques et le commerce de détail.[342]	Site internet du SAS.[343]	L'intelligence artificielle (IA) permet à des machines **d'apprendre par l'expérience**, de **s'adapter à de nouvelles données** et de **réaliser des tâches humaines**. La plupart des exemples d'IA qui font les gros titres de nos jours (des ordinateurs jouant aux échecs aux voitures autonomes) reposent fortement sur le deep learning et le traitement du langage naturel. Grâce à ces technologies, il est possible de former des ordinateurs à **effectuer certaines tâches** en **traitant de vastes quantités de données** et en en **dégageant des tendances**. [Nos soulignements]
Oracle offre des systèmes de gestion de données et des plateformes infonuagiques pour les entreprises.[344]	Pour définir l'intelligence artificielle, Oracle renvoie à un article	L'Intelligence artificielle (IA) est la **simulation de processus d'intelligence humaine par des machines**. Plus particulièrement, des systèmes informatiques. Ces processus comprennent trois phases. Tout d'abord, l'**apprentissage**, c'est-à-dire **l'acquisition de l'information** et ses règles d'utilisation. Puis, le **raisonnement**, soit l'utilisation de règles pour tirer des conclusions approximatives ou définitives. Et enfin, **l'autocorrection**. Les applications

340 European Commission, *supra* note 235 à la p 7.

341 Donc, une lecture rapide, efficace et visuellement attrayante.

342 SAS, « SAS : Analytique, Intelligence artificielle et Data Management », en ligne : <https://www.sas.com/fr_ca/home.html>.

343 SAS, « Intelligence artificielle - Présentation et intérêt », en ligne : <https://www.sas.com/fr_ca/insights/analytics/what-is-artificial-intelligence.html>.

344 Oracle, « Dynamisez vos activités grâce à l'intelligence artificielle », en ligne : <https://www.oracle.com/ca-fr/artificial-intelligence/>.

Industrie	Source	Définition
	de blogue *d'actualité informatique*.[345]	particulières de l'IA comprennent la narrow AI, la reconnaissance faciale et la vision par ordinateur. [Nos soulignements]
Emerj aide les organisations mondiales à développer des stratégies et des initiatives en IA.[346]	Présenté sous forme de blogue sur le site Emerj, le responsable de la recherche présente une définition de l'IA à partir d'une revue de littérature.[347]	L'intelligence artificielle est une **entité** (ou un **ensemble collectif** d'entités coopératives), capable de recevoir des contributions de l'**environnement**, d'**interpréter** et d'**apprendre** de ces entrées, et de présenter des **comportements et des actions connexes et flexibles** qui aident l'entité à atteindre un **but ou un objectif** particulier sur une période de temps donnée. [nos soulignements]

Bien que certaines de ces définitions se recoupent, cet exercice de recensement par secteur permet de souligner le fait que l'intelligence artificielle est un terme utilisé dans plusieurs contextes et pour représenter différents intérêts. Pourtant, à force de vouloir la définir, la notion même d'intelligence artificielle est galvaudée à un tel point qu'elle semble couvrir à la fois tout et rien et ne représente finalement plus grand-chose. C'est pourquoi, il est important de proposer une définition qui représente l'état actuel de la recherche en IA. Pour ce faire, il convient de cibler les définitions qui s'accordent le plus avec la taxonomie présentée à la sous-section 3.2.

4.5 Définir l'intelligence artificielle

À partir des définitions recensées, des fonctionnalités présentées précédemment et des éléments mis de l'avant dans le document de travail d'AI Watch : *Defining Artificial Intelligence : Towards an Operational Definition and Taxonomy of Artificial Intelligence*[348], les caractéristiques suivantes nous semblent être les éléments les plus pertinents à intégrer à une définition de l'intelligence artificielle :

- la réalisation d'une tâche spécifique ou d'un objectif prédéfini;
- la perception de l'environnement;
- le traitement de l'information par la collecte et l'interprétation des données à travers une intervention humaine;
- la prise de décision à partir de méthode de raisonnement logique ou d'apprentissage automatique qui permet aux systèmes d'agir, d'apprendre ou de raisonner;
- et comment ces systèmes peuvent être incarnés dans le monde matériel.

---

345 « Qu'est-ce que l'Intelligence artificielle (IA) ? », (2020), en ligne : *Actualité Informatique* <https://actualiteinformatique.fr/intelligence-artificielle/qu-est-ce-que-intelligence-artificielle-ia>.

346 Emerj, « Emerj AI Opportunity Landscape Service », en ligne : <https://emerj.com/ai-opportunity-landscape-inquiry/>.

347 Daniella Faggella, *supra* note 325.

348 European Commission, *supra* note 235 à la p 8.

Ces caractéristiques, en plus condensées, s'accordent avec la taxonomie et les concepts présentés à la sous-section 3.2. Une définition de l'IA qui réussit à intégrer ces caractéristiques couvre les principales fonctionnalités associées au paysage de l'intelligence artificielle et nous rapproche d'une définition précise de la discipline. De plus, ces caractéristiques demeurent assez générales pour capturer l'évolution rapide de l'intelligence artificielle. Finalement, il est important de mentionner que même si nous proposons une définition de l'intelligence artificielle, celle-ci n'est pas figée dans le temps et doit être revue et confirmée au fil des progrès technologiques afin de répondre à notre objectif de précision. Bref, nous pensons que la définition du Groupe d'experts en intelligence artificielle[349] est celle qui définit avec le plus de précision l'intelligence artificielle :

> Les systèmes d'intelligence artificielle (IA) sont des **systèmes logiciels** (et **éventuellement matériels**) conçus **par des êtres humains** et qui, ayant **reçu un objectif** complexe, agissent dans le **monde réel ou numérique** en **percevant leur environnement** par **l'acquisition de données**, en interprétant les données structurées ou non structurées collectées, en appliquant un **raisonnement aux connaissances**, ou en **traitant les informations**, dérivées de ces données et en décidant de **la/des meilleure(s) action(s) à prendre pour atteindre l'objectif donné**. Les systèmes d'IA peuvent soit utiliser des **règles symboliques, soit apprendre un modèle numérique**. Ils peuvent également **adapter leur comportement** en analysant la manière dont l'environnement est affecté par leurs actions antérieures.[350] [nos soulignements]

Cette définition est parmi les plus complètes, car non seulement elle est une des rares définitions à intégrer toutes les caractéristiques mentionnées ci-dessus, mais elle reflète aussi les éléments discutés dans la présente section. De plus, le Groupe d'experts en intelligence artificielle comprend des représentants du monde universitaire, de la société civile et de l'industrie[351], ce qui fait en sorte que cette définition a été proposée en prenant en compte les différents points de vue de ces acteurs.

En ce qui a trait à la définition elle-même, parler de l'intelligence artificielle comme d'un système logiciel permet de nous éloigner de l'idée d'un robot aux apparences humaines et de nous rapprocher de l'IA telle qu'elle est : un ensemble de systèmes ayant été programmé pour accomplir une tâche prédéfinie. Ensuite, aucune comparaison n'est faite avec l'intelligence naturelle et il est mentionné que ces systèmes sont conçus par des êtres humains. Cette mention est importante, car elle évite tout malentendu : l'intelligence artificielle demeure sous notre contrôle. Finalement, il est précisé ce que peuvent faire ces systèmes (percevoir, raisonner, traiter de l'information, adapter leur comportement) et où ils peuvent agir (monde réel ou

---

349 A Definition of AI: Main Capabilities and Disciplines, *supra,* note 241.

350 *Ibid.,* à la p 6.

351 Commission européenne, « High-Level Expert Group on Artificial Intelligence », (2018), en ligne : *Shaping Europe's digital future - European Commission* <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>.

numérique). À partir de cette définition, il est possible de savoir ce qu'est capable de faire l'intelligence artificielle et comment elle peut être intégrée à notre quotidien.

Il est vrai que cette définition peut paraître longue et technique. Cependant, si l'on souhaite discuter de l'intelligence artificielle et pouvoir évaluer son impact sur la société, prévoir ses développements et justifier son investissement, il faut être en mesure de savoir de quoi l'on parle et cela est directement lié au choix de définition que l'on décide de mettre de l'avant. C'est le point de départ pour une meilleure compréhension commune de l'IA.

Pour aller plus loin :

European Commission, AI watch: defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence, 2020, en ligne : <https://publications.jrc.ec.europa.eu/repository/handle/JRC118163>.

Gingras Y., L'intelligence sociologique confrontée à l'« intelligence artificielle » *Ateliers SociologIA*, 2019, En ligne : <https://www.cirst.uqam.ca/nouvelles/2019/ateliers-sociologia-conferences-disponibles-en-ligne>.

High-Level Expert Group on Artificial Intelligence, *A definition of AI: Main Capabilities and Disciplines*, 2019, en ligne : <https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>.

Pagel, J.F. et P. Kirshtein, *Machine Dreaming and Consciousness*, Academic Press, Elsevier, 2017

Russell, Stuart J. Rus et P. Norvig, *Artificial intelligence: a modern approach*, Englewood Cliffs (N.J), Prentice Hall, 1995.

4.6 Conclusion

En conclusion, définir l'intelligence artificielle n'est pas une tâche facile. La définir dépend des points de vue — par exemple, déterminer si l'approche symbolique doit en être exclue ou encore, s'il est nécessaire de la comparer à l'intelligence naturelle — et dépend des objectifs des entités qui décident de la définir. Cela fait en sorte que la notion même d'intelligence artificielle est trop souvent imprécise. Afin d'être sensible aux impacts de l'intelligence artificielle sur la société et le droit, il faut d'abord parvenir à une compréhension minimale de la discipline. Pour ce faire, il est nécessaire de proposer une définition qui emploie des termes précis, même si ceux-ci sont techniques, afin de présenter l'IA telle qu'elle est. Cette approche sert de point de départ afin d'être en mesure d'évaluer les différents enjeux qui concernent l'utilisation de l'IA, mais aussi ses opportunités. À partir de cette compréhension générale de la discipline, il est ensuite possible de lier la discipline à des compétences transversales. Notamment, les notions d'éthique, mégadonnées, robotique, Internet des objets, etc.

Conclusion

Le premier chapitre de ce document de travail avait pour objectif de présenter l'objet de l'IA comme discipline (le « quoi »). À la section 1, nous avons vu les doctrines et les disciplines qui constituent le fondement même de son existence. La philosophie a donné à l'IA son raisonnement. La biologie a fourni, à son tour, le schéma d'organisation. La linguistique a partagé avec l'IA comment articuler et partager l'information. Les mathématiques ont offert un moyen de reconstituer et traduire les divers états et opérations mentaux. L'ingénierie informatique, quant à elle, a fourni la technologie nécessaire pour construire l'IA elle-même. Enfin, les médias ont capté la curiosité du public.

À la section 2, nous avons abordé les éléments marquants de l'histoire et du développement de l'intelligence artificielle. Pour ce faire, nous avons présenté en quoi l'opposition entre le courant symbolique et connexionniste a servi à façonner et structurer le domaine de la recherche en IA. Des balbutiements de l'informatique aux systèmes experts, l'IA symbolique a participé à jeter les bases de la discipline telles que la pose de diagnostics et la prise de décision. De son côté, l'approche connexionniste, longtemps reléguée à l'arrière-plan, a trouvé enfin un (re)gain de popularité en 2010 dans l'apprentissage profond. Notamment, grâce aux progrès des infrastructures informatiques. Marqué par deux hivers dans le développement de l'IA, cette section a aussi été l'occasion de discuter des incidences des promesses gonflées par les chercheurs sur le développement de la discipline. Aujourd'hui, les deux approches — symbolique et connexionniste — tendent à se rejoindre dans l'implémentation d'algorithmes hybrides pour en optimiser la performance.

À la section 3, il a été question de proposer une définition de l'intelligence artificielle. Nous avons d'abord présenté en quoi l'intelligence artificielle est différente des concepts connexes auxquels elle est souvent associée. En effet, depuis la première apparition du terme « intelligence artificielle », plusieurs définitions de l'IA ont été suggérées par les secteurs politiques et institutionnels, de la recherche et de l'industrie, ce qui participe à créer de la confusion dans notre compréhension de la discipline. Il nous semblait donc important de faire une revue de ce que l'IA n'est pas et de présenter une taxonomie de la discipline. Finalement, avant de nous arrêter sur la définition proposée par le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle, nous avons suggéré que l'expression « intelligence augmentée » devrait être préférée. L'expression « intelligence artificielle » peut être trompeuse, car elle sous-tend l'idée d'une super-intelligence (forte) capable de surpasser l'intelligence humaine et nous éloigne de ce que l'IA est réellement : un outil informatique qui permet d'augmenter notre intelligence en optimisant (en temps, en précisions et en possibilités) le traitement de l'information et l'exécution de certaines tâches.

CHAPTER 2
POTENTIALS A0ND LIMITATIONS OF ARTIFICIAL INTELLIGENCE (AI) METHODOLOGIES AND APPLICATIONS

Section 1 - From Symbolic Systems to Machine Learning: A Task-Oriented Approach

Quick facts

Most artificial intelligence today is narrow artificial intelligence, focused on a certain task rather than attempting to recreate human intelligence.

The symbolic approach and expert systems were popular in the 1980s. They rely on encoded rules based on explicit knowledge. This works well for some problems but is difficult to apply to many real-world problems.

Machine learning relies on algorithms to automatically build models of correlations in data. Deep learning, a subfield of machine learning, works very well on many tasks and has rekindled the interest in artificial intelligence.

As we have previously described, it is difficult to find a unified definition of artificial intelligence. Describing the entire field of artificial intelligence is therefore a difficult task, and a lot of effort could be spent debating whether a certain technological method can be seen as artificial intelligence, or which specific subclass of artificial intelligence a method belongs to. Instead, this chapter will take a more practical approach. We aim to provide the reader with a high-level understanding of the broad ideas that underpin different artificial intelligence systems. It seems that there are a few large "schools" of artificial intelligence methods. While these could be overlapping in terms of methodologies and timeframes, they will here be treated as distinct ideas for simplicity. Further, the focus of the description is not to provide an in-depth explanation of the algorithms involved, but rather to provide the reader with practical, relevant examples of such systems, which will hopefully provide the reader with an intuition about the functioning and application areas of the methods.

This description is not exhaustive. With tens of thousands of papers being released in artificial intelligence each year, making an in-depth description of the different technologies being developed is beyond the scope of this report. Rather, it focuses on a few of the tasks with the most focus in the community, and the large schools of thought for methodologies.

There are multiple angles that could be used to classify artificial intelligence systems. In this work, we have chosen to rely on two such angles, namely that of tasks and technologies.

By classifying the artificial intelligence system by the tasks they perform, we see the systems as solutions to specific problems. Advances in artificial intelligence can serve to develop better solutions to problems, or even to solve entirely new problems. Looking at the systems in terms of tasks therefore allows us to understand which specific problems we can tackle using artificial

intelligence, and how the performance on these tasks has improved. This way of classifying artificial intelligence will be explored in section 2 of this chapter.

However, artificial intelligence systems can also be classified by technology. An interesting feature of artificial intelligence methods is that the same technology can often be used to tackle multiple problems and solve multiple tasks. The neural network, for example, has been compared to a universal software, that can in theory emulate any computer program.[352] We have chosen to focus on two schools of thought within the field of artificial intelligence, namely Symbolic systems and machine learning. While portrayed as separate here, they are often overlapping in time and in terms of influential ideas. We have aimed to describe the areas as accurately as possible, and to quickly elaborate on history, technologies used, main application areas and criticisms levelled against the approaches.

1. Tasks

Quick facts

Most artificial intelligence today is narrow artificial intelligence, focused on a certain task rather than attempting to recreate human intelligence.

Some tasks are easy for computers to accomplish, while others have only recently gotten feasible to attempt with AI methods. Some tasks are impossible, for AI systems or even for both humans and AI systems.

Different AI methodologies might be appropriate for different tasks.

1.1 Introduction

This section describes some common tasks that artificial intelligence has been applied to. This presentation has been chosen to facilitate looking at artificial intelligence systems as tools, rather than "intelligences". The first section will describe some background and information about tools in general. The second section will present a non-exhaustive list of possible tasks, including the input data and the desired outcome.

1.2 General Artificial Intelligence vs Narrow Artificial Intelligence

First of all, a distinction has to be made between general artificial intelligence and narrow artificial intelligence. Artificial General Intelligence (AGI) is the kind of AI that is able to rival human intelligence. Many researchers believe that we are not close to this level of artificial intelligence, and that the path to arrive at this class of AI is not evident. While AI is very strong

352 Andrej Karpathy, "Software 2.0", (13 June 2018), online: *Medium* <https://medium.com/@karpathy/software-2-0-a64152b37c35> at 2.

and able to rival humans in certain tasks, it does not share our ability to effortlessly adapt to new tasks.[353]

Further, the concept of AGI is somewhat vague from the start. There are many philosophical questions which surround the concept. For example, is consciousness required for an AI to be seen as general? Another question is how to determine whether an AI system should be seen as AGI. A number of tests have been suggested, including the Turing test, which suggests that a judge should communicate with either an AI or a human via a textual interface, and determine whether they are talking to a computer or not. If the judge fails to distinguish the two, this can be seen as a testament that the AI can pass as a human, and is thus general. Another example of a test is the coffee-cup test, which requires an AI controlled robot to walk into an unknown house and brew a cup of coffee.[354]

Recently, there has been a lot of development in large-scale language models that are pre-trained on massive corpora of texts to build a model of language, and are then able to write seemingly coherent texts based on a short prompt. The most recent such model, GPT-3, has been proven to be able to perform a wide variety of different tasks based on textual prompts.[355] There has been a large discussion on whether this constitutes a step towards AGI.

In general, however, the focus of the past few years has been on the development of so-called Narrow Artificial Intelligence. These systems focus on solving a specific task.[356] While not as general as AGI, this kind of systems still holds a tremendous potential in affecting society, and are starting to be used across a large range of industries and institutions.

1.3 What are tasks?

Due to the focus on the development of narrow artificial intelligence, it is important to understand what kind of tasks can be solved using the AI systems. Here are a few problems that have been attempted to solve by artificial intelligence systems :

- Proving logical theorems
- Recognizing numbers in images
- Playing chess
- Recognizing hate speech
- Recognizing objects in images and their location

---

353 Future of Privacy Forum, *supra* note 253 at 5.

354 *Ibid*.

355 Tom B Brown et al, "Language Models are Few-Shot Learners" (2020), online: <http://arxiv.org/abs/2005.14165>; Gwern Branwen, "GPT-3 Creative Fiction" (2020), online: <https://www.gwern.net/GPT-3>.

356 Future of privacy forum, *supra* note 253 at 6.

The artificial intelligence systems created here are usually heavily constrained by their tasks. Mastery of one task usually does not imply success at another task. Of course, the methods might be re-used – if a method works well for recognizing dogs, it might also be trained to recognize cats. However, a machine able to play chess well is not necessarily able to play another game, such as Go. Further, it gives no indication whatsoever over whether the system will be good at riding a bike.

Below, we discuss a few tasks that can be considered to require "intelligence" in computers to solve. While some of these can be solved best using modern deep learning algorithms, others might be solved more efficiently using expert systems or other more traditional methods (we will discuss these below). To tackle the tasks in the best possible way, it can therefore be important to have several possible approaches in mind. While some today see the term artificial intelligence to exclusively refer to the newer methods of machine learning and deep learning, historically it has encompassed a much wider variety of methods, many of whom are still useful and in use today.

1.4 Easy vs AI-complete tasks

Of course, not all tasks are equally challenging. Some tasks, such as adding two numbers together, can be seen as intelligence as they involve some computation. However, they are readily solvable using simple algorithms. Other tasks, such as recognizing objects in an image or spoken voice, are much more challenging, and often require methods such as machine learning to solve properly.

In fact, there is a discussion of whether it is possible to perfectly solve some tasks at all without general artificial intelligence, that is equal to human intelligence. Such problems are referred to as "AI-complete"[357]. The argument goes that we rely on our model of the entire world when, e.g., translating a sentence from one language into another. The intended meaning of a word might only be clear by having experience around the object or feeling the word refers to, and thus be impossible to translate for anything other than a human. Other tasks speculated to be AI-complete include question answering, common sense planning and image understanding.[358] The question of whether some tasks are AI-complete, and to which extent, is still open. However, in recent years many AI systems have achieved scores in limited domains that might be considered AI-complete by some that equal the scores of humans, such as machine translation[359].

---

357 Roman V Yampolskiy, "Turing Test as a Defining Feature of AI-Completeness" in Xin-She Yang, ed, *Artificial Intelligence, Evolutionary Computing and Metaheuristics* Studies in Computational Intelligence (Berlin, Heidelberg: Springer Berlin Heidelberg, 2013) 3 at 4.

358 Roman Yampolskiy, "AI-Complete, AI-Hard, or AI-Easy: Classification of Problems in Artificial Intelligence" (2012) The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA at 8.

359 Stuart J Russell & Peter Norvig, *Artificial intelligence: a modern approach*, fourth edition ed, Pearson series in artificial intelligence (Hoboken: Pearson, 2021) at 29.

Of course, artificial intelligence is not magic. There can be tasks that are simply impossible to solve using the data used to attempt to tackle it. For example, predicting the weather in a week by looking at the sky in one location would likely be impossible for both humans and artificial systems. The signal contained to devise the answer is simply not contained in the data fed to the system, or the correlation behind the data might be too complex to understand.

1.5 Examples of tasks

To get an idea of the types of tasks that can be accomplished with artificial intelligence, below you will find a non-exhaustive list of tasks that could potentially be solved with artificial intelligence.

1.5.1 General Prediction

- How much will the house sell for, based on the properties of the house?
- How much will a stock be worth tomorrow, based on historical price movements?
- Is a person susceptible to a certain disease, based on their age, blood values and diet?
- Is a certain event, measured by sensors, unusual?

1.5.2 Recommendations

- Which movies might a user be interested in, based on their viewing history?
- Is a certain email likely to be spam, based on the sender and content of the email?

1.5.3 Computer Vision

- Which objects are contained in an image?
- What is the location of a certain object in an image?
- How can the events in a video be described in human language?

1.5.4 Natural Language Processing

- Is a sentence of positive or negative sentiment?
- What does a certain word mean in a sentence?
- What does a sentence say translated into another language?
- What does a person say, based on a recording of their voice?

– What is the likely next word in a sentence?

1.5.5 Robotics

– How should a car react to an object seen by a camera?

– How should a robot dog move the motors to walk?

1.5.6 Playing games

– What is the most promising next move in chess or Go?

– What should the computer do to obtain a high score in Super Mario or Space Invaders?

– What action should the computer take to beat a human user in a strategy game, such as StarCraft II?

1.6 Conclusion

When talking about artificial intelligence, it is important to note that almost all of AI research today focuses on so-called narrow artificial intelligence, that is focused on solving specific tasks. Some such tasks might be possible to solve with regular algorithms, while others might require more advanced machine learning algorithms, or might currently be out of scope for machine learning. Next, let us take a look at the technologies that have historically been used to tackle artificial intelligence tasks.

2. Symbolic systems

2.1 Introduction

Quick facts

- Symbolic systems were an early type of AI that relied on the logical manipulation of symbols, i.e. concepts represented in a computer system.
- The systems usually work by providing the algorithm with explicit knowledge encoded as symbols, and rules about how these can be manipulated, to achieve goals.
- Expert systems, a later iteration of symbolic systems, shifted the focus to expert knowledge encoded in rules in computer systems. These created a large industry and are still in use in several domains.
- Finally, the hopes of creating fully intelligent systems using the symbolic approach fell short, partially due to the large investment required to create the system, difficulties of representing knowledge as symbols and effort require to integrate the systems into workflows.

This section will describe a class of systems known as symbolic artificial intelligence, or Good Old-Fashioned Artificial Intelligence (GOFAI)[360]. These systems were an early type of artificial intelligence, designed in the 1950s. They relied on reasoning with symbols, i.e. manifestations of ideas and objects, as well as their relationships, encoded into computer systems. The systems are then able to reason with the symbols to arrive at conclusions. At the time, these approaches were seen as very promising, with many researchers believing that the systems would rival human intelligence in just a few years' time[361]. However, these expectations did not pan out, as a number of shortcomings of the approach became obvious[362]. While today, the method has fallen somewhat out of favor, in comparison to sub-symbolic approaches (such as deep learning), it remains in use through expert systems. Recently, research combining new neural methods and the symbolic systems have gained popularity[363].

We will take a look at both symbolic systems and expert systems, which can be seen as an extension of the symbolic systems. It should be noted that this section does not aim to give an exhaustive picture of the methodologies involved in the symbolic approach. The field was a huge research undertaking, involving hundreds of researchers and many different sophisticated methodologies. Instead, we aim to capture some of the ideas involved and give a few examples of how such systems could look. Hopefully, these will serve as an illustration of the ideas, their advantages and shortcomings for the reader.

2.2 Symbolic systems

2.2.1 Introduction

The birthplace of artificial intelligence, and the symbolic approach, is often said to be a two-month workshop organized at Dartmouth college, in the summer of 1956. The basic idea of the conference was the idea that intelligence can be described precisely enough that it can be simulated using machines[364]. The conference brought together many of the major figures of the early research and established the field of artificial intelligence as a separate discipline from the rest of computer science[365].

---

360 John Haugeland, *supra* note 93.

361 Stuart J Russell, Peter Norvig & Ernest Davis, *Artificial intelligence: a modern approach*, 3rd ed, Prentice Hall series in artificial intelligence (Upper Saddle River: Prentice Hall, 2010) at 21.

362 James Lighthill, *supra* note 147; Hubert L Dreyfus, "Alchemy and Artificial Intelligence" (1964), online: <https://www.rand.org/pubs/papers/P3244.html>.

363 Marta Garnelo & Murray Shanahan, "Reconciling deep learning with symbolic artificial intelligence: representing objects and relations" (2019) 29 Current Opinion in Behavioral Sciences 17–23; Jude W Shavlik, "Combining symbolic and neural learning" (1994) 14:3 Mach Learn 321–331.

364 Stuart J Russell, Peter Norvig & Ernest Davis, *supra,* note 361 at 17.

365 *Ibid.,* at 18.

The research quickly made what seemed like huge amounts of progress. Several systems were built that were able to, among other things, solve logical theorems and play board games such as checkers, or solve specific problems that require intelligence[366].

These early successes lead to a huge enthusiasm, and expectations that the systems would advance to tackle more difficult, real-world problems in the very near future. However, soon the field realized that the solutions often failed to scale to the more complex real-world context. This led to a significant dampening of the enthusiasm[367].

2.2.2 Fundamental concepts

Symbolic systems are based upon the idea of reasoning with symbols. This can be traced back to Aristotle and Descartes, who discussed the separation between the body and the mind. Early AI research borrows many ideas from this view and aims to recreate the world as reconstructed from sensory inputs inside the mind, divorced from the physical world[368]. The theory is that the mind works by somehow representing concepts in the real worlds, including objects, plans, decisions, hopes etc., and manipulating these representations through logical processes. The mind is therefore seen as a computer[369].

Several of the researchers in the early days of artificial intelligence believed that, in fact, human intelligence can be *fully* described as symbols manipulated as physical patterns by the brain[370]. Allen Newell and Herbert A. Simon postulated the Physical Symbol System Hypothesis:

A physical symbol system has the necessary and sufficient means for general intelligent action[371].

This hypothesis implies that a system reasoning with symbol is sufficient to achieve general intelligent action, and further that it is the *only* way to achieve general action.

In order to recreate this idea of intelligence in computers, researchers began building systems that could take symbols (that can be seen as representing cognitive states)[372] and manipulating such symbols and their relationships, without necessarily being concerned with the object behind

366 *Ibid.,* at 18, 19.

367 *Ibid.,* at 21, 22; Stephen F Davis & William Buskist, *supra* note 19.

368 H R Ekbia, "Artificial Dreams" (2008) 418 at 22, 23.

369 *Ibid.,* at 23–26.

370 Paul Smolensky, "Connectionist AI, symbolic AI, and the brain" (1987) 1:2 Artificial Intelligence Review 95–109 at 98; H R Ekbia, *supra* note 368 at 29.

371 Allen Newell & Herbert A Simon, "Computer science as empirical inquiry: symbols and search" (1976) 19:3 Commun ACM 113–126 at 116.

372 H R Ekbia, *supra,* note 368 at 28.

the symbols[373]. By using a number of logical rules, systems can be built that are able to act on symbols, for example to achieve certain goals. McCarthy describes a hypothetical example of the advice taker, which knows symbols for a person, a car and an airport, and their relationship. By providing the system with the goal of going to the airport, the system would in theory figure out the steps the person would have to perform to get to the airport, including leaving the desk, going to the car and driving to the airport[374].

Since the symbols are distinct from their meaning, one could then translate many different real-world scenarios into symbols and use them same methods to achieve goals. Since all the required knowledge must be entered into the system in the forms of symbols and rules, this can be seen as a top-down approach - knowledge is encoded into the system explicitly, rather than learnt by the machine from the ground up[375].

A brief illustration of how logical reasoning can act on symbols shall illustrate this style of intelligence. One such rule, known as Modus Ponens, is such a rule that can be used to infer knowledge from premises[376] :

Premise 1: If it rains, the street is wet.

Premise 2: It rains.

Conclusion: The street is wet.

As you can see, the argument starts with two premises, that contain information about symbols (raining, the street being wet) and their relationship (rain leading to the street being wet). Using these premises, we can deductively arrive at the conclusion, in a way creating new knowledge (that the street is wet). Further, since the argument is acting on symbols, it can be abstracted to the following form[377] :

Premise 1: A -> B

Premise 2: A

Conclusion: B

---

373 Andre Vellino, “Artificial intelligence: The very idea" (1986) 29 Artificial Intelligence 349–353; John Haugeland, *supra* note 93; Paul Smolensky, *supra* note 370 at 98.

374 John McCarthy, “PROGRAMS WITH COMMON SENSE”, (1959) Paper 1-1, National Physical Laboratory, Teddington, Middlesex, at 8.

375 Stephen F Davis & William Buskist, *supra* note 19 at 487.

376 *Ibid*.

377 J Walmsley, *Mind and Machine* (Springer, 2016) at 33 Google-Books-ID: O3cYDAAAQBAJ.

We can substitute any symbols with the same relationships for A and B, and the reasoning with still hold true, based on the premises[378]. As such, a computer system developed to reason in this style, would theoretically be able to exhibit intelligence in any domain, encoded into these symbols. Winograd referred to this as developing a formalism that describes knowledge[379]. However, this only works as long as the knowledge can be expressed in terms of symbols with clear relationships, which might not always be the case[380]. Later, we will see how this approach contrasts with the sub-symbolic approach, where the machines are not provided with explicitly represented knowledge, but rather learn the representations themselves by looking at examples.

Modus Ponens is only an example of a method for reasoning with symbols. In reality, the systems developed used a wide variety of different methods, such as other predicate logic, fuzzy logic, semantic networks and Bayesian networks[381]. However, the reasoning with symbols seem to be at the basis of the approach.

2.2.3 The General Problem Solver – an example of a symbolic reasoning system

One of the most notable early symbolic artificial intelligence systems is called General Problem Solver (GPS), developed by Newell, Simon and Shaw. It expands on their previous work from the Dartmouth summer AI conference[382]. The researchers observed students and their process for solving problems, and then formalized this procedure into a computer system[383].

GPS takes problems in the form of objects and operators. Objects seem to be similar to symbols and can stand for anything from chess pieces to mathematical expressions. Operators transform these objects into other objects. This could be, for example, chess moves that move the pieces to other places on the board[384]. The user of the system then has to define a goal, for example to transform one type of object into another (e.g. transforming positions of pieces on a chess board into positions so that the opponent is checkmated)[385].

The system then tries to arrive at the goal, by devising subgoals that are easier than the overall goal and using various heuristics to devise and execute a plan to apply operators to the objects

---

378 *Ibid*; H R Ekbia, *supra* note 368 at 25.

379 Hubert L Dreyfus, "From micro-worlds to knowledge representation: AI at an impasse" (1981) Mind design 161–204 at 149.

380 Stephen F Davis & William Buskist, *supra* note 19 at 487.

381 *Ibid* at 487, 488.

382 "P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf", online: <http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf> at 2.

383 *Ibid.,* at 3.

384 *Ibid.,* at 3, 4.

385 *Ibid.,* at 6.

to arrive at the stated goal[386]. In doing so, it attempts to mimic the way a human would solve a problem to arrive at a goal[387].

The General Problem Solver is an impressive display of the ingeniousness of the early AI researchers. It is one of the first systems to separate the problem-solving method from the domain. GPS was one of the first system to separate the knowledge (symbols and their relationships etc.) from the problem-solving algorithm. Providing that a problem could be expressed in terms of symbols comprehensible to GPS, it should therefore in theory be applicable to all kinds of problems. However, the system also illustrates the issues that the symbolic approach would run into, that would prevent it from achieving its lofty ambitions. Notice the difficulty of describing certain kinds of tasks in terms of objects and operators – how do you describe a task such as riding a bike, or recognizing an image[388] ? Further, while the method should in theory be able to solve many complex problems (such as playing chess), in practice the enormous number of possibilities makes this approach unfeasible for this kind of problems. Beyond certain kind of simple problems, the GPS therefore faces tremendous problems, as did many symbolic AI systems.

2.2.4 Other examples

Besides the General Problem Solver, there were a wealth of different systems implemented in the early years of symbolic AI[389]. A field that received a lot of attention, and is well suited to analysis using symbolic systems, is the proving of mathematical theorems. The system is here tasked with starting with a set of axioms and generating a number of intermediary steps to arrive at a theorem, thereby proving it[390]. The Logic Theorist worked by starting from the problem and working backwards to find a connection between the axioms and the theorem[391]. Gelernter at IBM developed a system able to solve prove theorems in elementary Euclidian plane geometry[392]. The system could, for example, mathematically prove that segments of a triangle were equal based on the angles in the triangle[393].

There were also a number of programs that aimed to understand and utilize human language. An early example of such a system is ELIZA, developed by Joseph Weizenbaum.[394] ELIZA presented

386 *Ibid.,* at 8–24.

387 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 18.

388 John McCarthy, "Generality in artificial intelligence" (1987) 30:12 6 at 1031.

389 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 18.

390 Herbert Gelernter, *Realization of a geometry theorem proving machine.* (1959) at 136.

391 *Ibid*.

392 *Ibid.,* at 158.

393 *Ibid.,* at 147.

394 Joseph Weizenbaum, "ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine" (1966)9:16 Communications of the ACM 36-41.

itself as a conversational partner, that could be communicated with via a teletype, like a chatbot today. It scanned the text sent by the user and detected certain keywords. These words were then manipulated and reassembled into a response. If no keyword is found, the system would go back to an earlier prompt or post a generic response[395]. The system was built to mimic the way certain psychotherapists talk[396]. Despite the relatively simple rules, it was able to create quite complex and real-sounding conversations[397]. These might trick people into believing that they were interacting with a real sympathetic listener, while in reality they were interacting with a set of scripts aiming to identify certain words. Some researchers refer to this effect of overestimating the capability of artificial systems as the Eliza effect[398].

Other systems were focused on solving issues contained in so-called micro-worlds, e.g. small constrained coded environments. Terry Winograd, for example, developed a system able to reason about objects, and answer to questions in natural language, such as "Find a block which is taller than the one you are holding and put it in the box"[399]. While the "world" the system worked in was very limited and simplified, the idea was that solving them would nonetheless allow the modelling of real-world phenomena[400]. Dreyfus criticized this plan[401].

2.2.5 Discussion

This section will discuss some of the advantages and disadvantages associated with symbolic AI systems.

Symbolic AI systems seem relatively quick to get to work. Even with the meager computational resources of the sixties, a number of systems were developed that had results that seemed impressive. ELIZA, for example, was able to mimic human conversations to some extent. Several systems able to prove certain mathematical formulas were developed. These are certainly activities that would be considered as intelligence if they were performed by humans.

Despite these early impressive systems, the symbolic approach does not seem to have led to the heralded advances in artificial intelligence. Several reports, such as the Lighthill report[402], were released criticizing the approach, leading to an AI winter of decreased interested.

---

395 *Ibid.,* at 3–5.

396 *Ibid.,* at 6.

397 *Ibid.,* at 1, 2.

398 H R Ekbia, *supra* note 368 at 8.

399 Hubert L Dreyfus, *supra* note 379 at 144; Russell, Norvig & Davis, *supra* note 361 at 20.

400 Hubert L Dreyfus, "From micro-worlds to knowledge representation", *supra* note 379 at 146–148.

401 *Ibid.,* at 149, 150.

402 James Lighthill, *supra* note 147.

One issue with the symbolic approach was that the systems did not seem to be able to escape their created micro-worlds, to solve real-world problems. While they worked well in some very limited domains, applying them to the complexities of the real world seemed doomed to fail[403].

One of the reasons for this is the reliance of the systems on symbols. This works well for some high-level problems[404], such as chess pieces or mathematical formulas. These are fields where there are clearly defined rules for symbols, and what they can do. However, there are many things that we consider to be part of intelligence that is not as easily defined in terms of symbols[405]. Which symbols, for example, define how to ride a bike? How can the fact of an image containing an apple or an orange be translated into symbols? Even language, which could be considered to be made up of symbols (words), can be vague and depend on context, and is therefore difficult to treat with symbolic systems[406]. These approaches rely on a logical representation of facts being possible, and that treating this logical representation itself is enough to obtain intelligence. In the real world, however, there are rarely such clear representations.

Even where it is possible to represent knowledge in terms of symbols, the systems would run into another issue: the combinatorial explosion. This is an issue where very simple systems often become immensely complex based on the number of possible actions. A chess board, for example, contains a certain number of pieces with different possible actions. In theory, therefore, a system such as the General Problem Solver should be able to be given a representation of the chess board and devise the optimal plan to arrive at a certain other state, e.g., winning the game by checkmating the opponent. However, in practice, fully calculating all the possible moves on a chess board is beyond the possibilities of computers[407]. While such an approach would show very promising results for simpler games, such as tic-tac-toe or problems set out in microworlds, scaling to chess would not be possible.

2.2.6 Conclusion

This section has described some early symbolic systems. These use processes of manipulating symbols in order to create systems that are supposedly intelligent. In the beginning, the approaches led to impressive results in solving issues in constrained domains, leading to great expectations on future developments. These, however, never materialized, as the systems proved difficult to adapt to real world situations. This, together with the inflated expectations, resulted in an AI winter, where research in the area was drastically decreased.

---

403 Terence Horgan & John Tienson, “Representations without Rules” (1989) 17:1 Philosophical Topics 147–174 at 152.

404 Stephen F Davis & William Buskist, *supra* note 19 at 487.

405 Hubert L Dreyfus, *supra* note 379 at 179.

406 Terence Horgan & John Tienson, *supra* note 403 at 151.

407 “Combinatorial explosion”, online: *Oxford Reference* <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095626338>.

However, the area of artificial intelligence saw an insurgence in the seventies, when a new approach focused on incorporating expert knowledge into computer systems took hold. This will be described in the next systems.

2.3 Expert Systems

2.3.1 Introduction

In the seventies, new approaches arose, that incorporated domain-specific knowledge to create systems able to handle typical, real-world problems[408]. These so-called expert systems were less focused on creating general methods to solve problems, and more focused on incorporating human expert knowledge to create working systems. These systems were used to perform medical diagnosis and attempts to understand language, and lead to an industry boom in artificial intelligence[409]. The funding cut during the previous AI winter was restored, and countries and companies began to research the use of expert systems. Artificial intelligence, now dubbed Intelligent Knowledge-Based systems by some, was soon a billion-dollar sector[410].

However, soon shortcomings became evident, and another AI winter with a decreased interest took hold[411] Researchers investigated the market for expert systems in 1995 and found that two thirds of the systems were no longer maintained, with many being inaccessible. They found that managerial reasons, and difficulties of maintaining and integrating the systems into the companies was a prevalent reason for this decline in use[412].

2.3.2 Technological explanation

Expert systems can be seen as a kind of logical reasoning artificial intelligence systems. The idea is that experts typically approach problems in a structured, logical manner. By working with the experts to gather and formalize this information, these approaches can be encoded in computer programs, that are able to tackle the tasks usually performed by experts. Expert systems are seen to be knowledge-intensive, meaning that they are structured around collected knowledge specific to a domain. Experts systems typically involve several important parts.

(1) The Knowledge Base

The knowledge base of a system is where the expert system stores its knowledge. The creation of this database typically involves the interview with experts of a field, to find out the rules they

---

408 Bruce G Buchanan, *supra* note 22.

409 T Grandon Gill, "Early expert systems: Where are they now?" (1995) MIS quarterly 51–81 at 51.

410 Stuart J Russell, Peter Norvig & Ernest Davis, *supra,* note 361 at 24.

411 *Ibid.,* at 22–24.

412 T Grandon Gill, *supra* note 409 at 68.

usually use to solve an issue. This knowledge is then taken by "knowledge engineers" and encoded into a format that the computer can use. This process can take a long time and be expensive, with costs ranging from thousands to millions of dollars, depending on the complexity of the domain[413].

A big question in designing these systems is deciding how the knowledge should be represented in the computer system. Typically, the systems define knowledge declaratively, i.e., in an explicit manner that can be understood by other computer programs[414].

Typically, they involve so-called production rules. These are rules that typically contain an "IF" clause and a "THEN" clause. A simple expert system could look like this[415] :

IF (fever) THEN (predict infection)

A knowledge base of an expert system might contain many hundreds to thousands such rules[416].

(2) The Inference engine

Once the rules are encoded into the system, the system needs to find a way to reason about them to arrive at a conclusion that could be helpful for the user. This part can be referred to as an inference engine. The engine needs to consider some context (such as a sensor reading or answers to questions previously posed by the system) and then use the rules to arrive at an answer to a problem[417].

Many of the steps involved in the inference engine use logic, just like the systems described in the previous chapter[418]. However, they might also use other concepts, that enable them to handle uncertainty, for example[419]. Just like in some of the earlier symbolic systems, the inference engines are often distinct from the knowledge base, meaning that the inference engine can in theory be used to support expert systems in different domains. Of course, the knowledge basis is usually specific to a domain and task[420].

An important consideration is in which order the knowledge available to the system should be considered. Researchers distinguish between data-directed (forward) and goal-directed

---

413 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 19, 20; Ekbia, *supra* note 368 at 95.

414 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 10.

415 Paul Smolensky, *supra,* note 370 at 98.

416 Bruce G Buchanan & Reid G Smith, *supra,* note 24 at 27, 28.

417 *Ibid.,* at 17–19.

418 *Ibid.,* at 11, 12.

419 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 23.

420 Bruce G Buchanan & Reid G Smith, *supra,* note 24 at 14.

(backward) reasoning. Forward reasoning starts with the facts, and reasoning through them to arrive at the required conclusions[421]. Backwards reasoning, on the other hand, starts with the goals, and reasons backwards through the facts that are required to fulfill this goal. This can work well in a conversation, as the system asks the user for facts only when they are needed[422].

(3) User Interface

Once the system is created, it also needs a way to interact with the user. Early expert systems tended to use textual interfaces, prompting the user for their answers[423]. Newer systems also used graphical interfaces. Buchanan and Smith highlight the importance of the user interface for the end-user utility of the system[424]. A possible mistake might be to assume that the terminology and point of view of the experts is the same as the end-users of the product, limiting the usefulness of the systems[425].

Since expert system arrive at conclusions based on explicit rules, they can often provide an explanation of their reasoning process. Many expert systems offer a system to explore the reasoning steps that lead to a certain answer through the interface[426].

2.3.3 MYCIN – an expert system for medical diagnosis

A notable expert system is MYCIN, developed at Stanford University. It aimed to support physicians by giving antimicrobial therapy advice[427]. It consists of three programs: a consultation system, an explanation system, and a rule acquisition system[428].

An example of a decision rule from the MYCIN system[429]

IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) THE AEROBICITY OF THE ORGANISM IS ANAEROBIC
THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE IDENTITY OF THE ORGANISM IS BACTEROIDES

421 *Ibid.,* at 16.

422 *Ibid.,*at 17.

423 *Ibid.,*at 21.

424 *Ibid.,*at 34.

425 *Ibid* at 20.

426 *Ibid.,* at 21, 22.

427 Edward H Shortliffe et al, "Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system" (1975) 8:4 Computers and Biomedical Research 303–320 at 304.

428 *Ibid.,* at 305.

429 *Ibid*.

The consultation system consisted of encoded rules and a method to traverse these rules, in order to recommend an action. The rules were created by looking at representative case histories, and contained premises and an action part, encoded in the "Lisp" programming language. Figure 1 shows an example of such a rule. There were several hundred such rules. In each session, the system would ask the physician whether the premises applied, and then provide them with possible organisms and recommended treatments[430]. The system also provided an explanation system, which allows the physician to explore how and why a rule was chosen[431]. It was also possible for physicians to interactively add new rules to a system[432].

MYCIN is an illustration of how powerful expert systems could be. Compared to GPS, it is much more based on domain knowledge, which is encoded into hundreds of rules using Lisp. Eventually, the system seemed to be able to perform better than junior doctors[433]. However, it also illustrates some of the difficulties of these systems. Even in the medical domain, where knowledge should be quite structured, the researchers found that physicians were often not sure how they arrive at a certain conclusion, making data acquisition difficult[434]. MYCIN was in the end never used in a real-life scenario. It is unclear whether this was due to legal and ethical risks or difficulties getting the system to be accepted by the doctors[435].

2.3.4 Other examples

There are many further examples of expert systems. This section will briefly list a few such systems. Of course, this list is not exhaustive – expert systems were a large industry and were used for many tasks, including data interpretation, equipment diagnosis and maintenance, credit authorization, inventory control, design and scheduling, among others[436].

A very early example of expert systems is the DENDRAL program, which aimed to infer molecular structure from readings of a mass spectrometer. Doing so in a brute-force manner, by generating all possible molecules compatible with a specific reading, is impossible, as there are too many possibilities. Instead, the researchers learned which common, well-known patterns analytical chemists were typically looking for, and encoded these rules into the system[437].

---

430 *Ibid.,* at 307–309.

431 *Ibid.,* at 309–315.

432 *Ibid.,* at 315–318.

433 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 23; R Duda & E Shortliffe, "Expert Systems Research" (1983) 220:4594 Science 261–268 at 264.

434 Shortliffe et al, "Computer-based consultations in clinical therapeutics", *supra* note 427 at 319.

435 David A Teich, "Artificial Intelligence (AI), Healthcare and Regulatory Compliance", online: *Forbes* <https://www.forbes.com/sites/tiriasresearch/2018/03/20/artificial-intelligence-ai-healthcare-and-regulatory-compliance/>; John McCarthy, "SOME EXPERT SYSTEM NEED COMMON SENSE" (1984) 11.

436 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 4–7.

437 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 22, 23.

Another notable expert system is XCON (Expert CONfigurer). It was used by the Digital Equipment Corporation to configure orders of systems. The input to the system was an order by the customer. The system would then apply a number of rules to decide how the system should be configured, and display the spatial relationships between the different components[438]. It was one of the early commercially successful systems[439]. Eventually, the system contained 6,000 different rules[440]. It was credited with saving the company 15 million USD between 1980 and 1985[441].

Several expert systems have been constructed in the legal domain. Here, the knowledge is often contained in legal rules, which are similar to production rules to some extent. Legal rules are often structured as criteria, containing requirements and consequences. This is similar to how production rules work. A notable example of such systems is HYPO, developed to provide explainable arguments for legal outcomes based on a database of encoded case law[442]. The Cyberjustice Laboratory is developing a tool to provide decision support to parties to landlord and tenant conflicts in Quebec[443]. There are also tools that enable the creation of such expert systems with various features[444]. Finally, Turbotax and similar software, used by millions to file their taxes, can be seen as expert systems, in that they contain expert-created rules to generate documents for submission to the tax authorities[445].

2.3.5 Discussion

Expert systems provided a resurgence to the field of artificial intelligence and were used commercially in a number of firms. They turned artificial intelligence into a billion-dollar industry. Despite this, they are no longer considered to be the most promising way to achieve "true" artificial intelligence. This section will discuss some of the advantages and drawbacks of the knowledge-intensive approach to artificial intelligence.

---

438 John McDermott, "RI: an Expert in the Computer Systems Domain'" (1980) AAAI 18 at 269.

439 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 336.

440 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 27.

441 John J Sviokla, "An Examination of the Impact of Expert Systems on the Firm: The Case of XCON" (1990) 14:2 MIS Quarterly 127–140 at 127.

442 Kevin D Ashley, *Modelling Legal Argument: Reasoning with Cases and Hypotheticals* (PhD Thesis, University of Massachusetts, 1988) [unpublished].

443 Hannes Westermann, *ODRAI - JusticeBot, prototype and challenge* (2020); "JusticeBot", online: *Laboratoire de cyberjustice* <https://www.cyberjustice.ca/en/justicebot/>.

444 "Docassemble", online: *Docassemble* <http://docassemble.org/>; "A2J Author", online: <https://www.a2jauthor.org/>.

445 "TurboTax® Official Site: File Taxes Online, Tax Filing Made Easy", online: <https://turbotax.intuit.com/>; Lorelei Lard, "Expert systems turn legal expertise into digitized decision-making" (17 March 2016) online: *ABA Journal* <https://www.abajournal.com/news/article/expert_systems_turn_legal_expertise_into_digitized_decision_making>.

(1) Advantages

(a) Easy to get started

One advantage of the systems is the ease of getting started in creating useful systems. There are a number of systems that provide the tooling necessary to quickly start entering rules and exploring the output[446]. In fact, this exploratory way or building expert systems is considered an important step of creating expert systems, as it allows the developed to construct and evaluate prototypes of the system rapidly[447].

(b) Well suited to encode a certain kind of information

Another advantage of expert systems is that they are naturally quite suited to certain domains, where there is a lot of formalized knowledge. In the legal and medical domain, for example, the knowledge is often explicitly encoded in laws and books. This knowledge takes the shape of criteria and conclusions, which can lead to other conclusions or motivate actions. This kind of knowledge is well suited to be encoded into the production rules required by the expert systems to function.

(c) Explainable

Another advantage of expert systems is that they are quite easy to explain. Since they rely on encoded declarative knowledge, the rules are written in a way that can be understood by humans. It is thus possible to trace the reasoning steps of a system, to see why it arrived at a certain conclusion. Further, many systems provide an interactive mode, where commands can be executed in the context of the expert system session, which allows the user to query the steps taken in reasoning and see how alternative actions might play out[448].

(2) Disadvantages

(a) Difficulty of creating, maintaining

Unfortunately, while expert systems are typically easy to get started with, the complexity rises once systems are designed to be put to real use. Often, there are huge numbers of rules in a domain. Encoding these takes time and effort, costing potentially millions of dollars[449]. As the creators of MYCIN found out, experts often rely more on intuition than they expect, making the

446 Bruce G Buchanan & Reid G Smith, *supra,* note 24 at 22–24.

447 *Ibid.,* at 20.

448 *Ibid.,* at 26.

449 *Ibid.,* at 19.

recording of rigorous rules difficult[450]. This difficulty, together with legal and ethical issues, prevented MYCIN from being used in practice[451].

Further, the effort does not go away once the system is complete. In traditional software development, the main effort is often the creation of the program. In expert system development, however, rules often have to be updated and maintained to remain accurate. In legal systems, for example, new legislation or court cases might change the way the law is applied. The creators of the XCON system claim that 50% of the system had to be rewritten annually to keep up with changes in requirements[452]. This makes development difficult. A report found that issues with maintaining and integrating expert system development at firms were significant reasons the systems were abandoned[453].

Once an expert system has been created, it can further be difficult to validate that it works. Buchanan and Smith argue that there are three main dimensions that an expert system can be judged on – computational, psychological and performance. The computational axis involves the speed and extensibility of a program. The psychological axis includes how easy to use and natural a system is. Performance refers to the competence of the system and whether the advice is correct, and if time and money is saved. This kind of analysis might be done rarely, they claim[454].

(b) Difficulty of generalizing

One issue that expert systems might run into is the difficulty of generalizing. Generalizing is the fact of performing well on new, unseen situations. This can be difficult in expert systems. Expert systems reason by traversing a set of explicit rules. If a case falls outside of a rule, the system will not be able to deal with the issue, causing the system to fail[455]. Further, the system itself would not be aware that it would fail in a certain instance, providing incorrect results without an indication that it might be wrong[456]. Unlike experts, expert system do not have a notion of common sense or intuition that allows them to solve problems in the absence of an explicit rule[457]. Ekbia describes how an expert system decided to give a job to a teenager who claimed to have worked a job for 20 years. While a human expert would have probably spotted this as an issue, the system failed to do so in the absense of a specific rule[458].

---

450 R Duda & E Shortliffe, *supra* note 433 at 265.

451 David A Teich, *supra* note 435.

452 John J Sviokla, *supra* note 441 at 137.

453 T Grandon Gill, *supra* note 409 at 68.

454 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 24.

455 *Ibid.,* at 14.

456 *Ibid.,* at 22; John McCarthy, *supra* note 388 at 1032.

457 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 15.

458 H R Ekbia, *supra,* note 368 at 96, 97.

Cyc, a large-scale research project, aimed to overcome this issue by encoding millions of common-sense facts into an expert system. The idea was that this base of knowledge would be useable to give common sense to any expert system, thereby breaking the knowledge acquisition bottleneck[459]. While it is unclear exactly how well this approach works, it seems like the system still has trouble grasping some fundamental notions[460]. Ekbia believes this might be related to the difficulty of representing tacit knowledge, meaning, change and context in terms of declarative rules[461].

Unlike real experts and machine learning models, expert systems are further unable to learn on their own. All of the knowledge is encoded by a person. If a system applies a rule in a way that leads to an incorrect outcome, it will do the same mistake on the next iteration. Being able to figure out new rules is also not in the scope of these systems[462].

(c) Difficulty of dealing with implicit knowledge

Finally, there is a large limitation that is inherent to expert systems. Just as the name implies, they seek to emulate the kind of reasoning an expert would apply. This works well for knowledge that is structured and formulated in hierarchies[463].

However, there are a lot of cases where this kind of reasoning is not applicable. Issues that require any kind of intuition or learning are often out of the scope for these systems. In determining whether an object we see is an apple, or how to ride a bike, people are unlikely to use rigid rules[464].

Further, it seems like a lot of expert reasoning has some aspect of intuition. In the legal domain, for example, people often have to interpret vague concepts to determine whether a rule should be applied or not[465]. In doing this, the judge often has to weigh common sense, the aim of legislation and even societal needs to arrive at a relevant conclusion, in the absence of a specific rule. Implementing this kind of reasoning into an expert system, which by their very nature rely on rules, is very difficult. Dumouchel refers to this aspect of reasoning as "judgment" and holds it as one of the main ways machine intelligence differs from human intelligence[466]. This limits the applicability of expert systems to a part of the reasoning of an expert.

---

459 *Ibid.,* at 91, 92.

460 *Ibid.,* at 110–115.

461 *Ibid.,* at 118–125.

462 Bruce G Buchanan & Reid G Smith, *supra* note 24 at 21.

463 *Ibid.,* at 30.

464 H R Ekbia, *supra* note 368 at 119.

465 H L A Hart, "Positivism and the Separation of Law and Morals" (1957) 71 Harv L Rev 593–629 at 607.

466 Paul Dumouchel, *supra,* note 35 at 256.

2.3.6 Conclusion

In this section we discussed expert systems, which can be seen as a class of systems using symbolic reasoning. Expert systems aim to capture and encode the reasoning steps performed by experts in certain areas into computer systems able to perform these steps by themselves, saving time and money. Experts systems brought back the interest in artificial intelligence in the eighties, and lead to the creation of a billion-dollar industry. However, soon shortcomings of these systems became obvious, and today most of the early systems have disappeared. Despite this, many of the systems still exist and are used by millions to, for example, file their taxes.

2.4 Conclusion

Is human intelligence based on symbols, e.g. representations of objects, their relationships, plans and goals in our brain? Some argue that this is the main defining characteristic of humans[467]. In this section, we have presented two approaches to artificial intelligence that seek to emulate human intelligence, and produce intelligence computer systems, by treating intelligence as the manipulation of symbols.

The early symbolic approaches relied on a variety of systems, using logic to manipulate symbols. They worked remarkably well for small scale problems and seemed to open the door to truly intelligent computers. However, they ran into a wall when being applied in a real-world context.

Expert systems, on the other hand, aimed to incorporate rules into systems able to reason like an expert in a circumscribed domain. While useful in some domains, they fell somewhat out of favor after the complexity of building large, adaptable systems became apparent, and were also unable to deal with implicit knowledge, which seems to play an important part in human intelligence.

Looking at intelligence purely in terms of symbolic systems might therefore not be the key to unlocking intelligence in computer systems. Instead, artificial intelligence research shifted to the sub-symbolic level, to let the computer learn representations for itself rather than providing it with the information it needs in a declarative fashion. We will describe these approaches in the subsequent sections.

---

467 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 25.

3. Machine Learning

3.1 Introduction

Quick facts

- Machine learning is an approach to artificial intelligence that relies on algorithms autonomously detecting patterns and correlations in data, to build a model that can be used to predict unseen data. This approach is very flexible and can be used to tackle a huge variety of tasks.
- While the systems learn from the data autonomously, there is still some human expertise required to choose a task, collect the required data, build and evaluate a model and integrate it into a workflow.
- Recently, deep learning has arisen as the most powerful way of doing machine learning, and lead to a number of breakthroughs in a variety of fields. They are especially well suited to dealing with textual data, such as images, video and text.
- It is important to be aware of limitations of machine learning and deep learning when employing them, especially in sensitive contexts. The systems reason very differently from humans, and may arrive at unexpected, bad or harmful solutions to problems.

After the difficulty of using expert systems to perform advanced tasks became apparent, another AI winter appeared. During this time, the methods used in artificial intelligence shifted. The focus landed on machine learning, methods that are able to autonomously learn from data, rather than requiring the hand-coding of rules like the symbolic approach. These methods were more connected to previous research in mathematics and statistics and had a much stronger focus on evaluation on realistic datasets[468]. As increases in computing power and larger data sets became available, this approach flourished, and rekindled the waning interest in artificial intelligence[469].

Over the past few years, a new class of algorithms for machine learning, known as deep learning, has emerged. This technology has taken the world by storm and is now the most widely used approach for many tasks. Unlike traditional machine learning, deep learning is less reliant on feature engineering – often it is able to create its own complex representation of the input data, which allows it to capture the complexity of the real world to a much larger extent than previously used models. This makes it especially suitable for complex, unstructured data such as images, text and videos[470].

In this section, we will present the methodology behind building machine learning systems, from choosing a task, creating a dataset, choosing and training a machine learning model to eventually

468 Stuart J Russell & Peter Norvig, *supra* note 14 at 24–26.

469 *Ibid.,* at 26.

470 *Ibid.,* at 750.

deploying the model in the real world. We will also discuss a few application areas and shortcomings of the models used.

3.2 Machine Learning, from dataset to model

In order to explore and explain the different steps usually involved with creating a machine learning model with a certain goal, let us step through the steps usually involved in creating such a model. These steps are[471] :

- deciding on a task – Deciding what the model should be used for (see section 2) and deciding how the model should be constructed to achieve these goals;

- selecting or creating a dataset – Identifying a dataset that could be used to train a model to achieve the forementioned task or creating such a dataset from scratch. This includes the tasks of choosing which samples to include in the dataset, and how the sample should be represented in the dataset in terms of features and labels, and performing this representation;

- data preparation – After the data has been collected, it needs to be translated into a format that the computer can understand, usually in the form of a vector. A number of choices have to be made on how to represent the samples in this format. The data also has to be split into data used to build the model, and to evaluate the model (training and testing data);

- selecting and training a model – Once the dataset has been selected, it is typically split into two parts, one for training and one for testing. The training part is used to train a machine learning model. There is a wealth of models to choose from, and many methods to optimize and adjust the training process;

- evaluating the model – The remaining part of the dataset, the test part, is used to evaluate the performance of the model, to see how well it performs on its task;

- deploying the model – The model is deployed and integrated into a worksflow. This often includes the continuous evaluation of the performance of the model.

Each of these steps contains unique challenges and may overlap in terms of time. They are the most typical for “supervised” learning, where the target value is known in advance.

---

471 Compare Yufeng G, *supra* note 32; Harini Suresh & John V Guttag, *supra* note 32; Chanin Nantasenamat, *supra* note 32; Nithya Sambasivan et al, "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI" (2021) 15 at 6.

In order to explain the different steps, we have chosen to follow the example of the Iris plant database[472]. This is a dataset of 150 Iris flower plants belonging to three subspecies (Iris-Setosa, Iris-Versicolour and Iris-Virginica), each with a measurement of the sepal and petal width and length. The goal is usually to predict the subspecies of the plant based on these measurements. For a machine learning task, this dataset has a rather small number of samples and features. However, it is one of the most well-known datasets in machine learning, and is often included as a "toy" dataset with frameworks to learn the general steps involved with machine learning[473]. Therefore, it will serve as an excellent example to illustrate the tasks associated with machine learning.

Figure 1 - Iris Setosa flower

3.2.1 Deciding on a task

The first step in building machine learning models is deciding on the task the model should solve[474]. This task could arise, for example, as the result of a business need or as a research project. In the business context, the models might aim to support a human in accomplishing a certain task more efficiently or in a better way, or even to replace human labor completely on some tasks. Many models are also created in the academic context. Here, the researcher may wish to explore whether a certain novel problem can be solved using machine learning, or develop a way to create models that perform better at solving certain kinds of problems, thereby advancing the state of science.

In our example, we want to create a model that can identify the species of a plant based upon the measurement of its sepals and petals. This could be used to support people in classifying plants they discover on the field.

(1) Types of Tasks

Part of figuring out the task is deciding on the machine learning mode the model will use. These are generally split into supervised, unsupervised or reinforcement learning[475]. However, the distinction is not always clear, and some modes (such as self-supervised learning) combine aspects of supervised and unsupervised learning[476].

---

472 Ronald A Fisher, "The use of multiple measurements in taxonomic problems" (1936) 7:2 Annals of eugenics 179–188.

473 "7.1. Toy datasets — scikit-learn 0.24.1 documentation", online: <https://scikit-learn.org/stable/datasets/toy_dataset.html#iris-dataset>.

474 Stuart J Russell & Peter Norvig, *supra* note 14 at 704.

475 *Ibid.,* at 705.

476 *Ibid*.

(a) Supervised Learning

Supervised learning asks the model to predict a *label*, based upon a set of features[477]. The label (or target) can be seen as the question we are asking of the model, i.e. the property we ask it to predict or establish[478]. In order to build the model in a supervised manner, we would give it many *samples* (individual datapoints) together with the label. Each sample is represented by a number of *features,* i.e. data describing the sample. A feature could be, for example, how many windows a house has, the length of a petal, the color of a pixel in a certain location of an image, or a word in a sequence.

When given the samples (comprised of features and labels) the algorithm will then (ideally) learn to identify the correlations between the features and the label. Once this is done, we can give the model a new sample that it has never seen, and the model will predict the correct label based on the provided features. However, at no point do we tell the model how the features are connected to the label – the model is able to learn this by itself. This is the big distinction to the symbolic approach, where we explicitly tell the model the rules of how to solve the task.

This simple idea of training a model on previous data, with features and a label, and then use this model to predict values, is remarkably versatile and used to accomplish many important tasks in many sector. Here are a few examples of tasks that can be accomplished using supervised machine learning, and which features and labels might be used:

Task	Features	Label
Predict which object is in an image	The image with the object	The type of object in the image
Predict whether an online review is positive or negative	The text of the review	Positive/Negative
Predict whether an email is spam	The text of the email, the address of the sender, whether the email contains links etc.	Spam/Ham (not spam)
Predict the price of a stock	The price of the stock on previous days, price of other similar stocks, news coverage of the company	The price of the stock

Supervised Learning is likely to be the most commonly used style of machine learning. However, it has some shortcomings, as determining the label value for training the model can be very hard work, as we will see below.

(b) Unsupervised Learning

Unsupervised Learning is another kind of learning. Unlike supervised learning, unsupervised learning does not rely on the label values in its prediction. Instead, the algorithm is given only

477 Future of Privacy forum, *supra* note 253 at 10.

478 Stuart J Russell & Peter Norvig, *supra* note 14 at 653.

features, and asked to spot patterns in this data on its own[479]. A common such task is, for example, the *clustering* of datapoints[480]. Here, the algorithm identifies clusters of similar samples[481]. This has the advantage of not requiring the expensive labeling of datapoints and can further spot correlations that might not have an explicit label. However, it also means that the model might not identify the categories the author desires. For example, when clustering a dataset of fruits, the developer might want the model to group fruits of the same kind together. However, the model might instead group images by the number of fruits they contain, or whether an image contains leaves or not.

There are some tasks where Unsupervised learning is required. For example, a bank might want to detect unusual transactions to see whether they are fraudulent. Unsupervised machine learning methods can learn what typical transactions look like, and then notify the user when a transaction is unusual. Another common use case is for recommendations. The algorithm can automatically cluster products that are frequently bought at the same time together, and provide recommendations for users based on these patterns[482].

Unsupervised learning is often seen as very promising for the future, as it requires less expensive data collection (see below). Yann LeCun, one of the founder of the modern deep learning methodology, believes that unsupervised learning is the future of artificial intelligence[483].

(c) Reinforcement Learning

In supervised and unsupervised learning, the algorithm learns from data by passively observing it, and building a model of the data. However, if we compare this to how humans learn, we can see that it is quite different. While us humans can learn by passively observing, we are also part of the world, and can interact with it to see what happens in response to our actions[484].

Reinforcement learning more closely mimics this kind of learning. Here, the model is referred to as an agent, with the power to act. The agent is placed in an environment and given a number of actions it can perform to act upon this world. Further, the agent receives a reward when it acts in a way that the researcher wants to encourage[485].

---

479 Future of Privacy Forum, *supra* note 253 at 16.

480 Stuart J Russell & Peter Norvig, *supra* note 14 at 653.

481 Future of Privacy Forum, *supra* note 253 at 16.

482 *Ibid.,* at 16, 17.

483 Karan Hao, "The AI technique that could imbue machines with the ability to reason"(12 July 2019), online: *MIT Technology Review* <https://www.technologyreview.com/2019/07/12/65579/the-next-ai-revolution-will-come-from-machine-learnings-most-underrated-form/>.

484 Paul Dumouchel, *supra* note 115 at 247.

485 Stuart J Russell & Peter Norvig, *supra* note 14 at 789.

What could be an example of such an environment? Games are often used as environments to train these agents. For example, the agent would be given the state of the chess board at every turn and then decides how to act based on the position of the pieces. If the agent wins, it will receive a reward[486]. Another environment could be a simulated physical world where the agent has to learn how to walk by activating artificial muscles, with the reward being based on how far the agent is able to move forward in a certain time period[487]. Or, the agent might be tasked with getting as far as possible in a computer game, such as breakout[488] or even the strategy game Starcraft II[489]. Usually, using a virtual environment is much easier than a physical one, since they can be replayed very often very quickly, and do not break if the agent does something wrong[490]. Further, the agent can improve by playing against itself over and over again[491]. However, the technique has also been used to train real-world robots to walk[492].

At first, the system is likely to fail to obtain the reward, as it performs random actions. However, soon, a random action might lead to the agent obtaining a reward. Perhaps it falls forward and thereby moves a meter to the front. The next iterations of the agent can then learn from this experience, and further try to increase the reward, hopefully eventually learning to walk.

One of the powerful aspects of this approach is that the researcher does not have to teach the agents how to solve a problem. This means that the researcher does not have to be an expert in the field, or even know how it should be tackled[493]. The agents should discover this autonomously through exploration of the possibilities of the environments. This also means that the agents can surpass the level of proficiency of the human or find completely new ways of accomplishing tasks.

On the technical level, there are many different ways of performing reinforcement learning, whether with simple linear classifiers or deep neural networks. Some of these methods can be trained by giving some previous information, such as training it to play chess using thousands of matches played by humans, or by teaching it the rules of chess (i.e. which pieces can move which

---

486 David Silver et al, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play" (2018) 362:6419 Science 1140–1144.

487 Nicolas Heess et al, "Emergence of Locomotion Behaviours in Rich Environments" (2017) arXiv:170702286 [cs], online: <http://arxiv.org/abs/1707.02286> arXiv: 1707.02286.

488 Volodymyr Mnih et al, "Playing Atari with Deep Reinforcement Learning" (2013), online: <http://arxiv.org/abs/1312.5602>.

489 Oriol Vinyals et al, "Grandmaster level in StarCraft II using multi-agent reinforcement learning" (2019) 575:7782 Nature 350–354 Primary_atype: Researchpublisher: Nature Publishing GroupSubject_term: Computer science; StatisticsSubject_term_id: Computer-science;statistics.

490 Stuart J Russell & Peter Norvig, *supra* note 14 at 790.

491 *Silver et al, supra* note 486.

492 Tuomas Haarnoja et al, "Learning to Walk via Deep Reinforcement Learning" (2019), online: <http://arxiv.org/abs/1812.11103>.

493 Stuart J Russell & Peter Norvig, *supra* note 14 at 790.

way). Some methods start with nothing, meaning that the agent has to find out which moves are legal by trying and failing[494].

One of the difficulties in reinforcement learning relates to how to properly define the reward function, i.e. how to tell the agent that it is succeeding. In many games, such as chess, the reward can only be given at the very end, by giving points to the agent if it won the match. This is a so-called sparse reward, as the agent is only rewarded at the very end, meaning that it might be tricky for it to understand which particular moves contributed to an outcome. This problem is referred to as the credit assignment problem[495]. Other environments give more frequent rewards. For example, in the environment where the agent must learn to walk, the reward function can give a reward for the progress can be given at every step[496].

Further, one must take care that a given reward function actually rewards the behavior we are looking for. Otherwise, the agent might learn to do something undesired that still maximizes the reward function. For example, the agent in the walking simulator might find a trick in the physics engine that allows it to fall over and slide forwards, rather than actually learning to walk[497].

We will get back to reinforcement learning in the section about deep learning, since this technology applied to the reinforcement learning has led to a number of significant breakthroughs.

(2) Example

In our example, the task is to predict flowers from the length and width of sepals and petals. Let us look at a few samples, one for each of the flower types:

Sepal Length	Sepal Width	Petal length	Petal width	Flower type
5,1 cm	3,5 cm	1,4 cm	0,2 cm	Setosa
7 cm	3,2 cm	4,7 cm	1,4 cm	Versicolor
6,3 cm	3,3 cm	6 cm	2,5 cm	Virginica

We have a dataset with datapoints consisting of the features of sepal length and width, and the corresponding label values, i.e. whether a flower is an Iris-Setosa, Iris-Versicolour or Iris-Virginica. Therefore, it is a supervised learning task. Further, the label is discrete – either Setosa, Versicolor or Virginica. Therefore, the task is one of classification.

494 *Ibid.,* at 790–803.

495 *Ibid.,* at 807.

496 *Ibid.,* at 790.

497 "Specification gaming: the flip side of AI ingenuity", (9 July 2020), online: *Deepmind* <https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity>.

In our case, we already have the dataset corresponding to this task. However, in the real world, this might not be the case, and the dataset has to be created from scratch, or adapted from another dataset. Let us look at how this occurs.

3.2.2 Selecting or creating a dataset

As previously mentioned, machine learning methods rely on the existence of datasets to learn from. Often, a big part of the work in creating machine learning algorithms is in assembling these datasets. This can be both tricky and time-consuming. For other research, there might be existing datasets that support the desired task, which can greatly speed up the process.

The way the dataset is created has the potential to lead to harmful outcomes once the model is deployed. Suresh and Guttag have created an overview over such potential issues, which we will refer to below[498].

(1) Creating a new dataset

If no existing dataset covers the envisioned task, it is up to the developer to create their own dataset to train the machine learning system. This means that they need to decide which samples to include in the dataset, how to represent these datapoints in terms of features, and what the prediction label should be. In all of these steps, the developer has to make certain choices. How these are taken is important for the functioning of the system. However, they can also introduce unintended harmful consequences into the system. A common saying in data science circles is ”Garbage In, Garbage Out”[499], meaning that if the data is of bad quality, the model trained on that data will not work well.

(a) Which samples should be in the dataset?

The first step in building a dataset is to determine which samples to include in the dataset. From all the possible examples in the world, they need to select a population for the purposes of creating the dataset. There could be a wide variety of sources for this data, depending on the envisioned task. These include interviews with people, medical data, websites or images crawled from the internet, measurements from Internet-of-things devices, historical arrest data etc.

In collecting this data, it is important to be aware of potentially harmful effects that could be introduced by the collection. One such effect is what Suresh and Guttag refer to as “Historical bias”. This occurs when historical factors introduce some bias in the data that will be reflected in

498 Harini Suresh & John V Guttag, *supra* note 32.

499 “Clipped From The Times”, *The Times* (10 November 1957) 65.

the data[500]. As an example, they mention that CEOs of large companies are overwhelmingly men. However, displaying almost only men when searching for images of CEOs might cause harm[501].

It is important that the collected data accurately reflects the distribution of the real world, and the problem we are looking to solve. For example, if we are aiming to predict the price of a house, we need to include all different kinds of houses in the data. If, for example, we only include houses sold by a certain real estate agent that focuses on expensive property, the resulting model will not work well in reality as it encounters other kinds of houses. This can also lead to harmful consequences, referred to as "representation bias"[502]. For example, if an algorithm is trained to recognize faces, but is only trained on faces of a certain ethnicity, it will fail to accurately recognize faces of other ethnicities, with potentially harmful consequences[503].

In our example, since we want to predict the type of the Iris flower from measurements of the flower, it would be reasonable to include examples of this flower in the dataset. Since it is impossible to include all flowers, we will need to select a subset of flowers to include. In this case, the researcher chose to include 150 measured flowers, 50 of each type.

(b) Which features should be used for the data?

Once we have decided which samples to include in the dataset, it is important to decide on how these samples should be represented in the dataset. This can be seen as the decision on how the samples should be measured.

For some types of samples, this might be quite simple. If the goal is to detect which object is in an image, the obvious feature to include is the image itself, i.e. the individual pixel color values at different places. If the aim is to identify the sentiment of a sentence, the sentence should be included as data.

But what if the goal of the project is to predict housing prices? How should the house be included in the model? A number of features could be used, such as square footage of the house, number of windows, crime rate per capita in the surrounding area, garden square footage etc. All of these could have an impact on the house price, and the more of these are included, the more sophisticated and potentially accurate the resulting model will be.

However, we have to be careful with the features that we choose. Would, for example, the average electricity use of a household have an impact on the price of the house? It could, but could also vary widely with the current inhabitants. In that case, including the feature in the

500 Harini Suresh & John V Guttag, *supra* note 32 at 4.

501 *Ibid.,* at 5.

502 *Ibid*.

503 Joy Buolamwini & Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (2018) Conference on Fairness, Accountability and Transparency 77–91.

dataset has the potential of confusing the model, leading to decreased performance[504]. This is especially the case if there are not a huge number of samples in the dataset[505].

Further, by choosing the features, we are implicitly making a choice over what matters about a sample. However, features only describe a certain aspect of a sample, and cannot capture the full complexity of the real world. For example, there might be features that affect the house price that are not included in the above list (such as the presence of a pool), but that matter in individual cases, leading to inaccurate predictions. The features might also be monitored differently across different groups, leading to potentially discriminating outcomes. This has been referred to as ”Measurement bias”[506].

(c) Which labels should be used for the data?

Finally, we have to decide what the label for a prediction should be. What is the aspect we want to predict? This, of course, depends wholly on the aim of the experiment. Possible examples of labels include (as listed above) whether a sentence is positive or negative, what type of object an image contains, which words were spoken based on a voice recording or the price of a stock on the next day. For translation, the “label” could be the English sentence, based on the input of a French sentence.

There is a large variety in the types of labels than can be used. Labels are frequently separated by whether they are discrete or continuous. If the label is discrete, the model should predict a class (such as the type of the flower)[507]. If another class is predicted, the prediction is wrong. This kind of prediction is referred to as *Classification*[508]*.* Other labels, on the other hand, might be continuous, such as the price of a stock. The prediction can be more or less wrong, depending on how far from the real price the prediction is. This type of prediction is referred to as *Regression*[509]*.*

There is often a difference between the label we teach the model to predict, compared to the label we actually want to predict. Often, the real label is not captured anywhere, which means that we have to resort to a proxy, that hopefully captures the desired label. For example, if we try to predict whether a certain individual has a disease, we usually only have access to people who were *diagnosed* with a disease. If certain groups are more likely to be misdiagnosed, this can be a biased proxy[510]. In cases of predictive policing, where the system aims to predict whether an individual is likely to re-offend, the real label is who will re-offend. However, the data we have access to is only who was re-arrested. Again, if certain minorities or areas are more

504 Future of Privacy Forum, *supra* note 253 at 9.

505 Pedro Domingos, *supra* note 265 at 82.

506 Harini Suresh & John V Guttag, *supra* note 32 at 5.

507 Future of Privacy Forum, *supra* note 253 at 12.

508 Stuart J Russell & Peter Norvig, *supra* note 14 at 652.

509 Future of Privacy Forum, *supra* note 253 at 11; Stuart J Russell & Peter Norvig, *supra* note 14 at 652.

510 Harini Suresh & John V Guttag, *supra* note 32 at 5.

highly policed, they will be overrepresented in terms of being re-arrested, leading the system to learn a biased representation. In essence, the system does not learn who will re-offend, but rather who will be arrested again[511].

Once more, the label may also represent a simplification of the real world. If we build a model to predict academic success, and pick the GPA as a label, we will miss aspects of internships, work experience, social connections etc.[512]. The model will learn to predict the GPA only, ignoring the other factors, which could paint a misrepresentative picture if this value is then used to signify academic success.

(d) Collecting and annotating the data

Once the decisions on how the data should be represented and labeled have been made, the actual labeling still remains to be done. This can be a massive undertaking, since the number of samples required to build accurate machine learning models can easily range into the thousands or even millions. The process can consist of employees that sit by a computer and select the desired label for each sample[513]. Depending on the data, the task can also be crowdsourced, where individuals are paid online to select the appropriate label[514].

Labelling samples can be a significant bottleneck in the creation of machine learning systems[515]. However, it is a very important part of the process. Usually, the more data the machine learning system has access to, the better it will be at spotting patterns and learning to predict new examples. In many cases, having more data is even more important than the type of machine learning algorithm used[516]. However, the quality of the data is also important – mistakes in the labelling of data could confuse the model or even make it learn erroneous correlations. Since a part of the same data is usually also used for evaluation of the model (see below) this can lead to significant problems in understanding the performance of a model. Researchers discovered that in 10 commonly used datasets, an average of 3.4% of the labels are wrong[517].

---

511 *Ibid.,* at 5, 6.

512 *Ibid.,* at 5.

513 Dave Lee, "Why Big Tech pays poor Kenyans to programme self-driving cars", *BBC News* (3 November 2018), online: <https://www.bbc.com/news/technology-46055595>.

514 Stuart J Russell & Peter Norvig, *supra* note 14 at 705.

515 Ciarán, "'I'm Not A Robot': Google's Anti-Robot reCAPTCHA Trains Their Robots To See", (25 October 2017), online: *The World's Number One Portal for Artificial Intelligence in Business* <https://aibusiness.com/recaptcha-trains-google-robots/>; Lee, *supra* note 162; Pedro Domingos, *supra* note 265 at 85.

516 Cynthia Rudin & David Carlson, "The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis" (2019), online: <http://arxiv.org/abs/1906.01998> at 10; Pedro Domingos, *supra* note 265 at 84, 85.

517 Curtis G Northcutt, Anish Athalye & Jonas Mueller, "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks" (2021), online: <http://arxiv.org/abs/2103.14749>.

Researchers have devised a number of tricks to obtain training data in a more efficient way. When a website asks us to identify whether an image contains a boat to ensure that we are human, the data is later used as labelled data for machine learning algorithms[518]. Active learning is a research direction that seeks to make annotation of data more efficient, by asking the algorithm to suggest pertinent examples to label next[519]. In research conducted at the Cyberjustice Laboratory jointly with researchers from the United States, we devised a method to use machine learning methods in the process of annotation, by showing similar sentences across a corpus[520].

(2) Using an existing dataset

Another possible way of obtaining a dataset is to use an existing public dataset. This has the advantage of not requiring the expensive and time-consuming data collection and annotation steps, as described above. However, depending on the task at hand and the purpose of developing the machine learning model, publicly available datasets might not cover the envisioned purpose of the model. Further, it is important to be aware of possible shortcomings in the datasets, such as biases or errors.

The availability of datasets has significantly increased over the last few years. Google, which offers a search engine specifically to find datasets online[521], estimates that there are over 30 million datasets publicly available on the internet[522]. There are also entire communities, such as Kaggle, that are based around the sharing, analysis and competition in datasets[523].

Below is a list of a few notable public datasets, together with the task linked to them, features, labels and number of samples.

---

518 Ciarán, *supra* note 515.

519 *Active Learning Literature Survey*, Technical Report, by Burr Settles, minds.wisconsin.edu, Technical Report (University of Wisconsin-Madison Department of Computer Sciences, 2009).

520 Hannes Westermann et al, "Sentence Embeddings and High-Speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents" (2020) Legal Knowledge and Information Systems 164–173.

521 "Dataset Search", online: <https://datasetsearch.research.google.com/help>.

522 "An Analysis of Online Datasets Using Dataset Search (Published, in Part, as a Dataset)", online: *Google AI Blog* <http://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html>.

523 "Kaggle: Your Machine Learning and Data Science Community", online: <https://www.kaggle.com/>.

Name	Task	Features	Label	N samples
MNIST[524]	Recognize numbers in small images	Images of numbers	The number in the image	70K
ImageNet[525]	Recognize objects in images	Links to images	Which of ~22K synsets (i.e. concepts) are shown in an image	~14 million
Mozilla Common Voice[526]	Transcribe text to speech	Recordings of spoken voice	The spoken words in text	~7,300 hours of recordings in more than 60 languages
The Winograd schema Challenge[527]	Recognize the interpretation of ambiguous terms based on contexts	Sentences, such as: The trophy doesn't fit in the brown suitcase because it's too big. What is too big?	Answer 0: **the trophy** Answer 1: the suitcase	100

These datasets are incredibly important in the academic context when researchers want to explore and develop ways to build machine learning systems. Not only do they provide data to train the system that might otherwise be difficult to obtain. They also serve as a benchmark, where multiple different algorithms or methods can be compared to determine the best approach to a problem. Multiple large advancements have come out of competitions around datasets. In 2012, for example, the AlexNet architecture vastly improved the state-of-the-art scores for the ImageNet Large Scale Visual Recognition Challenge, a challenge based around ImageNet[528]. This was a significant breakthrough and arguably the point where the interest in deep learning methods began to really take off[529]. Today, new model architectures are today usually released together with their performance on public datasets and benchmarks[530]. If models are able to achieve state of the art on the datasets, they are often seen as the new standard in a field, and adopted by subsequent research and application.

---

524 Li Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]" (2012) 29:6 IEEE Signal Processing Magazine 141–142.

525 Jia Deng et al, "ImageNet: A Large-Scale Hierarchical Image Database" (2009) 2009 IEEE Conference on computer Vision and Pattern Recognition 248-255, doi: <10.1109/CVPR.2009.5206848>.

526 "Common Voice by Mozilla", online: <https://commonvoice.mozilla.org/>.

527 Hector J Levesque, Ernest Davis & Leora Morgenstern, *The winograd schema challenge* (2012) Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning 552-561.

528 Alex Krizhevsky, Ilya Sutskever & Geoffrey E Hinton, "ImageNet Classification with Deep Convolutional Neural Networks" in F Pereira et al, eds, *Advances in Neural Information Processing Systems 25* (Curran Associates, Inc., 2012) 1097.

529 Md Zahangir Alom et al, "The history began from alexnet: A comprehensive survey on deep learning approaches" (2018) arXiv preprint arXiv:180301164 at 11; Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 782.

530 See i.e. Ashish Vaswani et al, "Attention Is All You Need" (2017), online: <http://arxiv.org/abs/1706.03762> at 8.

However, it is important to note that performance on these datasets does not guarantee that a model will work well in the real world. It is possible to produce a model that scores very well on the datasets (such as recognizing objects in the ImageNet dataset) but fail when applied in the real world[531]. It can be hard to know whether a model has actually learned the underlying task or is taking some "shortcut" that works well on the dataset but not beyond. An example of such an occurrence is a model trained to detect whether a skin lesion is malignant[532]. The authors found that the model partially based the prediction on the existence of a ruler in the image[533]. Doctors would use a ruler when they were particularly worried about a lesion. However, if used in the wild by people taking pictures of their skin, this additional data would not be present, which might cause the model to perform worse.

Further, while models might learn well when they have access to millions of images, many tasks do not have as much training data available. Developers often do not have the resources to collect these images. Recently, a method known as transfer learning[534] has emerged that allows the training of deep learning models on some huge dataset, on some invented task such as predicting a missing word in a sentence. After training, the models have learnt a general understanding of the underlying data (such as language) and can be trained to do specific tasks using a comparatively small amount of additional data. This approach has been very successful in natural language processing[535] and image recognition[536].

3.2.3 Data preparation

We now have a dataset that we want to train the model on. The next step is to prepare this data for use in algorithms. The data we have can come in many different shapes and forms, such as a collection of images, text, video, categories or numerical values. However, most machine learning models require the data to be composed of vectors, essentially lists of numbers. How can we represent the data we have in a way that is useful to the computer?

---

531 Alexander D'Amour et al, "Underspecification Presents Challenges for Credibility in Modern Machine Learning" (2020), online: <http://arxiv.org/abs/2011.03395>.

532 Andre Esteva et al, "Dermatologist-level classification of skin cancer with deep neural networks" (2017) 542:7639 *Nature* 115–118.

533 Akhila Narla et al, "Automated Classification of Skin Lesions: From Pixels to Practice" (2018) 138:10 *Journal of Investigative Dermatology* 2108–2110.

534 Sinno Jialin Pan & Qiang Yang, "A Survey on Transfer Learning" (2010) 22:10 IEEE Trans Knowl Data Eng 1345–1359.

535 Jacob Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2019), online: <http://arxiv.org/abs/1810.04805>.

536 Hoo-Chang Shin et al, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning" (2016) 35:5 IEEE Transactions on Medical Imaging 1285–1298.

An example of the kind of data structure many machine learning models require

Sample	Feature 1	Feature 2	Feature 3	…
Sample 1	0.2	1	1	…
Sample 2	0.7	3	0	…
Sample 3	0.5	2	1	…

There are several different ways of doing this, depending on the data and requirements by the model. It can also include methods to enhance the data, such as by adding additional features or samples that can help the model. This process, known as "feature engineering", is one of the parts of machine learning that can take the most time and domain-specific knowledge, as well as creativity[537]. It is further an iterative process, with features being used to train a model, and re-adjusted based on the performance of the model[538].

(1) Exploratory data analysis

The first step of deciding how to represent data is often to perform exploratory data analysis to gain a better understanding of the distributions of the data and the different variables involved. This can help the researcher develop an intuition for how the data looks and how best to proceed[539].

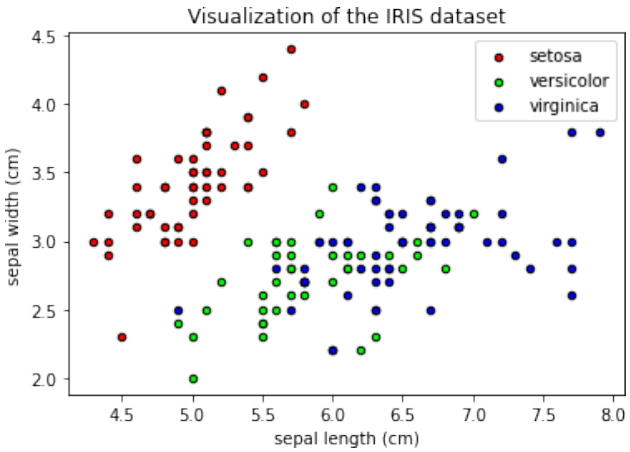


Figure 4

A visualization of the samples in the IRIS dataset, positioned by sepal size

In Figure 4, you will see a visualization performed on the IRIS dataset. Here, we map the different flower samples to a graph. Their coordinates correspond to the size of the petal of the flowers. As you can see, the setosa flowers occupies a distinct space. However, the versicolor and virginica

537 Pedro Domingos, *supra* note 265 at 84.

538 *Ibid*.

539 Stuart J Russell & Peter Norvig, *supra* note 14 at 708, 709.

variants occupy the same space. Finding a model that can reliably distinguish by the plants by just the sepal size therefore seems near impossible.

On the other hand, Figure 5 shows what happens when we instead use petal length and sepal lengths. While there is still some overlap, the three flower types occupy more or less distinct areas, making it possible for a machine learning model to learn the pattern in the data. If we had to choose 2 features, using petal length and sepal length would therefore be a better idea than using sepal length and sepal width.

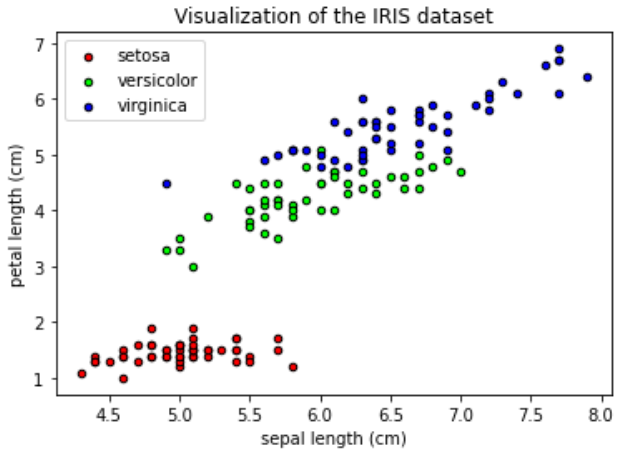


Figure 5
A visualization of the samples in the IRIS dataset, positioned by sepal and petal length

(2) Data representation

Then, the developer has to decide on a list of features to use to give to the model. In some cases, like our example of IRIS data, we can simply use the lengths of the petals as features. Some models might benefit from adjusting the data to fall into certain bounds. In other cases, it might make sense to group numbers into bins (such as ages 0-18, 18-35 etc). Encoding categorical data might require the creation of a row for each possible value of the categories, known as One-hot encoding[540]. Further, the data often has to be cleaned, by removing wrong or misleading examples or converting the data into a format easily treatable by the computer, such as extracting it from PDF documents. Beyond this, the process depends on the kind of data the model should treat, and of course the task the model aims to accomplish. Let us discuss how to represent images and text, both of which are common data sources for machine learning models.

For *images*, this process could involve turning the image into a list of pixel values. This has been done in the MNIST dataset, which deals with recognizing digits from small images. Here, the pixels have been converted into a vector of brightness, such that the model sees a list of 784 values, whether the pixel is black or white[541]. It is important that the image is of fixed size – if the input

540 *Ibid.,* at 707, 708.

541 Li Deng, *supra* note 524.

images have different dimensions, part of the preprocessing involves cropping and rescaling the image to be a consistent size.

For *text*, the process of turning data into features is a bit trickier. How can we turn words into numbers? A simple approach is to count the number of occurrences of each word in a text (such as a sentence) and using this as features. This approach is referred to as "Bag of Words"[542]. The model does not know what the words mean, but it can devise the correlation between the desired label and certain words (or combinations thereof) occurring in the text. Here is an example of a few texts being vectorized with this approach:

Text	Example	Of	Text	Another	Lots
Example of text	1	1	1	0	0
Another example	1	0	0	1	0
Lots of text	0	1	1	0	1

Thus, to the computer the phrase "example of text" looks like the vector (1,1,1,0,0). This is one of the simplest ways of turning text into vectors, but it can be tuned and enhanced in many different ways, such as by excluding so-called stop-words (such as "a", "of" etc.), by looking at combinations of words as well as single words (so called n-grams), or by weighing the words by how frequently they appear overall (since more frequent words across all documents are less likely to be important). The data could also be enhanced with additional information, such as whether a certain word is a verb or a noun. More recently, more advanced deep learning models have been developed that can represent words in a way that maintains their order and incorporates the linguistic similarity of different terms[543].

The final step of data preparation is splitting the data into training and testing data. The training data is used for training the model, while the testing data is used after to evaluate how well the model works. This is important to understand how well the model works on data it has not seen before[544]. Often, a portion of 20% or 30% of the data is set aside for testing. There are also methods that allow the use of all data by training multiple models on different portions[545].

3.2.4 Choosing and training a model

Once the data has been collected and prepared, it is time for the fun part – building the model! At this stage, we will select a suitable model for the task, and train it using the training data assembled before. This step is perhaps the most "characteristic" of the machine learning process,

---

542 Yin Zhang, Rong Jin & Zhi-Hua Zhou, "Understanding bag-of-words model: a statistical framework" (2010) 1:1–4 *International Journal of Machine Learning and Cybernetics* 43–52.

543 Jacob Devlin et al, *supra* note 535.

544 Yufeng G, *supra* note 32.

545 Stuart J Russell & Peter Norvig, *supra* note 14 at 666.

but in terms of the time required, it can sometimes pale in comparison to the steps of data collection and feature engineering[546].

Choosing which model to use can be a tricky question. Of course, the type of model is constrained by whether the task is supervised or unsupervised. Further, certain models are more suitable for tackling certain types of data. There can further be multiple trade-offs to models, including the time to train and run them, how easy the results are to explain and how well they perform[547]. However, luckily, one often does not have to choose – since the models can generally take the data in similar formats, it is often relatively easy to train multiple different types of models and see which one works best. Sometimes, the best approach might even be to combine multiple types of models to create ensemble models, that can combine the advantages of different types of models[548].

There are many more types of models than can be described in this work, and different models can further be enhanced and optimized in a huge number of ways. However, we thought it would be pertinent to look at a few well-known types of models, and explain their basic functioning, and the resulting models. While not an exhaustive description, we hope this will give an intuition over how the models work and what they can accomplish.

(1) K-nearest neighbors

One of the simpler methods of building a machine learning model is the K-nearest neighbor lookup. The idea is quite simple but can be very effective. Each sample is represented as a point in a geometric space. The location is defined by the features of the sample. To get the label for a new point, we simply obtain the *k* (for example, 5) nearest points, based on this location, and see what the majority of their labels are[549]. In a way, we are relying on the assumption that samples with similar features are also likely to share a label.

This method can work very well when there are not that many features and many samples. However, as the number of features rises, the method stops working. In spaces of hundreds or even thousands of dimensions, every point is somewhat similar to every other point. This can mean that complex problems are hard to tackle with this method[550].

---

546 Pedro Domingos, *supra* note 265 at 84.

547 Stuart J Russell & Peter Norvig, *supra* note 14 at 709.

548 Pedro Domingos, *supra* note 265 at 85, 86.

549 Stuart J Russell & Peter Norvig, *supra* note 14 at 687, 688.

550 *Ibid.,* at 688; Pedro Domingos, *supra* note 265 at 82, 83.

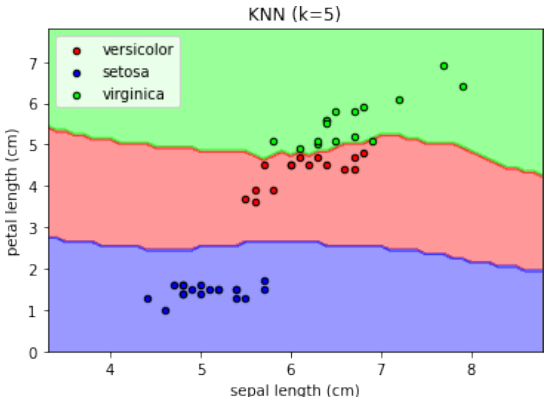
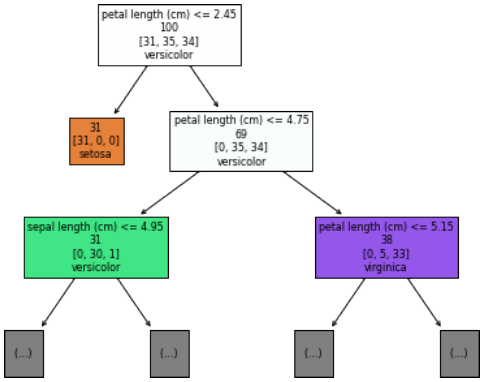


Figure 6
A visualization of the predictions by a KNN

Let us look at how the k-nearest neighbor method works with the IRIS data. For this example, we trained a K-nearest neighbor algorithm using the sklearn library[551]. We use 67% of the data to train the model, with the petal and sepal length as features (see above). The background colors show how the model would predict certain areas of the space, while the dots show samples from the test data. The model has not seen these examples before. If as many of these as possible are on the matching color field, it means that the model would generalize to these datapoints. As we can see, the model does a good job of finding the decision boundaries, and separating the different classes from each other[552].

(2) Decision trees

Another way of building a machine learning algorithm is that of decision trees. These are similar to flowcharts, where the computer runs through a number of questions about the features. At the end of this path is a prediction for the label of that sample. This might sound similar to expert systems as described in the previous chapter. However, the difference is that the tree is here not designed by a human. Rather, it is learned by the computer itself based on the training data presented. There are different algorithms to accomplish this, such as splitting the data by the most important features, and continuing to do so down each resulting branch[553].



551 F Pedregosa et al, "Scikit-learn: Machine Learning in Python" (2011) 12 *Journal of Machine Learning Research* 2825–2830.

552 Code adapted from "Plot the decision boundaries of a VotingClassifier — scikit-learn 0.24.2 documentation", online: <https://scikit-learn.org/stable/auto_examples/ensemble/plot_voting_decision_regions.html>.

553 Stuart J Russell & Peter Norvig, *supra* note 14 at 658, 659.

Decision trees are a simple method that is easy to understand, by looking at the branches and following along to understand why the algorithm chose a certain method. However, they are not the most accurate method[554].

Figure 7 shows the decision tree resulting from training on the IRIS data. As we can see, the model correctly identifies that all setosa plants have a petal length of under 2.45 cm. Most versicolor plants have a petal length of under 4.75 cm, but so do some virginica plants. Here, the model goes further to create new distinctions (not shown). As we can see, the decision boundary works well in this case. However, it is important to note that this case is a simple one, with two dimensions and few datapoints. The straight lines we see in the decision boundary graph might not be suitable for problem with hundreds or thousands of features.

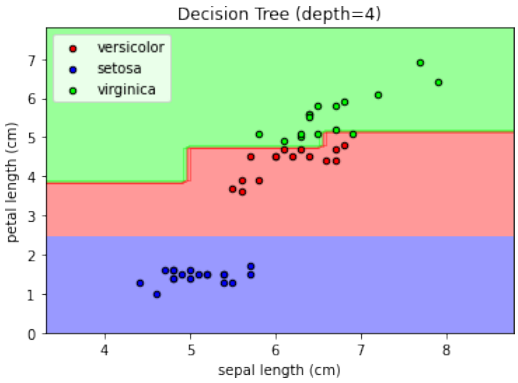


Figure 7
Visualizations of a decision tree trained on the IRIS dataset

(3) Random Forest[555]

A way to overcome the weaknesses of Decision Trees is to aggregate them into random forests. This is what is known as an ensemble model, as they consist of multiple decision trees, working in conjunction to make an accurate prediction. Each individual tree can be assigned a random selection of the training data for training (bagging) and chooses a random set of features at each split point. This randomness makes the forest more resistant to overfitting (see below)[556].

554 *Ibid.,* at 665.

555 *Ibid.,* at 697.

556 *Ibid.,* at 697, 698.

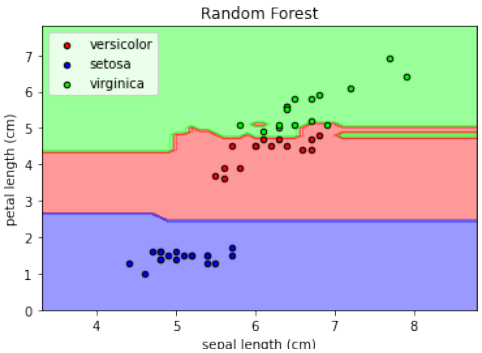


Figure 8

The decision boundaries resulting from training a Random Forest model on the IRIS dataset

Random forests are a strong model, and was for a long time favored as a one-size fits all model. They have seen uses in a wide variety of contexts[557]. Figure 8 shows the decision boundaries resulting from training a random forest model on the IRIS dataset. As we can see, the model manages to create a sophisticated decision boundary. In this case, the model seems to have overfit the data, by focusing too much on the training data resulting in a too complex decision boundary for the data. However, for real-world problems, the capacity of the model to create complex boundaries allows it to adapt to complex data to some extent.

(4) Support Vector Machines (SVM)[558]

Another popular method is Support Vector Machines. This method constructs a so-called maximum margin separator between samples in different classes. In 2-d space, it can be visualized as a line drawn between samples of different classes. This line aims to be as far away as possible to examples, which helps it perform well when applied to unseen examples. In datasets where it is not possible to separate samples with straight lines (such as data where the samples are aligned in a circle shape) SVMs make use of something called the kernel trick – by projecting the points into a higher dimension, the data is suddenly possible to separate using the separator[559].

557 *Ibid.,* at 698.

558 *Ibid.,* at 692.

559 *Ibid.,* at 693–695.

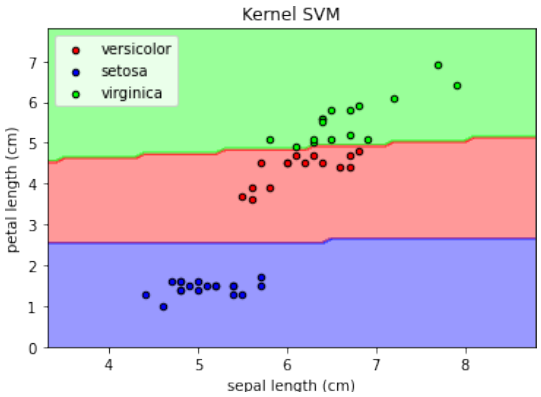


Figure 9
A visualization of the decision boundaries produced by a SVM when trained on the IRIS data

Figure 9 shows how powerful this method can be. The model has created a simple boundary, that nonetheless precisely classifies the previously unseen data into their respective classes.

(5) Neural networks and Deep Learning

Neural networks are another type of model. Over the past few years, they have grown tremendously, and are responsible for a huge number of breakthroughs across the field of machine learning. Neural networks can be used for supervised learning, but also have applications in unsupervised learning and reinforcement learning.

Unlike the models we previously described, deep learning is less reliant on feature engineering – often it is able to create its own complex representation of the input data, which allows it to capture the complexity of the real world to a much larger extent than previously used models. This makes it especially suitable for complex, unstructured data such as images, text and videos[560].

Beyond traditional supervised learning as described in the previous chapter, deep learning further allows the tackling of problems that were difficult to achieve with traditional machine learning models, such as models outputting a possible word following a sequence of words, based on a sophisticated understanding of natural language, or networks that generate images.

(a) Feedforward neural networks

The feedforward neural network is at the heart of most deep learning architectures[561]. These start with a number of "inputs", i.e. values being passed into the system. These values can be seen as the features from the dataset, see previous chapter. These inputs are then connected to

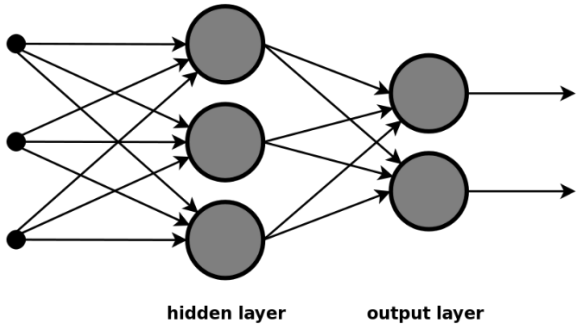
560 *Ibid.,* at 750.

561 Ian Goodfellow, Yoshua Bengio & Aaron Courville, *Deep learning*, Adaptive computation and machine learning, Cambridge, Massachusetts, The MIT Press (2016) at 168.

a number of layers inside of the network. Each layer contains a number of artificial neurons, loosely inspired by how neurons work in the brain[562].

Each neuron takes every single output of the previous layer, multiplies them by individual weights, adds them together and applies an activation function to the resulting value to obtain an output. This output is then passed on to each neuron in the next layer, and so on[563]. Each layer between the input layer and the final layer is referred to as a "hidden layer"[564].

When the data reaches the final layer of neurons, called the output layer, the data is fully "transformed", and the resulting outputs are the predictions of the network[565]. In classification, for example, there would be one neuron per class we want to predict. The neuron with the highest activation corresponds to the class the network predicts[566]. The name, "feedforward network", comes from the direction of the flow of the data, from the input through the hidden layers to the output layer[567].



So far, we have talked about a quite simple neural network. However, one of the specialties of neural networks is that they are very flexible and scalable. In this example, we used a network with a single hidden layer and a few inputs and outputs. Modern neural networks for dealing with images or texts can have tens of layers and millions or even billions of individual weights. How many layers are used, their type, and how many neurons should be used are referred to as the "architecture" of the neural network. How specifically to design the architecture of a network to best solve a certain tasks is hard to tell, and often arrived at through experimentation[568].

562 Stuart J Russell & Peter Norvig, *supra* note 14 at 750.

563 *Ibid.,* at 751–752; Ian Goodfellow, Yoshua Bengio & Aaron Courville, *supra* note 561 at 197.

564 Ian Goodfellow, Yoshua Bengio & Aaron Courville, *supra* note 561 at 169.

565 *Ibid.,* at 181.

566 Compare Stuart J Russell & Peter Norvig, *supra* note 14 at 758.

567 Ian Goodfellow, Yoshua Bengio & Aaron Courville, *supra* note 561 at 168.

568 *Ibid.,* at 198.

However, an interesting insight from the field is that "deeper" networks, containing more layers, are generally better at learning to accomplish tasks[569].

(b) Training

Of course, at first, the predictions generated by the network will be completely wrong. Usually, the weights of the neurons (i.e. the number that determine how strong the output should depend on the inputs of that neuron) is initialized randomly. Just like traditional machine learning methods, neural networks have to be *trained* to be useful.

Therefore, just like before, we give the network a number of samples from the training split of the data. We let the network calculate its prediction, which will almost surely be wrong at the beginning. We use a so-called loss function to determine how bad the prediction was[570]. We can then step backwards through the entire network, and slightly adjust the weights at each stage to be closer to being correct. This process is called backpropagation, since the error is passed back and adjusted through each layer[571].

After running through the samples a number of times, hopefully the network will have learnt a good model for the connection between the inputs and the outputs. Now, when we feed new, previously unseen data (such as the measurements of a flower not used during training) through the network, it should return the desired output for the new data.

Training a neural network can involve a lot of computation. However, using Graphics Processing Units (used in many computers to display games), the process can be sped up tremendously, allowing us to train networks with billions of parameters. Nowadays, there are even Tensor Processing Units, computers designed to train neural networks extremely quickly[572].

(c) Example

Let us go back to the example of the previous chapter, of predicting the type of a flower from the petal and sepal measurements. We would probably use the petal and sepal length and width as input features. There would be three neurons in the output layer, one each for setosa, versicolor and virginica. The network would then receive the size of the flowers, perform the calculations described above, and finally give an output of a number for each of the three neurons in the output layer. The prediction of the network would be the class corresponding to the neuron in the output layer with the highest output.

---

569 *Ibid.,* Stuart J Russell & Peter Norvig, *supra* note 14 at 769.

570 Ian Goodfellow, Yoshua Bengio & Aaron Courville, *supra* note 561 at 178.

571 Stuart J Russell & Peter Norvig, *supra* note 14 at 755.

572 *Ibid.,* at 763.

3.2.5 Evaluating the model

Once the model has been trained, the next step is to evaluate its performance. During the training process, the model only has access to the training data. At some point, depending on the configuration of the model, the training process will have completed, leaving us with a model that has attempted to learn the patterns contained in the training data.

However, this does not tell us how well the model would work in the real world. It is possible that the model was simply not able to grasp the patterns in the data, meaning that it would fail in the real world (*underfitting*)[573]. Likewise, it is possible that the model has learnt the training data by heart instead of learning the patterns behind the data (*overfitting*)[574]. We can see both of these in the example models we trained in the previous section. The decision tree learns a quite simple boundary, that separates the points by a few straight lines. This is an example of underfitting, as the model does not properly learn the decision boundaries. The random forest, on the other hand, develops a very complex decision boundary, with small spots where the prediction suddenly changes. This is likely due to the model overfitting the training data, which means that it does not properly work when applied against the test data.

We are not really interested in how well the model performs on data it has already seen, but rather on estimating on how well the model would perform in the real world. This capability is referred to as *generalization*[575]*.* Estimating how well a model generalizes can allow us to make decisions on whether the model can be deployed or whether it needs to be further improved, or whether a change to the model made it better or worse for our task.

This is where the evaluation of the model comes in. In the data preparation phase, we removed a part of the dataset, and withheld it until after the model is trained. Now, we can use the model to predict the label of this portion of the data, and see how it compares to the real labels of these samples. Since the model has never seen these examples, this procedure allows us to assess how well the model would work in the real world. Understanding the performance of the model allows us to assess whether it works well enough to be deployed into a final product, or if it needs further work. It can also help us decide between multiple models, to pick the one that works best.

There are many different metrics that can be used for assessing a machine learning model. Which of these to pick depends to a large extent on which task the model is designed to perform. Picking the right metric is very important, as picking an unsuitable metric can lead us to optimizing a system in a wrong direction. Further, which metric to pick also depends on the type of task one is aiming for, i.e. regression or classification. Below, we will discuss a few metrics for classification.

---

573 *Ibid.,* at 655.

574 *Ibid*. Pedro Domingos, *supra* note 265 at 81.

575 Pedro Domingos, *supra* note 265 at 80.

(1) Types of errors

Let us consider a sample that can be either part of a class (positive) or not (negative). Our model has to predict whether the sample is part of the class (predicted positive) or not part of the class (predicted negative). In this configuration, there could be a number of different scenarios:

	Positive	Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Ideally, all predictions would be true positives and true negatives, i.e. the predictions of the model would be correct. However, in reality, the model will often make mistakes. There are a number of metrics to measure these errors.

(2) Precision

One of the most basic metrics is that of precision. It is calculated by taking the number of true positives and dividing it by the combined number of true positives and false positives. It can be said that it calculates the number of correct predictions of a class divided by the total predictions for that class.

Precision is an important metric, as it shows the capability of the system to find positive examples. However, it does not show all information. Let us think of a situation where an algorithm has to spot infections in images. Let us say that there are 100 images in total, 10 of which contain infections. A model that spots one of the infections correctly, but predicts all other examples as negative, would here receive a perfect 1.0 in precision, since all of the examples it found where correct, Of course, the model is unlikely to be very useful, since it only finds a single of the positive examples. To measure how many of the example are found, recall is used instead.

(3) Recall

Recall measures how many of the true positives are found by the system. This is calculated by taking the number of True Positives divided by the number of true positives and false negatives. In essence, it calculates how many of the existing positive examples were found by the algorithm.

Looking at the above example, a system that correctly identifies only one out of ten infections would receive a recall of 0.1. Depending on the use case, this might be a more informative score. However, a system that would classify all images as containing infections would receive a perfect recall score of 1.0, since it would capture all positive examples. Of course, the precision here would be very low.

(4) Recall, Precision and F1-score

So, what is more important, precision or recall? There is no answer to this – it depends on the application. In the example above, our algorithm was looking for infections. Here it might be advantageous to aim for high recall, as missing an infection could have bad consequences, while

a doctor would verify predicted infections to make sure that treatment is necessary. On the other hand, in some instances precision might be more important. This could be the case, for example, if the prediction immediately triggers a reaction that could adversely affect humans, such as deleting an email message that the model considers to be spam.

Since both recall and precision are important, but do not tell the whole picture, the F1-score is often used as a single metric of performance. This is a type of average between precision and recall, designed to give a picture of overall performance.

(5) Other metrics

There are also many more metrics, focusing on different use-cases. For example, the Matthews Correlation Coefficient (MCC) is a way to measure classification performance that some believe is more informative than the previously discussed metrics[576]. Hamming distance or Jaccard distance can be used to evaluate datasets where each item can have multiple labels, such as images containing multiple objects[577].

For tasks that are not classification, the metrics used are also different. For regression, for example, mean squared error (MSE) can be used. This measures the difference between the predicted value and the real value[578]. For some tasks there might be even more specialized metrics - the BLEU-score, for example, is a metric used for assessing automatic translation of documents[579].

(6) Potential issues with metrics

It can sometimes be tricky to evaluate models based on a single metric. As described above, one such issue might be cases where the dataset is very unbalanced, i.e. certain labels are much more common than others. Further, a high score on a metric might hide other underlying problems, such as the model only working well on a certain subgroup of the data[580]. For example, if algorithms for facial recognition are evaluated on datasets that contain very few images of dark-skinned females, the model might score highly on the metrics while underperforming on these images[581].

---

576 Davide Chicco & Giuseppe Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (2020) 21:1 *BMC Genomics* 1–13.

577 Valentin Gjorgjioski, Dragi Kocev & Sašo Džeroski, "Comparison of Distances for Multi-Label Classification with PCTs" (2021).

578 "Mean Squared Error (MSE)", online: <https://www.probabilitycourse.com/chapter9/9_1_5_mean_squared_error_MSE.php>.

579 Kishore Papineni et al, *Bleu: A Method for Automatic Evaluation of Machine Translation*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics (2002).

580 Harini Suresh & John V Guttag, *supra* note 32 at 6.

581 *Ibid*.

Another issue is that the test data might focus certain samples. In a dataset for face recognition that consists mostly of faces of people of certain ethnicities, for example, the model would be able to perform very well on the metrics by just learning to recognize faces of that ethnicity. An F1-score of 0.90 might hide the fact that the model performs much worse on certain ethnicities than others.

More broadly, the process relies on the fact that the ability of the model to generalize can be assessed by splitting a part of the dataset from the rest. Of course, this only works if the data in the that the model will be working with in real life closely corresponds to the testing data. If the data is different from the testing data, the model might score very high on any metric, but still completely fail at carrying out its task when deployed[582].

(7) Example

In our above example, we want to evaluate how well our model does in distinguishing types of flowers. We trained the model on 70% of the data. We then use the remaining 30% of the data to evaluate the performance of the model. Let us look at the performance of our classifiers trained above when evaluated on the remaining data.

Classifier	Avg Precision	Avg Recall	Avg f1-score
Decision Tree	0.98	0.98	0.98
KNN	0.98	0.98	0.98
SVM	1.00	1.00	1.00
Random Forest	0.98	0.98	0.98

The table shows the precision, recall and F1-score for the classifiers. Since there are three classes, we average the individual scores for each class to receive an overall score. As you can see, the scores are very high. This is due to the IRIS dataset essentially being a toy dataset, rarely in the real world will prediction scores be this high. However, we can see that the SVM model performs slightly better than the other three models, that classify one of the virginica colors as versicolor. Can you see where the mistake occurs in the graphs?

3.2.6 Deployment of model

Once the model has been evaluated and deemed to work well enough for the desired task, it is time to deploy it into the real world. It can be integrated into many different places, such as an app, a website or even a self-driving car. This can be a challenging process – machine learning models rely on a lot of data and heavy computation. Making sure these work well when put on a

582 *Ibid*. Joy Buolamwini & Timnit Gebru, *supra* note 503.

server can be very tricky. Estimates indicate that most data science projects never make it into production[583].

The work does not stop here. While the model may have worked well on the testing data, this is no guarantee that it will work equally well in the real world. Despite being trained on millions of examples, the data the system recognizes in the real world might differ from the data used in training, leading to unpredictable results. Monitoring the system to make sure that it runs well, and maintains high performance, is an important step of the process[584].

Further, while the trained model is usually static, the world around it keeps changing. If the type of data that the model sees changes, the predictions will not work well. Retraining the model with new data is also important to make sure that the model stays up to data. the model is, in most cases, static, while the world is constantly adapting and changing.

Beyond these issues, it is important to note that deployed models are often part of a larger system, involving other computer systems and humans. Even a model that works very well when evaluated might cause issues when it comes into contact with humans, who are subject to or act based on the outputs of the system. This is a growing area of investigation. Many researchers are investigating how to make models that are explainable, so that people can understand why an output is given and incorporate this into their decision-making procedure[585].

Therefore, it is very important to evaluate the entire resulting system, rather than just the machine learning component. It is also important to be mindful of how a model was trained, and the how it is used in practice. A model trained to predict the likelihood of crime, for example, could be harmful when used for other purposes, such as determining the length of a prison sentence[586].

3.3 Application areas

Aiming to describe the application area of machine learning system is a bit like aiming to describe the application areas of software. There is an enormous amount of potential applications and describing all of them would be unfeasible. Instead, we will describe a few notable applications and examples of how these were built with the previously described machine learning systems.

---

583 Cristiano Breuel, "ML Ops: Machine Learning as an Engineering Discipline", (3 January 2020), online: *Medium* <https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f>; Rising Odegua, "How to put machine learning models into production", (12 October 2020), online: *Stack Overflow Blog* <https://stackoverflow.blog/2020/10/12/how-to-put-machine-learning-models-into-production/>.

584 Stuart J Russell & Peter Norvig, *supra* note 14 at 712.

585 Stuart J Russell, Peter Norvig & Ernest Davis, *supra* note 361 at 711, 712.

586 Harini Suresh & John V Guttag, *supra* note 32 at 6, 7.

Examples of neural models in the real world[587]

3.3.1 Spam detection

One of the traditional applications for machine learning is for spam detection. Estimates indicate that hundreds of billions of spam emails are sent each day, compared to tens of billions of legitimate emails[588].

Luckily, this is a problem where machine learning can help out. Many of the large email providers, such as Gmail, Yahoo and Outlook, implement systems that automatically classify emails as spam or legitimate. To do this, they often have sophisticated rules that look at the sender of the email, the subject and text, whether the email contains links etc. These models can be continuously refined based upon whether users mark emails as spam[589].

3.3.2 Computer Vision

Recognizing images has long been a very important task in artificial intelligence and robotics. Images, in pixel form, can be tricky to deal with. There is usually a lot of data, since millions of pixels make up a single image. Further, each pixel does not describe a single concept, such as the number of windows on a house. Rather, each pixel makes up the image in connection with the other pixels in the image. The objects in an image are further independent from the location – a balloon is a balloon, no matter where in an image it appears[590].

A solution to this problem is Convolutional Neural Networks (CNN). They work by sliding a special network known as a "kernel" over the image. Each kernel is trained to recognize a certain feature, such as a line or a curve in the image. Since the kernel is slid over the image, it does not matter where in the image an object occurs – the CNN can recognize it. Further, it decreased the number of parameters required[591].

The first big breakthrough in Convolutional Neural Networks was AlexNet, which managed to win the ImageNet Large Scale Visual Recognition Challenge, a challenge centered around the recognition of objects in millions of images[592]. Since then, a large number of improvements have

---

587 Stuart J Russell & Peter Norvig, *supra* note 14 at 782.

588 "Email and Spam Data || Cisco Talos Intelligence Group - Comprehensive Threat Intelligence", online: <https://talosintelligence.com/reputation_center/email_rep>.

589 Emmanuel Gbenga Dada et al, "Machine learning for email spam filtering: review, approaches and open research problems" (2019) 5:6 Heliyon e01802.

590 Stuart J Russell & Peter Norvig, *supra* note 14 at 760.

591 *Ibid.,* at 760–763.

592 Alex Krizhevsky, Ilya Sutskever & Geoffrey E Hinton, *supra* note 528.

been made. For example, there are now images able to recognize the position of objects as well as their type[593], or even precisely segment an image into different portions[594].

Even more recently, the Transformer architecture (described below as used in natural language processing) has been applied very successfully in computer vision tasks[595].

The power of neural networks for recognizing images has made a number of interesting applications possible. For example, a number of large companies, such as Tesla[596] and Waymo[597], have been working on using the image recognition prowess of the models to create self-driving cars[598].

3.3.3 Natural Language Processing

Another key task for machine learning is understanding natural language. This is another domain where deep neural networks have recently made massive advantages to the state of the art.

One of the key features of a feedforward neural network is that data always moved forward. In the case of natural language processing, however, it would be useful to feed data back into the network after a single prediction is done. When translating a sentence, for example, the translation for each individual word does not only depend on the word itself, but also the words that came before.

This is where recurrent neural networks (RNNs) come in. In addition to input data for the current step, the RNN also receives its own output for the previous time step. This allows it to have a sort of "memory", being able to refer back to previous data and predictions[599]. Thus, the network can look at one word at a time in a sentence, but still incorporate the previous context.

---

593 Joseph Redmon & Ali Farhadi, "YOLOv3: An Incremental Improvement" (2018), doi: <https://doi.org/10.48550/arXiv.1804.02767>.

594 Liuyuan Deng et al, *CNN based semantic segmentation for urban traffic scenes using fisheye camera* (2017) IEEE Intelligent Vehicles Symposium (IV), doi: <10.1109/IVS.2017.7995725>.

595 Alexey Dosovitskiy et al, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2021), online: <http://arxiv.org/abs/2010.11929>.

596 "Autopilot AI", online: *Tesla* <https://www.tesla.com/autopilotAI>.

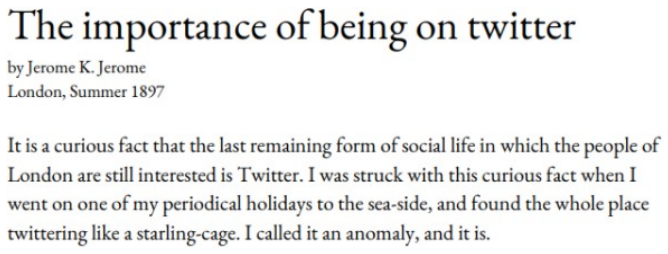
597 Andrew J Hawkins, "Inside the lab where Waymo is building the brains for its driverless cars", (9 May 2018), online: *The Verge* <https://www.theverge.com/2018/5/9/17307156/google-waymo-driverless-cars-deep-learning-neural-net-interview>.

598 Qing Rao & Jelena Frtunikj, *Deep learning for self-driving cars: chances and challenges*, New York, NY, USA, Association for Computing Machinery (2018).

599 Stuart J Russell & Peter Norvig, *supra* note 14 at 772, 773.

Since the introduction of the recurrent neural network, there have been a number of improved versions. The Long short-term memory (LSTM) Network, for example, is able to remember data for longer than traditional RNNs[600].

The current state-of-the art architecture for this kind of tasks is called the Transformer. It uses a mechanism called attention to decide which words to pay attention to when predicting something at the current step[601]. This results in more sophisticated models, able to pay attention to words that come hundreds of words before the current word. This has resulted in a number of very sophisticated language models, such as BERT, which can solve a number of language specific tasks with very little additional training[602]. It has also led to the creation of GPT-3, an enormous model with 175 billion parameters that can write quite coherent and thematically relevant texts based upon a short prompt[603]. The system is currently being used in hundreds of applications,[604], and has been found to be able to perform well in a wide variety of contexts, including translation[605], poetry and creative fiction[606], automatic generation of textual game worlds[607] and the automatic generation of computer code, based on descriptions of what the code should achieve[608].



The importance of being on twitter

by Jerome K. Jerome
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is.

Figure 10
The beginning of an article written entirely by GPT-3.
It is provided with the title, author name and the first word "It" and generated the rest[609]

600 *Ibid.,* at 775.

601 Ashish Vaswani et al, *supra* note 530.

602 Jacob Devlin et al, *supra* note 535.

603 Tom B Brown et al, *supra* note 355.

604 "GPT-3 Powers the Next Generation of Apps", (25 March 2021), online: *OpenAI* <https://openai.com/blog/gpt-3-apps/>.

605 Tom B Brown et al, *supra* note 355 at 15.

606 Gwern Branwen, *supra* note 355.

607 Adam Nieri July 18 et al, "AI-written Scenario for Dungeons & Dragons Is Actually Quite Good", (18 July 2020), online: *Mind Matters* <https://mindmatters.ai/2020/07/ai-written-scenario-for-dungeons-dragons-is-actually-quite-good/>.

608 Ram Sagar, "OpenAI's GPT-3 Can Now Generate The Code For You", (20 July 2020), online: *Analytics India Magazine* <https://analyticsindiamag.com/open-ai-gpt-3-code-generator-app-building/>.

609 Mario Klingemann (@quasimondo), "Another attempt at a longer piece. An imaginary Jerome K. Jerome writes about Twitter. All I seeded was the title, the author's name and the first 'It', the rest is done by #gpt3

3.3.4 Generating media

Deep neural networks are, as we have seen, very strong in recognizing images. But what about generating images?

For this task, another architecture has been developed. It relies on two networks that are trained in conjunction. One of the networks is the "generator", which aims to create convincing fake versions of an image. The other network is the "discriminator", which aims to determine whether an image is real or fake. By training these two networks together, the generator network will eventually learn to create images that resemble real images but are entirely made up[610].

Figure 11
A few images of made-up faces produced by a GAN-network[611]

Another application of deep neural networks used for generation is so-called deepfakes. Using GANs[612] or another technology known as variational auto-encoders[613], these models are able to replace one person in a video with another person, using a few images of the second person. This has a number of creative uses, such as adding the actor Nicolas Cage to any movie scene[614].

Here is the full-length version as a PDF: https://t.co/d2gpmlZ1T5 https://t.co/1N0lNoC1eZ" (18 July 2020, 11:25 AM), online: <https://twitter.com/quasimondo/status/1284509525500989445>.

610 Ian J Goodfellow et al, "Generative Adversarial Networks" (2014), online: <http://arxiv.org/abs/1406.2661>; Stuart J Russell & Peter Norvig, *supra* note 14 at 780.

611 Tero Karras, Samuli Laine & Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks" (2019), online: <http://arxiv.org/abs/1812.04948>.

612 Egor Zakharov et al, "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models" (2019), online: <http://arxiv.org/abs/1905.08233>.

613 Diederik P Kingma & Max Welling, "Auto-Encoding Variational Bayes" (2014), online: <http://arxiv.org/abs/1312.6114>.

614 Usersub, "Nick Cage DeepFakes Movie Compilation", online: <https://www.youtube.com/watch?time_continue=25&v=BU9YAHigNx8>.

However, it can also have troubling uses – the networks have been used to create pornographic videos of famous actors or other subjects, and to create fake videos of politicians making certain statements[615].

3.3.5 Deep Reinforcement Learning

Above we described reinforcement learning, where an agent aims to accomplish certain tasks in an environment, such as winning a board game, moving forward without falling over or achieving a high score on a computer game, such as tetris[616].

By using neural networks as agents in reinforcement learning situations, major breakthroughs have been achieved, and agents able to learn and carry out sophisticated behavior have been created.

Perhaps the most famous use of deep reinforcement learning is AlphaGo, created by Deepmind, a part of Google[617]. In 2016, it astonished the world by beating the one of the champions of the game Go, Lee Sedol. This came as a great surprise to many, as Go was considered as the game that would be the most difficult for computers to become proficient in, due to its massive number of possible moves and different game states. Move 37 in the second game in particular was considered an incredible and surprising move that even the highest-level experts were not able to predict[618].

AlphaGo was trained on a large number of games by human champions, and then made to play against itself for millions of iterations[619]. Later, the researchers at DeepMind demonstrated AlphaGo Zero, a version that was able to surpass the previous systems without relying on any games played by humans, becoming arguably the best player of Go in the world, without relying on any human games[620]. Further versions have been shown that the system can learn other games, such as Chess and Shogi, using the same methodology[621].

---

615 Robert Chesney & Danielle Keats, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, SSRN Scholarly Paper, by Robert Chesney & Danielle Keats Citron, Rochester, NY, Social Science Research Network (2018); Rebecca Ruiz, "Deepfakes are about to make revenge porn so much worse", online: *Mashable* <https://mashable.com/article/deepfakes-revenge-porn-domestic-violence/>.

616 See above under "Reinforcement Learning" for a more in-depth explanation.

617 David Silver et al, "Mastering the game of Go with deep neural networks and tree search" (2016) 529:7587 Nature 484–489.

618 Cade Metz, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future" *Wired* (16 March 2016), online: <https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/>.

619 David Silver et al, *supra* note 617.

620 David Silver et al, "Mastering the game of Go without human knowledge" (2017) 550:7676 Nature 354–359 Primary_atype: Researchpublisher: Nature Publishing GroupSubject_term: Computational science;Computer science;RewardSubject_term_id: computational-science;computer-science;reward.

621 David Silver et al, *supra* note 617.

Another notable actor in the field is OpenAI. The organization has created a system called the OpenAI Gym which allows different researchers to test their algorithms in common environments. Many different environments are included, ranging from an environment that teaches an agent to balance a pole on a cart, to an environment that teaches an agent to walk on two legs, to an environment where agents have to play classic Atari games[622]. OpenAI has further developed its own solutions to many problems. For example, it developed an environment for agents to play tag cooperatively. The agents were able to figure out many smart strategies involving tools placed on the playfield, such as blocking the door to protect themselves from catchers. Some of these strategies were even unanticipated by the creators[623]. OpenAI has further trained systems to play the popular computer games Starcraft II[624] and DotA 2[625]. In both of these titles, the agents managed to beat many top human players.

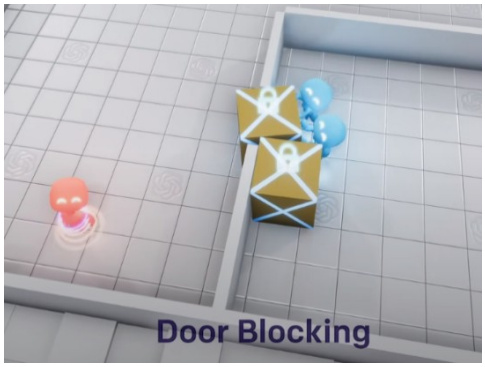


Figure 12
An example of a strategy used by agents in a catch environment

3.4 Discussion

Machine Learning is undoubtedly an incredibly powerful methodology, that has lead to incredible advances in many areas of artificial intelligence. However, it also has some shortcomings, many of which are being actively worked on in the research field. Being aware of these advantages and shortcomings is essential in utilizing machine learning and deep learning to its fullest potential.

---

622 OpenAI, "Gym: A toolkit for developing and comparing reinforcement learning algorithms", online: <https://gym.openai.com>.

623 "Emergent Tool Use from Multi-Agent Interaction", (17 September 2019), online: *OpenAI* <https://openai.com/blog/emergent-tool-use/>.

624 Oriol Vinyals et al, *supra* note 489.

625 "OpenAI Five Defeats Dota 2 World Champions", (15 April 2019), online: *OpenAI* <https://openai.com/blog/openai-five-defeats-dota-2-world-champions/>.

3.4.1 Advantages

(1) Less reliant on human knowledge

A huge advantage of machine learning when compared to symbolic systems is the fact that they rely less on human operators to achieve their target. In the symbolic systems, the rules on how to solve a problem had to be created and encoded explicitly by human operators. As we saw, this can be very time-consuming, as hundreds or thousands of rules might be required. As in many cases expert knowledge is required to encode the rules, it can also be an expensive operation. Further, while this way of encoding data works well for some cases where knowledge is explicit (such as in the medical and legal domains), it fails when data is more implicit and intuitive, such as how to recognize an image or win a match of chess.

Machine learning sidesteps these issues to some extent. Instead of encoding the rules on how to solve a problem, the machine instead requires previous data with features and a label. The machine then autonomously figures out the rules on how to solve the problem, which hopefully also works on future examples. This works very well, since giving examples of a problem is often more easy than describing how to solve the problem in a computer-comprehensible format. Imagine, for example, labelling fruits in images as compared to precisely describing the rules on how to recognize different fruits precisely. In some cases, the data might even be available to obtain from accessible sources, such as image descriptions available on the internet, or training using massive corpora of publicly available text[626].

Further, this approach can work with implicit knowledge. Since the machine learning model learns how to solve a task itself, it can hopefully learn the intuitive parts of how to solve the problem autonomously. This also means that the model is not limited by the approach and performance of the human that creates it – as we have seen, machine learning systems are able to surpass human performance in some limited domains, such as playing games including Go and Chess.

Of course, this does not mean that no expert input is required when building a machine learning system. First, of course, the labelling of the data can be very time-consuming and require a lot of human work. Further, throughout the construction of the machine learning model, a number of choices have to be taken to make the model work, including cleaning and preparing the data and building and evaluating the model. These are very important steps, and making the wrong choice can lead to a model that does not work, or worse, causes harm. Finally, integrating a model into a system comprised of humans can be a significant undertaking and require a lot of effort and domain expertise to be done successfully.

626 e.g. “Common Crawl”, online: <https://commoncrawl.org/>.

(2) Sophisticated models

Another advantage of the machine learning approach is the sheer level of sophistication these models can achieve. Symbolic systems can be constrained by the number of rules they can feasibly use to arrive at a result, perhaps due to the time required to enter these rules into the system. Further, the rules are often binary, and can thus have trouble dealing with vague data or uncertainties, making it difficult to generalize to unseen cases.

Machine learning models, on the other hand, are often probabilistic by their nature. Neural networks, for example, output a probability of their prediction, based on the combination of weights of potentially billions of neurons. This function can encode very complex functions, relying on many different signals to come to the right solution. Andrey Karpathy refers to neural networks as "Software 2.0" due to their ability to learn any function[627].

This sophistication means that many types of data that were previously out of reach can now be successfully treated with artificial intelligence. Images, sound and video are examples of data types that are very common, but also unstructured, and thus very difficult to treat with symbolic systems. Machine learning, and especially deep learning, has opened the door to models that can work with this data, which opens up massive data troves for use as training data or analysis. Further, more and more data is collected in society, whether through Internet-Of-Things sensors, self-driving cars, medical genome sequencing or media shared on social media. With machine learning, we have powerful tools to analyze this data and use it to make relevant decisions in the future.

As we have seen, the models can learn functions that are beyond the scope of human reasoning. AI is quite different from humans in the way it treats data. For us, it might be very difficult to look at billions of rows of data, with millions of columns. Deep learning models are able to do this, and often find important patterns in the data. As such, problems that may be out of scope for humans may be open to analysis using machine learning. An example might be AlphaFold 2, which managed to beat the previous state of the art of predicting how proteins will fold, a fundamental and crucial biological reaction[628].

3.4.2 Disadvantages

(1) Lack of explainability

Symbolic system, by virtue of the rules being written by people and traceable through the system. Some machine learning models are equally simple to explain. Decision trees, for example, learn binary rules to classify a sample. This process can be followed and understood, although it might not always be obvious why a certain decision criterion was chosen.

---

627 Andrej Karpathy, *supra* note 352.

628 "AlphaFold: a solution to a 50-year-old grand challenge in biology", online: *Deepmind* <https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>.

As models become more and more sophisticated, often their explainability suffers as a result. This is especially the case for large neural networks. The millions or billions of parameters in the networks interact in complex and entangled ways, over multiple layers with concepts that are unlikely to be decipherable for humans. All this means that tracing a decision over the different layers to understand how and why it was taken can be incredibly challenging.

In many fields, having an explanation is a requirement for the implementation of artificial intelligence into systems. This is the case, for example, in the judicial system. Only if a decision can be understood by the individual the question is facing can the individual defend themselves and appeal the decision. The same goes for administrative decisions and decisions around online content moderation – the trust for these systems will be largely diminished if the decisions cannot be understood. Regulators are reacting to this. In the European Unions data protection legislation, there is arguably a provision for a right to an explanation, where certain decisions have to be explained to the affected user[629]. Even in other, more mundane, decisions, having an explanation would be beneficial to understand what the system is doing and why certain errors occur, and how to correct these.

Many researchers are investigating ways of making neural networks more explainable. This is not always easy, however. What does an explanation mean? To some extent, neural networks can be explained by referring to the activations in the neurons when a certain input runs through the system. However, this is unlikely to be useful. Many approaches are being investigated, such as the posing of counterfactuals (if you would earn this much, your application for a loan would be granted)[630].

(2) The Alignment Problem

Another problem that faces developers of machine learning systems is the alignment problem. It posits that the solutions found by artificial intelligence systems are not necessarily aligned with human expectations and values. Developers provide AI systems with a task they want solved. The AI system will then find a way to accomplish this task. However, while there might be an obviously "right" way for humans to solve a certain task, the AI system has no idea what expectations the human has on it. This can prevent the system from generalizing to the real world, or even introduce harmful biases into the way the system reasons[631].

---

629 Merle Temme, "Algorithms and Transparency in View of the New General Data Protection Regulation" (2017) 3:4 *European Data Protection Law Review* 473–485.

630 Sandra Wachter, Brent Mittelstadt & Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (2018) *Harvard Journal of Law & Technolog*y, doi: <https://doi.org/10.48550/arXiv.1711.00399>.

631 David A Shaywitz, "'The Alignment Problem' Review: When Machines Miss the Point", *Wall Street Journal* (25 October 2020), online: <https://www.wsj.com/articles/the-alignment-problem-review-when-machines-miss-the-point-11603659140>; Eliezer Yudkowsky, "The AI alignment problem: why it is hard, and where to start" (2016) Symbolic Systems Distinguished Speaker.

Let us consider a few examples of this phenomenon. Researchers trained a model to distinguish between huskies and wolves. The system seemed to work well when tried against testing data, but failed to work well in the real world. What was going on? When the researchers analyzed the functioning of the neural network, they discovered that most of the training data had a particular bias – images with a wolf were much more likely to have snow in the background[632]. Since this is much easier to spot than the subtle difference between huskies and wolves, the network learned to tell them apart by the presence of snow in the image. On the testing data, this approach would work well, but in reality, it would completely fail, since the background is of course unrelated to the species of the animal captured on a camera. As we can see here, the network solved the task, but completely differently than the researchers had expected, and in a much less useful way.

This kind of mistake can also lead to harmful biases entering the system. There have been several such examples, such as high-paying jobs being more likely to be shown to men rather than women[633], a recruiting tool was found to specifically discriminate against women[634], and pre-trial sentencing tools were found to discriminate against certain minorities[635]. It is likely that these systems were trained on historical data to perform some useful task. Instead of identifying the capabilities expected and intended for by the researchers, the systems instead developed a problematic and biased understanding of the data.

Similar problems can be seen in Reinforcement Learning. Designing the reward function that guides the system towards the desired behavior can be very tricky, since the researcher may have trouble understanding how exactly the system will aim to obtain the reward. There have been many examples of systems learning to "trick" the system into rewarding points without actually accomplishing the task, such as a boat driving around in circles collecting powerups rather than finishing the race[636].

---

632 Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, « "Why Should I Trust You?" Explaining the Predictions of Any Classifier », *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016, en ligne : <arxiv.org/pdf/1602.04938v1.pdf>.

633 Julia Carpenter, "Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you." (6 July 2015), online: *Washington Post* <https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/>.

634 Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters* (10 October 2018), online: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

635 Julia Angwin et al, "Machine Bias" (23 May 2016), online: *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>; compare Sam Corbett-Davies et al, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.", *Washington Post* (17 October 2016), online: <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>.

636 "Specification gaming", *supra* note 497.

(3) Common sense, causality and embodiment

A lot of speculation has taken hold over whether machine learning systems are approaching the level of human intelligence. Dumouchel sees this question as misguided, arguing that even human intelligence cannot be defined or measured in total. Rather, it consists of a number of related skills. Therefore, he argues, comparing human intelligence to AI is not a useful comparison. AI should rather be seen as a tool, that can surpass human capabilities in some cases but not others[637].

Despite the comparison not being apt in all cases, there seems to be a few distinguishing features that sets human intelligence apart from the artificial counterpart. One frequently discussed difference is the notion of "common sense". This is somewhat hard to define, but humans seem to have an innate ability to integrate their experiences and sensory perceptions with common sense, a large body of understanding of how the world works. This gives us a high-level ability to understand when an outcome is absurd, and rethink our understanding. Further, it allows us to use experiences that we have previously learnt and apply these to new situations. An example from expert systems that illustrates the issue with a lack of common sense is a system that was programmed to … but was not able to surmise that a minor should not be considered under this regime. However, even deep learning often faces issues with this. For example, an uber self-driving car recognized a person crossing the road. However, since the person did not cross the road at a designated crossing, the system failed to predict their path, and ended up hitting the person[638]. Likewise, once of the largest and most sophisticated deep learning language generation models, GPT-3, was tested in the context of therapy sessions, and told the experimental user to consider killing themselves[639]. Integrating common sense, in order to minimize such blatant implausible results, might therefore be very important if models should be exposed to real world users.

Perhaps linked to this is the notion of causality. Yoshua Bengio sees this as one of the key remaining next step in creating stronger AI systems[640]. The idea here is that humans have some notion of what caused something else. Thus, we have some (perhaps imperfect) notion of hour things are connected, and why they occur. Current artificial intelligence systems, on the other hand, work purely by statistical correlation. Instead of having a theory of why something occurs, or why a certain object should be labeled in a certain way, the models rely on huge amounts of data to identify correlations. This distinction can be seen in practice. If a child is told that a certain

---

637 Paul Dumouchel, *supra* note 35.

638 Katyanna Quach, "Remember the Uber self-driving car that killed a woman crossing the street? The AI had no clue about jaywalkers", online: <https://www.theregister.com/2019/11/06/uber_self_driving_car_death/>.

639 Kevin Riera, Anne-Laure Rousseau MD, Clément Baudelaire, "Doctor GPT-3: hype or reality? - Nabla" (27 October 2020), online: <https://nabla.com/blog/gpt-3/>.

640 Payal Dhar, "Understanding Causality Is the Next Challenge for Machine Learning - IEEE Spectrum" (29 October 2020), online: *IEEE Spectrum: Technology, Engineering, and Science News* <https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/understanding-causality-is-the-next-challenge-for-machine-learning>.

object is a tree, it is likely to remember that tree, and also identify other trees, even if they look different. AI systems, on the other hand, need tens to thousands of images of trees to be able to identify them, and even then are likely to fail when applied on images that are radically different from the images it was taught on.

Paul Dumouchel posits that a part of the explanation for this distinction might be the notion of embodiment. Humans intelligence is extricably linked to an individual, manifested in a body. How we reason about objects is therefore also conditioned by this physical manifestation, our own goals etc.[641]. A door handle, for example, cannot be considered without our ability to open one with our hand. Further, a door handle takes on another meaning when we want to go through the door. All of this influences how we recognize, perceive and choose to act on objects in the world. Artificial systems, on the other hand, are not embodied but rather mathematical systems residing on some computer. They do not have their individual point of view as relating to a body, and none of the same tools or desires as us. Therefore, the argument goes, it is impossible for these systems to understand and interact with the world as we do.

3.5 Conclusion

In this section, we discussed machine learning. This is a popular way of building artificial intelligence systems that relies on using algorithms that autonomously learn from data. Recently, deep learning which rely on deeply interconnected artificial neurons, can build a sophisticated, multi-layered understanding of data from the real world. This makes them especially useful for unstructured data, like images, text, video and speech.

Machine learning and deep learning have had a huge impact on artificial intelligence research, and are today the dominating form of research in AI. They have achieved state of the art results in many fields. Further, new and innovative architectures are discovered on a regular basis, increasing the performance of the networks on certain tasks or enabling other improvements.

Despite these rapid advances and accomplishments, there are many areas left to research in the field. Neural networks face issues relating to common sense and causality, which makes them difficult to employ in certain fields. The next section will focus on current avenues to further advance the research in artificial intelligence, and how they compare to the current paradigm.

Further readings :

Dumouchel, Paul, “Intelligence, Artificial and Otherwise” (2019) 24:2 forphil 241–258.

Ekbia, H R, “Artificial Dreams” (2008) 418.

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton, “Deep learning” (2015) 521:7553 Nature 436–444.

Russell, Stuart J & Peter Norvig, *Artificial intelligence: a modern approach*, fourth edition ed, Pearson series in artificial intelligence (Hoboken: Pearson, 2021).

[641] Paul Dumouchel, *supra* note 35 at 246–252.

Section 2 - Vers une intelligence artificielle (IA) forte ?

Points saillants :

À quand le passage de l'« intelligence » à la « conscience artificielle » ? Dans notre quête incessante pour saisir l'« esprit dans la machine », la route vers l'intelligence artificielle (IA) forte serait balisée principalement par les avancées (attendues) suivantes dans la direction asymptotique de la singularité technologique :

- optimiser les résultats prédictifs en dépassant les limites imposées par la disponibilité des jeux de données ou la puissance computationnelle;
- optimiser la prise en compte du contexte dans l'apprentissage pour un traitement plus efficace des jeux de données;
- renforcer la transparence des algorithmes et l'intelligibilité des modèles d'apprentissage;
- optimiser les modèles d'apprentissage profond par un couplage toujours plus étroit et précis des mécanismes neurophysiologiques complexes sous-tendant notre apprentissage cortical;
- exploiter à fond les possibilités que présente l'apprentissage incarné dans le développement de la robotique cognitive;
- développer une pensée cognitive réflexive artificielle pour tendre vers une intelligence artificielle polyvalente et métacognitive.

***

À quand le passage de l'« intelligence » à la « conscience artificielle » ? Jusqu'où ira cet « esprit dans la machine »[642] ? Que nous ayons grandi avec les robots d'Asimov ou les Mécas de Spielberg, le mythe renouvelé du pantin sorti de la bûche brute semble annoncer ce jour J où l'« enfant-robot » que l'humain aura programmé de ses mains sera doué de vie après avoir « fait ses preuves de bravoure, de franchise, de loyauté et d'obéissance ».

À bien des égards, l'ambiguïté de l'expression « intelligence artificielle » serait aussi imputable à l'incertitude sémantique rattachée au concept de l'« intelligence » (renvoi au chapitre 1). Dans ses manifestations protéiformes, l'« intelligence naturelle » – tant individuelle que collective ou

642 Expression reprise de John Haugeland, L'esprit dans la machine. Fondements de l'intelligence artificielle, Odile Jacob, 1989.

de groupe[643], qu'elle se manifeste tant chez les humains que les animaux[644], voire dans le monde végétal[645] – reste une notion vague; ses embranchements extrêmement complexes n'ont d'égal que la diversité des points de vue multidisciplinaires qui tentent, sans grands succès, de l'asseoir sur des fondements théoriques plus solides[646]. Alors qu'il est plus ardu de justifier rationnellement ce qui distingue un « comportement intelligent » de ce qui l'est moins, il s'avère tout aussi certain que l'intelligence ne se contente pas d'un mode opératoire strictement quantitatif. Or, il semble que la cognition – notre cognition – artificiellement construite repose fondamentalement sur le paradigme probabiliste, lequel nécessite « bêtement » la collecte d'un volume colossal de données devant être traitées à l'aide d'une puissance de calcul tendant vers l'infini.

---

643 Parmi une littérature abondante sur le sujet, voir entre autres : Chao Yu, Yueting Chai et Yi Liu, « Literature review on collective intelligence : a crowd science perspective » (2018) 2:3 International Journal of Crowd Science, doi : <doi.org/10.1108/IJCS-08-2017-0013>; Federico Ast, « A Short Literature Review on Collective Intelligence », *Medium* (13 septembre 2015), en ligne : <medium.com/astec/a-brief-literature-review-on-collective-intelligence-2b7f7e4f4561>. Le concept d'intelligence collective ne se limite pas aux comportements de groupe qui peuvent être observés chez les humains, voir notamment : Iain D Couzin, « Collective cognition in animal groups » (2008) 13:1 *Trends in Cognitive Sciences* 36, doi : <doi.org/10.1016/j.tics.2008.10.002>; Ofer Feinerman et Amos Korman, « Individual versus collective cognition in social insects » (2017) 220 *J Exp Biol* 73, doi : <doi.org/10.1242/jeb.143891>.

644 Parmi une littérature abondante sur le sujet, soulignons quelques-unes des revues et découvertes les plus emblématiques : Thomas R Zentall, « Animal intelligence » dans Robert J Sternberg et Scott Barry Kaufman, dir, *Cambridge handbooks in psychology. The Cambridge handbook of intelligence*, 3e éd, Cambridge University Press, 2011, 309, doi : <doi.org/10.1017/CBO9780511977244.017>; Nathan J Emery, « Cognitive ornithology : The evolution of avian intelligence » (2006) 361:1465 *Philosophical Transactions of the Royal Society B Biological Sciences* 23, doi : <doi.org/10.1098/rstb.2005.1736>; Culum Brown, « Fish intelligence, sentience and ethics » (2014) 18:1 *Animal Cognition*, doi : <doi.org/10.1007/s10071-014-0761-0>; Ann Downer, *Smart and Spineless : Exploring Invertebrate Intelligence*, Twenty-First Century Books, 2015.

645 Simon Worrall, « There Is Such a Thing as Plant Intelligence », *National Geographic* (21 février 2016), en ligne : <www.nationalgeographic.com/science/article/160221-plant-science-botany-evolution-mabey-ngbooktalk>; Anthony Trewavas, « The foundations of plant intelligence » (2017) *Interface Focus*, doi : <doi.org/10.1098/rsfs.2016.0098>; Andrea Morris, « A Mind Without A Brain : The Science of Plant Intelligence Takes Root », *Forbes* (9 mai 2018), en ligne : <forbes.com/sites/andreamorris/2018/05/09/a-mind-without-a-brain-the-science-of-plant-intelligence-takes-root/?sh=376eeeb476dc>; André Geremia Parise, Monia Gagliano et Gustavo Maia Souza, « Extended cognition in plants : is it possible? » (2020) 15:2 *Plant Signaling & Behavior*, doi : <doi.org/10.1080/15592324.2019.1710661>.

646 Richard E Nisbett, « Intelligence : new findings and theoretical developments » (2012) 67:2 Am Psychol 130, doi : <doi.org/10.1037/a0026699>; Janet E Davidson et Iris A Kemp, « Contemporary Models of Intelligence » dans Robert J Sternberg et Scott Barry Kaufman, dir, *The Cambridge Handbook of Intelligence*, Cambridge, University Press, 2011, 58, doi : <doi.org/10.1017/CBO9780511977244.005>; Paul De Boeck et al, « An Alternative View on the Measurement of Intelligence and Its History » dans Robert J Sternberg, *The Cambridge Handbook of Intelligence*, Cambridge University Press, 2020, 47, doi : <doi.org/10.1017/9781108770422.005>.

D'une certaine manière, notre parcours vers l'intelligence forte ou augmentée (renvoi au Chapitre 1) repose principalement sur le dépassement du paradigme probabiliste[647]. Après tout, le fonctionnement du cerveau humain ne se fie pas qu'aux inférences probabilistes[648], et l'intelligence, dans ses étincelles protéiformes, ne se résout pas qu'à un travail (computationnel) même des plus acharnés.

Qu'il s'agisse de l'approche symbolique, connexionniste ou hybride (renvoi au Chapitre 1), le développement de l'intelligence artificielle « s'inspire de certaines capacités cognitives humaines pour les appliquer à des machines » (renvoi à la page citée au Chapitre 1). L'exercice d'assimilation est d'apparence trompeuse (renvoi au Chapitre 1). Plus que d'imiter l'humain ou sa cognition dans toutes ses splendeurs et misères, comment apprendre à la machine à performer plus efficacement et astucieusement que l'homme ? Cette question fondamentale, riche en expectatives, n'appelle pas encore de réponses définitives au sein de la communauté technoscientifique.

Les pistes de réflexion s'articulent autour d'une vision opératoire de l'intelligence artificielle en tant qu'un problème d'optimisation des modes de traitement de l'information. Jusque-là, ce traitement – intelligent – des données se veut être axé sur les tâches à accomplir (*task-oriented*) (renvoi au Chapitre 2) comme la perception, la détection de fraudes et la recommandation de contenus. Les défis de l'optimisation se situent sur les plans des résultats du traitement, de sa transparence ainsi que de son adaptabilité. Il importe tout d'abord d'optimiser les résultats du traitement pour qu'ils soient généralisables à un grand nombre de situations futures (1) tout en minimisant les coûts – en données et en calcul – de ce traitement (2). Il s'agit ensuite d'expliquer, de comprendre et de disséquer les facteurs pris en compte par les algorithmes d'apprentissage profond pour parvenir à leurs conclusions (3). L'adaptabilité enfin, faculté qui distingue le mieux – en biologie évolutive – un être vivant de la matière inanimée, souligne la nécessité d'un ajustement fonctionnel et dynamique entre l'agent intelligent et son environnement, d'où l'importance d'un apprentissage autonome et évolutif (continuellement mis à jour) à partir de l'expérience passée. Cet apprentissage implique non seulement la mise en place de règles opératoires et de structures internes dictant le potentiel d'action ou de réalisation des systèmes intelligents (4), mais aussi des conditions externes qui sont propices à l'apprentissage (5). Alors que les progrès actuels se situent principalement au niveau de la « cognition artificielle », le parcours vers l'intelligence artificielle forte ne pourrait passer sous silence les premiers pas vers une intelligence sociale augmentée (6), ainsi que l'éventuelle atteinte de la singularité technologique (7).

---

647 Voir aussi H James Wilson, Paul R Daugherty et Chase Davenport, « The Future of AI Will Be About Less Data, Not More », *Harvard Business Review* (14 janvier 2019), en ligne : <hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more>.

648 Adam N Sanborn et Nick Chater, « Bayesian Brains without Probabilities » (2016) 20:12 *Trends in Cognitive Sciences* 883, doi : <doi.org/10.1016/j.tics.2016.10.003>; Daniel Williams, « Predictive coding and thought » (2020) 197 *Synthese* 1749, doi : <doi.org/10.1007/s11229-018-1768-x>.

1. De l'optimisation des résultats prédictifs

Le développement des algorithmes prédictifs s'inscrit dans la continuité de la démarche scientifique : la valeur probante d'une théorie à prétentions scientifiques s'apprécie à l'aune de la robustesse de ses prédictions. Il importe non seulement de comprendre l'existant, mais aussi d'anticiper ce qui peut advenir. Explications (de ce qui s'est passé) et prédictions (de ce qui adviendra) se chevauchent en effet dans un système de lois universelles au temps invariant.

L'un des atouts de la théorie des probabilités appliquée à l'intelligence artificielle (IA) est précisément cette possibilité de se prévaloir d'un traitement optimisé et beaucoup plus efficace des données existantes en vue de quantifier la probabilité d'occurrence d'événements futurs, qu'il s'agisse du risque de faillite et d'insolvabilité, du comportement d'achat des consommateurs, ou encore des chances de succès d'un recours introduit devant un tribunal voire un(e) juge en particulier.

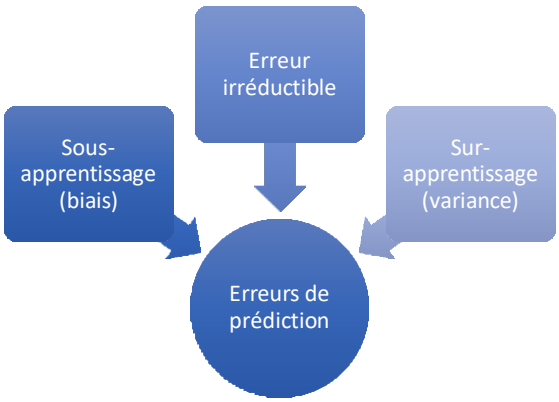
De même qu'un édifice n'est pas qu'un agrégat de matériaux, la qualité prédictive des algorithmes ne dépend pas que de l'accumulation des jeux de données ou de l'augmentation de la puissance computationnelle. Tant le sur-apprentissage (*overfitting*) que le sous-apprentissage (*underfitting*) sont susceptibles d'affecter la justesse prédictive des algorithmes intelligents.

- Le sous-apprentissage renvoie à l'incapacité de l'algorithme à saisir les corrélations existantes dans les données d'entraînement, incapacité qui se reflète dans des prédictions erronées relatives tant aux données d'entraînement qu'aux nouveaux jeux de données. En statistique, ce biais découle généralement d'hypothèses erronées quant à la pertinence des données de départ ou du modèle de corrélation. Ainsi, un modèle simple (p.ex. linéaire) est vulnérable au biais lorsqu'il est appliqué à l'analyse des phénomènes plus complexes.

- Un sur-apprentissage survient lorsque l'algorithme peine à « s'émanciper » des données d'entraînement de sorte qu'il n'obtient que des prédictions médiocres sur de nouveaux jeux de données. En statistique, cette variance s'explique par une trop grande sensibilité du modèle aux fluctuations aléatoires observées dans les données initiales qui nuisent à la généralisation.

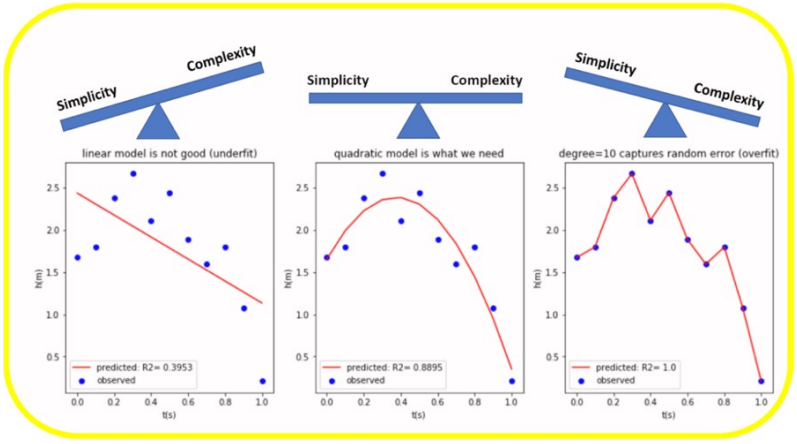
Les deux cas de figure posent la question du choix des données d'entraînement pertinentes. Il faut tout d'abord modéliser l'existant Ce dilemme biais-variance (*bias-variance tradeoff*) – auquel il convient également d'ajouter l'erreur ou le bruit irréductible sur lequel on ne peut pas agir (p.ex. données non mesurées ou imprécises) – est spécifique à l'apprentissage supervisé :

Figure 13
Erreurs de prédiction principales en apprentissage supervisé



La solution passe par un réajustement du modèle d'apprentissage en conjuguant simplicité et complexité :

Figure 14
Résolution du dilemme biais-variance : à la recherche du juste équilibre



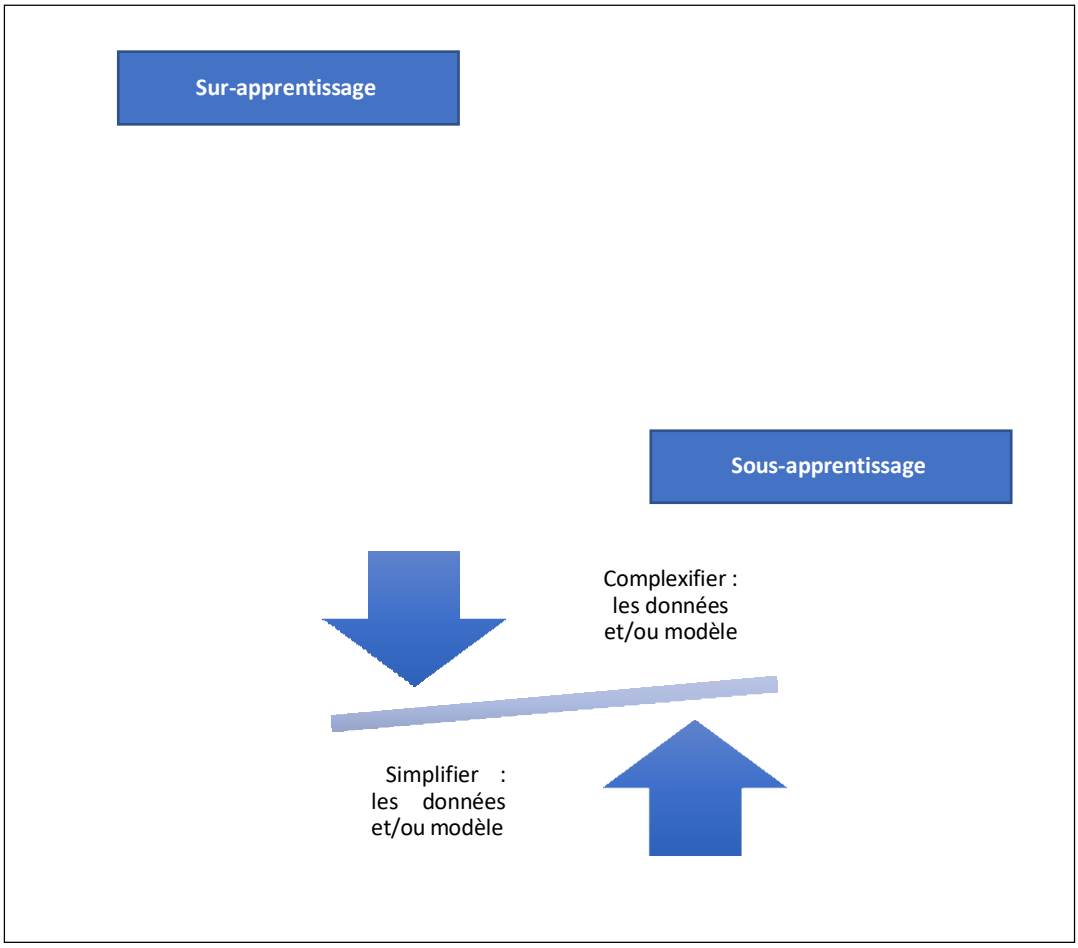
Source : Obi Tayo Ph.D., « Simplicity vs Complexity in Machine Learning – Finding the Right Balance », *Towards Data Science* (11 nov. 2019)[649]

En général, le sous-apprentissage peut être redressé en augmentant le jeu des données d'entraînement disponibles et/ou en complexifiant le modèle de corrélation. Le sur-apprentissage, de l'autre côté, appelle une simplification des jeux de données et/ou du modèle d'apprentissage par une restriction judicieuse de ses paramètres. À cet égard, le concept de parcimonie ou *sparsity* s'inspire du fonctionnement de notre cerveau dont les capacités multi-tâches s'accommodent d'un nombre limité de neurones (nœuds) et d'un temps limité

649 Benjain Obi Tayo, PhD, « Simplicity vs Complexity in Machine learning – Finding the Right Balance », *Towards Data Science* (11 novembre 2019), en ligne : <towardsdatascience.com/simplicity-vs-complexity-in-machine-learning-finding-the-right-balance-c9000d1726fb>.

d'apprentissage. La recherche d'un nombre optimal et parcimonieux de paramètres peut passer par plusieurs méthodes dites de régularisation du modèle d'apprentissage, lesquelles peuvent consister soit à imposer une contrainte sur les coefficients de pondération possibles du modèle (régulation Ridge ou Lasso), soit, dans le cas des réseaux de neurones, à désactiver une partie des neurones et leurs connections pendant l'entraînement pour éviter une co-adaptation excessive (*dropout*[650]) :

Figure 15
Ajuster les modèles d'apprentissage supervisé pour optimiser la généralisation



En tout état de cause, il importe moins pour l'algorithme d'apprivoiser un jeu de données en particulier que d'apprendre, par induction, à généraliser ses acquis dans un grand nombre de situations (à venir).

650 Nitish Srivastava et al, « Dropout : A Simple Way to Prevent Neural Networks from Overfitting » (2014) 15:56 Journal of Machine Learning Research 1929 : « The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. »

Pour en savoir plus :

Al-Masri, A., « What Are Overfitting and Uderfitting in Machine Learning ? », *Towards data science* (22 juin 2019), en ligne : <towardsdatascience.com/what-are-overfitting-and-underfitting-in-machine-learning-a96b30864690>

Bhande, A., « What is underfitting and overfitting in machine learning and how to deal with it. », *Medium* (11 mars 2018), en ligne : <medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76>

Tripathi, M., « Underfitting and Overfitting in Machine Learning », DataScience Foundation, 13 juin 2020, en ligne : <datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning>

2. De l'apprentissage contextuel : au-delà du compromis optimisation-force brute

Malgré des enviables[651], le modèle probabiliste appliqué aux algorithmes prédictifs est lui-même doté de domaines d'application, depuis le filtre anti-pourriel et les systèmes de recommandation jusqu'à la perception artificielle, intégrée notamment dans les véhicules autonomes.

En effet, la perception artificielle se calque sur la perception humaine en cherchant à modéliser par la machine une représentation du monde telle que perçue par nos cinq sens. Très tôt, la vision artificielle (par ordinateur) aspire à simuler la vision humaine dans l'interprétation, la reconstruction et la compréhension des images appliquées à la reconnaissance aérienne, à la robotique industrielle et à l'imagerie médicale. Propulsée à l'avant-plan des recherches en intelligence artificielle par le modèle du néocognitron[652], la vision artificielle est suivie de près par l'avènement des technologies en reconnaissance vocale numérique (années 1990) ainsi que des nez électroniques (années 2010). La perception artificielle vise ultimement à développer une conscience situationnelle (*situational awareness*) de la machine, soit une connaissance / conscience synthétique de l'information dynamique présente dans son environnement pour anticiper les actions à prendre et en optimiser l'efficacité.

Or, développer une conscience situationnelle artificielle s'avère ardu : l'intelligence situationnelle n'émerge au prix du traitement de grandes quantités de données et d'essais-erreurs grugeant la puissance computationnelle (malédiction de la dimension). L'optimisation par la « force brute » a ses limites lorsqu'il s'agit d'apprendre à l'algorithme à reconnaître, à partir d'un nombre limité de mises en situation, les données indicielles pertinentes à une représentation globalement juste qui soit transposable dans une multitude de contextes (p.ex. superposition d'objets, objets en mouvement), de conditions d'éclairage, d'exposition (p.ex.

---

651 Voir par exemple le mégamodèle de langage GPT-3 entraîné avec plus de 175 milliards de paramètres et presque l'ensemble des données indexables du net : Tom B Brown et al, *supra* note 355.

652 Kunihiko Fukushima, « Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position » (1980) 36 Biological Cybernetics 193, doi : <doi.org/10.1007/BF00344251>.

images partiellement visibles) et d'angles de prise de vue. Par conséquent, les systèmes autonomes sont mal adaptés pour réagir aux conditions opérationnelles imprévues[653].

Théorie des graphes (*graph theory*) et ses variantes – L'une des avenues de solution est l'approche graphique. Il s'agit d'entraîner l'algorithme non seulement à partir des données, mais aussi du contexte qu'il est possible de modéliser comme des représentations types sur les données existantes (nœuds) et les relations de voisinage entre les données (arêtes). Le compromis entre l'optimisation et la « force brute » se réalise par un apprentissage des propriétés structurelles caractérisant l'ensemble des données d'entraînement et leur réseau de relations, qu'il s'agisse :

- de la classification d'objets récurrents dans la reconnaissance d'images en général[654],
- de la détection (des schémas) de fraudes[655];
- de l'identification d'opportunités d'investissement[656];
- des interactions types entre les véhicules en circulation[657]; et
- du dépistage de drogues ou développement de médicaments aux structures et propriétaires moléculaires semblables[658].

Dans tous les cas, les propriétés d'ordre structurel imposent des contraintes sur l'algorithme d'apprentissage que certains qualifieraient de « biais inductifs relationnels » / « *relational inductive biases* »[659]. Il s'agit d'apprendre l'algorithme à reconnaître ce qui caractérise un réseau interrelié de propriétés plutôt que des caractéristiques (*features*) isolées, avec une efficacité marquée à l'égard de la reconnaissance de structures / situations invariantes qui apparaissent (très) différentes mais qui sont néanmoins structurellement identiques :

---

653 Shashi Phoha, « Machine perception and learning grand challenge : situational intelligence using cross-sensory fusion » (2014) *Front. Robot. AI*, doi : <doi.org/10.3389/frobt.2014.00007>.

654 Zhao-Min Chen et al., « Multi-Label Image Recognition with Graph Convolutional Networks » (2019) *Computer Vision and Pattern Recognition*, en ligne : <arxiv.org/abs/1904.03582>.

655 Jianguo Jiang et al, « Anomaly Detection with Graph Convolutional Networks for Insider Threat and Fraud Detection » dans *MILCOM 2019 – 2019 IEEE Military Communications Conference (MILCOM)*, Norfolk (VA), 2019, 109, doi : <doi.org/10.1109/MILCOM47813.2019.9020760>.

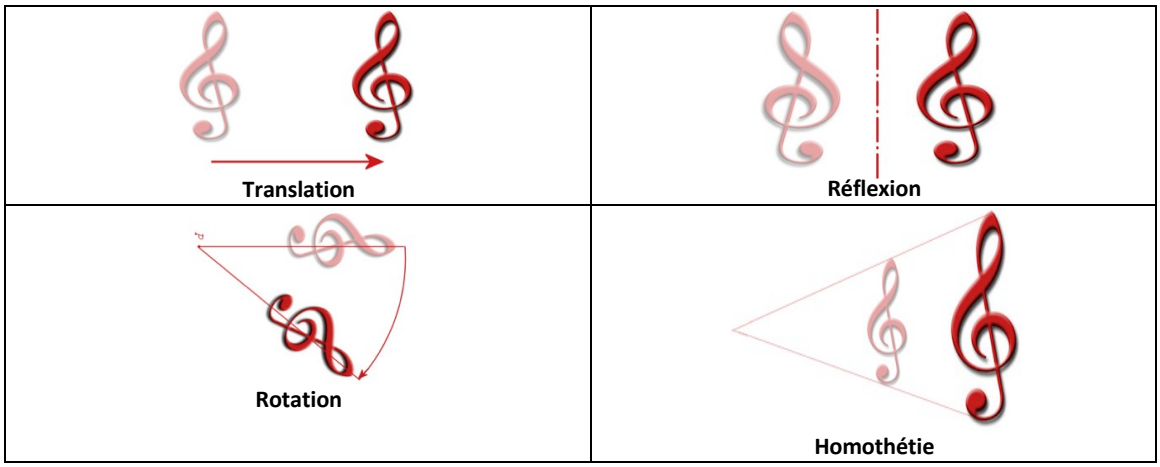
656 Thársis Souza, « Connecting the Dots : Using AI & Knowledge Graphs to Identify Investment Opportunities », *Towards Data Sicence* (28 mars 2019), en ligne : <towardsdatascience.com/knowledge-graphs-in-investing-733ab34abe>.

657 Donsuk Lee et al, « Joint Interaction and Trajectory Prediction for Autonomous Driving using Graph Neural Networks » (2019) *Machine Learning for Autonomous Driving NeurIP*S, en ligne : <arxiv.org/abs/1912.07882>.

658 Junying Li, Deng Cai et Xiaofei He, « Learning Graph-Level Representation for Drug Discovery » (2017), en ligne : <arxiv.org/abs/1709.03741>; Mengying Sun et al, « Graph convolutional networks for computational drug development and discovery » (2020) 21:3 *Briefings in Bioinformatics* 919, doi : <doi.org/10.1093/bib/bbz042>.

659 Peter W Battaglia et al, « Relational inductive biases, deep learning, and graph networks » (2018), en ligne : <arxiv.org/abs/1806.01261>.

Figure 16
Structures invariantes sous diverses transformations géométriques



Source : <www.intmath.com/blog/mathematics/music-and-transformation-geometry-5074>

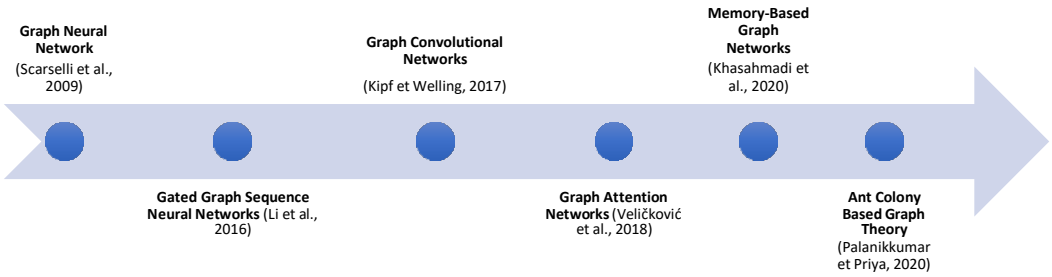
Selon plusieurs[660], le mécanisme d'apprentissage de notre néocortex reposerait davantage sur la reconnaissance de structures, de relations d'interdépendances et d'associations pertinentes plutôt qu'une analyse simplement conjointe mais isolée d'entrées sensorielles individuelles. Originellement proposé par Scarselli et ses collaborateurs[661], le modèle des réseaux de neurones graphiques (GNN : *graph neural network*) connaît dorénavant plusieurs variantes appliquant la théorie des graphes à l'apprentissage automatique[662].

660 Jeff Hawkins et Sandra Blakeslee, On Intelligence: How a New Understanding of the Brain will Leaad to the Creation of Truly Intelligent Machines, Times Books, 2004; Andy Clark, Surfing Uncertainty: Prediction, Action, and the Embodied Mind, Oxford Scholarship Online, 2016, doi : <doi.org/10.1093/acprof:oso/9780190217013. 001.0001>.

661 Franco Scarselli et al, « The graph neural network model » (2009) 20:1 IEEE Transactions on Neural Networks 61, doi : <doi.org/10.1109/TNN.2008.2005605>.

662 Yujia Li et al, « Gated Graph Sequence Neural Networks » (2016) ICLR, en ligne : <arxiv.org/abs/1511.05493>. Voir aussi : Thomas N Kipf et Max Welling, « Semi-Supervised Classification with Graph Convolutional Networks » (2017) ICLR, en ligne : <arxiv.org/abs/1609.02907>; Peter Veličković et al., « Graph Attention Networks » (2018) ICLR, en ligne : <arxiv.org/abs/1710.10903>; Amir Hosein Khasahmadi et al, « Memory-Based Graph Networks » (2020) ICLR, en ligne : <arxiv.org/abs/2002.09518>; D Palanikkumar et S Priya, « Ant colony based graph theory (ACGT) and resource virtual network mapping (RVNM) algorithm for home healthcare system in cloud environment » (2020) 79 Multimedia Tools and Applications 3743, doi : <doi.org/10.1007/s11042-018-6908-2>.

Figure *17*
La théorie des graphes et ses variantes appliquées à l'apprentissage automatique



Pour en savoir plus sur la théorie des graphes appliquée à l'intelligence artificielle, voir :

Anand, R., « An Illustrated Guide to Graph Neural Networks », *Medium* (30 mars 2020), en ligne : <medium.com/dair-ai/an-illustrated-guide-to-graph-neural-networks-d5564a551783>

Liao, W. et al., « A Review of Graph Neural Networks and Their Applications in Power Systems » (2021), en ligne : <arxiv.org/abs/2101.10025>

Sato, R., « A Survey on the Expressive Power of Graph Neural Networks » (2020), en ligne : <arxiv.org/abs/2003.04078>

Webber, J., « Graphs for Artificial Intelligence and Machine Learning », *Neo4j Blog* (18 février 2021), en ligne : <neo4j.com/blog/graphs-for-artificial-intelligence-and-machine-learning/>

Zhou, J. et al., « Graph neural networks : A review of methods and applications » (2020) AI Open 1 57, doi : <doi.org/10.1016/j.aiopen.2021.01.001>

Apprentissage non supervisé de la perception – Dans un autre ordre d'idées, plusieurs chercheurs ont expérimenté diverses méthodes d'apprentissage non supervisé appliquées à la perception artificielle[663]. L'apprentissage non supervisé présente tout d'abord l'avantage de ne plus nécessiter l'étiquetage manuel des données d'entraînement. En plus d'être un processus fastidieux et coûteux, cet étiquetage par l'humain[664] s'avère « incomplet » en ce qu'il ne permet pas à la machine de saisir des relations / associations latentes qui, quoique pertinentes, n'auraient pas été identifiées (étiquetées) d'avance. Bien des schèmes, structures et régularités latentes qui permettraient, par exemple, de distinguer un chat rasé d'une chauve-souris, s'avèrent difficilement étiquetables aux fins d'entraînement des algorithmes. Les possibilités

663 Filip Piekniewski et al, *Unsupervised Learning from Continuous Video in a Scalable Predictive Recurrent Network*, 2016, en ligne : <arxiv.org/pdf/1607.06854.pdf>. Voir aussi : Tejas Kulkarni et al, « Unsupervised Learning of Object Keypoints for Perception and Control » (2019) NeurIPS, en ligne : <arxiv.org/abs/1906.11883>; Katherine R Storrs et Roland W Fleming, « Unsupervised Learning Predicts Human Perception and Misperception of Gloss » (2020), doi : <doi.org/10.1101/2020.04.07.026120>.

664 Cade Metz, « A.I. Is Learning From Humans. Many Humans », *New York Times* (16 août 2019), en ligne : <www.nytimes.com/2019/08/16/technology/ai-humans.html>.

d'apprentissage sont aussi de ce fait élargies puisque la disponibilité des données non étiquetées surpasse de loin celle des jeux de données étiquetées[665].

À cet égard, il est possible d'affirmer que l'apprentissage non supervisé, par son exploration ouverte des divers facteurs présents dans les jeux de données, s'accorde davantage avec la manière synthétique, intuitive dont l'humain apprend dans son environnement. À titre d'exemples de méthodes d'apprentissage non supervisé appliquées à la perception artificielle, Storrs et Fleming[666] ont rendu compte de leurs travaux relatifs à l'apprentissage non supervisé appliqué à des réseaux de neurones génératifs pour la reconnaissance des rendus de surfaces lustrées. La performance de l'algorithme s'avère comparable et permet même de prédire les perceptions humaines jusqu'aux illusions d'optique engendrées par l'interaction entre les matériaux, les formes et l'éclairage.

Pour en savoir plus sur l'apprentissage non supervisé appliqué à la perception artificielle, voir :

Chen, W. et al., « Unsupervised Image Classification for Deep Representation Learning » (2020) ECCV, en ligne : <arxiv.org/abs/2006.11480>

Kulkarni, T. et al., « Unsupervised Learning of Object Keypoints for Perception and Control » (2019), en ligne : <arxiv.org/abs/1906.11883>

Olaode, A., G. Naghdy et C. Todd, « Unsupervised Classification of Images : A Review » (2014) 8:5 IJIP 325, en ligne : <www.cscjournals.org/manuscript/Journals/IJIP/Volume8/Issue5/IJIP-918.pdf>

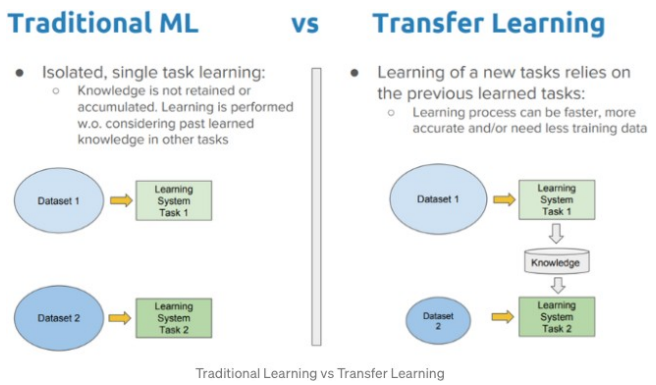
Van Gansbeke, W. et al., « SCAN : Learning to Classify Images without Labels » (2020) ECCV, en ligne : <arxiv.org/abs/2005.12320>

Apprentissage par transfert – L'attrait d'obtenir une performance meilleure ou équivalente avec moins de données d'entraînement a également motivé le développement de l'apprentissage par transfert. Au lieu de cantonner les acquis d'apprentissage en silos surspécialisés, il s'agit de développer chez la machine des compétences primaires qui puissent être facilement transposées d'un contexte à l'autre, soit en tirant parti des connaissances acquises dans des modèles génériques pour bâtir de nouveaux modèles plus spécialisés, soit en les appliquant dans de nouveaux contextes d'apprentissage :

665 Rob Toews, « The Next Generation of Artificial Intelligence », *Forbes* (12 octobre 2020), en ligne : <www.forbes.com/sites/robtoews/2020/10/12/the-next-generation-of-artificial-intelligence/?sh=39e7f5859eb1>.

666 Storrs et Fleming (2020), *supra* note 669.

Figure *18*
Apprentissage automatique traditionnel vs Apprentissage par transfert



Source : Sarkar (2018)[667]

En reconnaissance d'images, le cumul des couches dans les réseaux de neurones profonds permet d'optimiser l'utilisation des modèles pré-entraînés sur des tâches de classification à grande échelle afin de peaufiner (*fine-tuning*) l'apprentissage spécialisé des couches supérieures à l'égard des tâches spécifiques[668] (renvoi au Chapitre 2).

En effet, une intelligence artificielle forte (renvoi au Chapitre 1), au sens véritable du terme, suppose un système qui soit en mesure de fonctionner de manière autonome non seulement dans un domaine d'application prédéfini, mais plutôt dans divers domaines de fonctionnement, d'où l'importance d'acquérir et de développer des habiletés qui soient indépendantes des domaines et en même temps facilement transférables d'un domaine à l'autre[669].

---

667 Dipanjan (DJ) Sarkar, « A Comprhensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning », *Towards Data Science* (14 novembre 2018), en ligne : <towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a>.

668 Voir entre autres Mark Chen et al, « Generative Pretraining from Pixels » (2020), en ligne : <cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf>.

669 JF Pagel et Philip Kirshtein, *Machine Dreaming and Consciousness*, Academic Press, Elsevier, 2017 à la p 60 : « Strong AI is the concept of a system with the ability not only to operate autonomously within a predefined area of application, but to be completely autonomous within any general field of application. Such a system must have domain-independent skills necessary for acquiring a wide range of domain-specific knowledge. »

Pour en savoir plus sur l'apprentissage par transfert, voir :

Alto, V., « Transfer Learning for Computer Vision. An Implementation with Python », Medium (31 décembre 2020), en ligne : <medium.com/analytics-vidhya/transfer-learning-for-computer-vision-a1a8cd42d22d>

Marcelino, P., « Transfer learning from pre-trained models », Towards data science (23 octobre 2018), en ligne : <towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751#:~:text=Transfer%20learning%20is%20a%20popular,when%20solving%20a%20different%20problem.>

Ruder, S., « Transfer Learning – Machine Learning's Next Frontier », 21 mars 2017, en ligne : <ruder.io/transfer-learning/>

Sarkar, D., « A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning », Towards Data Science (14 novembre 2018), en ligne : <towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a>

Les transformeurs – Un autre modèle d'apprentissage prometteur est l'architecture des transformeurs, initialement proposée par Vaswani et al.[670]. Excellant surtout dans les tâches de traitement du langage naturel, cette architecture repose sur les mécanismes d'attention mettant l'emphase sur la pertinence du contexte pour permettre un traitement parallèle (plutôt que séquentiel) et partant, plus rapide, des jeux de données d'entrée. Dans le domaine du traitement du langage naturel, cette architecture facilite la prise en compte du contexte d'usage dans l'interprétation des différents mots ou expressions. Son application la plus connue est sans doute le GPT-3, le modèle de langage le plus puissant qui a été développé à ce jour et qui a été entraîné avec 175 milliards de paramètres[671]. Plus récemment, l'architecture des transformeurs a également être mobilisée pour la reconnaissance d'images, avec des résultats encourageants[672].

670 Ashish Vaswani et al., *supra* note 530.

671 Brown et al (2020), *supra* note 355.

672 Alexey Dosovitskiy et al, *supra* note 595.

Pour en savoir plus sur le modèle des transformeurs, voir :

Adaloglou, N., « How Transformers work in deep learning and NLP : an intuitive introduction », *AI Summer* (24 décembre 2020), en ligne : <theaisummer.com/transformer/>

Adam, I., « The Rise of the Transformers : Explaining the Tech Underlying GPT-3 », 21 septembre 2020, en ligne : <www.linkedin.com/pulse/rise-transformers-imtiaz-adam/>

Agarwal, R., « What Are Transformer Models in Machine Learning ? », *Lionbridge AI* (9 septembre 2020), en ligne : <lionbridge.ai/articles/what-are-transformer-models-in-machine-learning/>

Alammar, J., « The Illustrated Transformer », blogue personnel de Jay Alammar, en ligne : <jalammar.github.io/illustrated-transformer/>

Lingyi, « GPT-3, transformers and the wild world of NLP », *Towards data science* (16 septembre 2020), en ligne : <towardsdatascience.com/gpt-3-transformers-and-the-wild-world-of-nlp-9993d8bb1314>

3. De l'IA explicable ou XAI (« *eXplainable Artificial Intelligence* ») : au-delà du compromis interprétabilité-performance

Au-delà du chiffre des prédictions et indépendamment de leur étonnante justesse, il importe en parallèle d'optimiser la transparence des algorithmes intelligents sur les paramètres pris en compte ainsi que le chemin que les algorithmes ont pris pour arriver aux résultats, au risque d'entretenir cette « mystique de la technologie » qui devient une autre Pythie.

À plusieurs égards, l'exactitude des prédictions n'est en effet pas la seule métrique devant être prise en compte pour évaluer la fiabilité du modèle. Les « bons » résultats peuvent avoir été obtenus à l'aide de « mauvaises » données (p.ex. non autorisées) qui auraient été incluses par inadvertance dans les données d'entraînement. Le modèle pourrait encore exploiter des corrélations aléatoires qui s'avèrent absurdes en réalité. La performance seule n'est donc pas suffisante tant pour étayer la fiabilité du modèle qu'inspirer confiance.

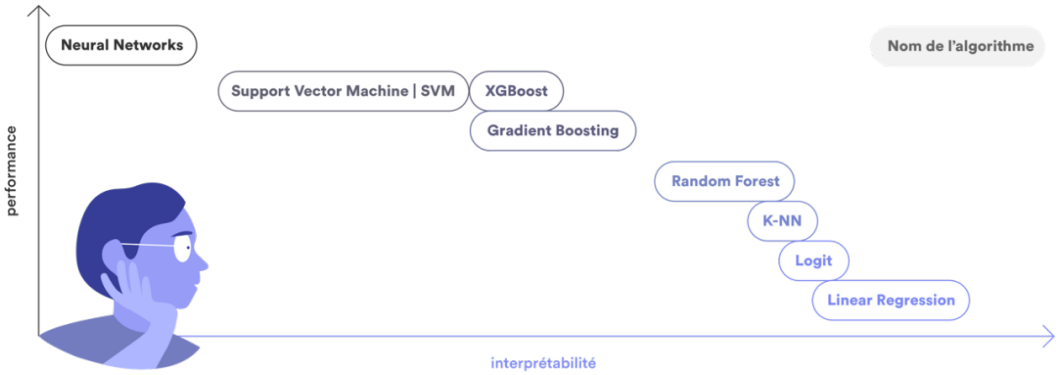
Le défi d'intelligibilité se corse en ce que l'interprétabilité des modèles (plus simples) nuit en général à la justesse de leurs prédictions. Moins un modèle comporte de paramètres et de coefficients de pondération complexes, plus il sera facile de l'expliquer et de l'interpréter. De l'autre côté, les progrès de l'apprentissage profond reposent justement sur la complexité du modèle pour améliorer sa performance.

Le spectre de l'interprétabilité – Hall, Ambati et Phan[673] proposent un spectre de l'interprétabilité des fonctions de réponse d'après certaines de leurs propriétés. À une extrémité du spectre de l'interprétabilité, les modèles linéaires et monotones sont les plus facilement interprétables : pour chaque (combinaison ou fonction de) variable indépendante, la fonction de réponse varie à

673 Patrick Hall, SriSatish Ambati et Wen Phan, « Ideas on interpreting machine learning. Mix-and-match approaches for visualizing data and interpreting machine learning models and results », *O'Reilly*, 15 mars 2017, en ligne : <www.oreilly.com/radar/ideas-on-interpreting-machine-learning/>.

un taux fixe et dans une seule direction. L'interprétabilité des modèles non linéaires et monotones ne pose pas non plus de difficultés particulières : les fonctions de réponse, quoique variant à un rythme différent suivant chaque (combinaison ou fonction de) variable indépendante, restent unidirectionnelles. La plupart des algorithmes intelligents, dont les réseaux de neurones, créent toutefois des fonctions de réponse non linéaires et non monotones. Les interpréter se complique du fait que pour chaque (combinaison ou fonction de) variable indépendante, les variables dépendantes peuvent fluctuer dans différentes directions et à des degrés variables :

Figure *19*
Compromis interprétabilité-performance



Source :Nathan Lauga, « IA et éthique : Comment comprendre son modèle ? », *Medium* (25 février 2019)

Le domaine de l'IA explicable ou XAI (« *eXplainable Artificial Intelligence* ») regroupe un ensemble de techniques et de méthodes qui permet d'interpréter les solutions intelligentes d'une manière qui soit compréhensible et qui paraisse sensé à l'humain. En attendant la maturation des modèles qui seraient intrinsèquement transparents sans préjuger de leur performance[674], la construction de modèles dits agnostiques (*agnostic models*) est privilégiée pour expliquer par proxy le fonctionnement interne des algorithmes dits de boîte noire. Généralement appréciées pour leur flexibilité et leur adaptabilité[675], les techniques agnostiques sont développées suivant trois grandes approches méthodologiques que sont les explications causales approximatives, les explications contrefactuelles et les modèles substituts.

Explications causales approximatives – On pourrait tout d'abord mesurer l'effet marginal d'une ou de plusieurs caractéristiques (*features*) sur les prédictions obtenues par l'algorithme et en estimer la relation – linéaire, quadratique ou plus complexe – entre une entrée et la sortie[676]

674 Voir entre autres : Jimmy Lin et al, « Generalized and Scalable Optimal Sparse Decision Trees » (2020) *ICML*, en ligne : <arxiv.org/abs/2006.08690>.

675 Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, « Model-Agnostic Interpretability of Machine Learning » (2016) *ICML Workshop on Human Interpretability in Machine Learning*, en ligne : <arxiv.org/abs/1606.05386>.

676 Jerome H Friedman, « Greedy Function Approximation : A Gradient Boosting Machine » (2001) 29 *The Annals of Statistics* 1189, doi : <doi.org/10.1214/aos/1013203451>.

(*dependence plots, Shapley values*, *SHAP*[677], *knowledge graphs*[678], *feature visualization*, *network /GAN dissection*[679]). Outre que l'activation de différentes unités de neurones peut encore être étudiée et interprétée localement, l'effet cumulé ou moyen de plusieurs caractéristiques sur les prédictions peut aussi être estimé par le biais de plusieurs techniques statistiques (*pair-wise interactions*, *accumulated local effects plot*[680]).

Explications contrefactuelles – Il est également possible de contre-vérifier les résultats obtenus à l'aide d'une approche dite contrefactuelle. Celle-ci consiste à raisonner à rebours de la chaîne causale, en se demandant si la conclusion aurait été la même ou différente n'eût été une hypothèse de départ donnée. Appliquée à l'apprentissage automatique, la méthode contrefactuelle cherche à déterminer la plus petite variation dans le modèle de corrélations qui affecterait les résultats, ce qui permettra de cerner les facteurs pris en compte par l'algorithme dans sa prise de décision[681].

Des modèles substituts (*Surrogate models*) – Des modèles substituts peuvent encore être développés pour approximer, à l'aide de modèles plus interprétables (p.ex. modèle linéaire, arbre de décision…), les prédictions et les étapes de raisonnement d'algorithmes de boîte noire. Un modèle substitut peut être entraîné à partir des données ou un sous-ensemble de données utilisées par l'algorithme à interpréter ainsi que de la fonction de prédiction de l'algorithme de boîte noire. Par ailleurs, plus d'un modèle substitut, utilisant tantôt la régression linéaire, tantôt l'arbre de décision, peut être développé à l'égard d'un même algorithme de boîte noire. Cela étant, il subsistera toujours une variance entre les modèles substituts et le modèle devant être interprété. Certains modèles substituts peuvent n'approximer adéquatement les résultats d'un algorithme de boîte noire qu'à l'égard d'un (sous-)ensemble de données seulement. Il s'agit alors d'apporter des « explications locales interprétables par modèle agnostique », plus connues sous l'acronyme anglais LIME (« *local interpretable model-agnostic explanations* »)[682].

---

677 Scott Lundberg et Su-In Lee, « A Unified Approach to Interpreting Model Predictions » (2017) *NIPS*, en ligne : <arxiv.org/abs/1705.07874>.

678 Freddy Lecue, « On the Role of Knowledge Graphs in Explainable AI » (2019) 11:1 *Semantic Web* 1, doi : <doi.org/10.3233/SW-190374>.

679 David Bau et al, « Network Dissection : Quantifying Interpretability of Deep Visual Representations » (2017) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, en ligne : <arxiv.org/abs/1704.05796>; David Bau et al, « GAN Dissection : Visualizing and Understanding Generative Adversarial Networks » (2018), en ligne : <arxiv.org/abs/1811.10597>.

680 Daniel W Apley et Jingyu Zhu, « Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models » (2019), en ligne : <arxiv.org/abs/1612.08468>.

681 Sandra Wachter, Brent Mittelstadt et Chris Russell, « Counterfactual Explanations without Opening the Black Box : Automated Decisions and the GDPR » (2018) *Harvard Journal of Law & Technology*, en ligne : <arxiv.org/abs/1711.00399>.

682 Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, *supra* note 632.

Pour en savoir plus :

Aguiar, R., « An overview of model explainability in modern machine learning », Towards data science (5 décembre 2019), en ligne : <towardsdatascience.com/an-overview-of-model-explainability-in-modern-machine-learning-fc0f22c8c29a>

Arrieta, A.B. et al., « Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI » (2020) 58 Information Fusion 82, doi : <doi.org/10.1016/j.inffus.2019.12.012>

Chou, Y.-L. et al., « Counterfactuals and Causability in Explainable Artificial Intelligence : Theory, Algorithms, and Applications » (2021), en ligne : <arXiv:2103.04244>

Molnar, C., A Guide for Making Black Box Models Explainable, 2019, en ligne : <christophm.github.io/interpretable-ml-book/>

Sarkar, D., « Hands-on Machine Learning Model Interpretation. A comprehensive guide to interpreting machine learning models », Towards data science (13 décembre 2018), en ligne : <towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608>

Verma, S., J. Dickerson et K. Hines, « Counterfactual Explanations for Machine Learning : A Review » (2020), en ligne : <arXiv:2010.10596>

4. Des modèles d'apprentissage plus étroitement couplés au fonctionnement de notre cerveau

Pour plusieurs, l'avenir de l'intelligence artificielle repose sur l'optimisation de l'apprentissage profond par un couplage toujours plus étroit et précis des mécanismes neurophysiologiques complexes sous-tendant notre apprentissage cortical[683], en tant que le système biologique le plus complexe étayant l'acquisition et la consolidation d'aptitudes cognitives des plus polyvalentes[684]. Le cerveau humain fonctionne de manière à la fois analogique et digitale (*cf.* potentiel d'action) tout en opérant massivement en parallèle sans processeur central en intégrant simultanément les signaux provenant des milliers de neurones. Tandis que, dans une mémoire d'ordinateur, toutes les combinaisons de 1 et de 0 sont nécessaires de sorte que la perte ou la modification d'un signal donné peut résulter en un résultat tout à fait différent, notre cerveau fonctionne à l'aide des distributions neuronales que Hawkins[685] explique comme étant clairsemées (*sparse*

683 Jeff Hawkins, « What Intelligent Machines Need to Learn From the Neocortex », *IEEE Spectrum* (2 juin 2017), en ligne : <spectrum.ieee.org/computing/software/what-intelligent-machines-need-to-learn-from-the-neocortex> [Hawkins (2017)]; Aras R Dargazany, *Deep Learning Research Landscape & Roadmap in a Nutshell : Past, Present and Future – Towards Deep Cortical Learning*, 7 août 2019, en ligne : <arxiv.org/pdf/1908.02130.pdf>.

684 On doit garder à l'esprit que cette imitation de la nature ne se limite pas qu'au système nerveux humain. Plusieurs algorithmes d'optimisation peuvent s'inspirer du comportement d'autres espèces – y compris certains agencements associatifs supraorganiques observés chez des animaux sociaux, tels que les algorithmes de colonies de fourmis (« *ant colony optimization* » ou ACO), les méthodes d'optimisation par essaims particulaires (« *particle swarm optimization* » ou PSO) et l'algorithme d'optimisation des troupeaux d'éléphants (« *elephant herding optimization* » ou EHO).

685 Jeff Hawkins (2017), *supra* note 689.

distributed representations ou SDRs) : seul un relativement petit nombre de neurones étant actifs à un moment donné, nos représentations sont intrinsèquement robustes face à la destruction ou l'inactivité de neurones isolés en plus de permettre des superpositions flexibles, une « riche compositionnalité »[686] entre des concepts similaires, voisins mais distincts.

Ces différences fondamentales subsistant entre un cerveau artificiel et notre cerveau ne doivent pas nous faire perdre de vue que l'histoire de l'apprentissage automatique consiste en une approximation incrémentielle, mais toujours plus étroite et fidèle, du fonctionnement de notre cerveau. Qu'il s'agisse du neurone formel[687], du perceptron[688] multicouche[689], ou du néocognitron[690], cet agencement toujours plus complexe de fonctions mathématiques en réseaux de neurones artificiels vise ultimement à reproduire un cadre structurel[691] éprouvé à l'émergence de processus intelligents. À ce jour, le couplage se base principalement sur le fonctionnement de notre néocortex, auquel s'ajoute une prise de conscience accrue de la nécessité de mobiliser d'autres structures, notamment sous-corticales, dans notre parcours vers une intelligence artificielle générale,

Néocortex – Le néocortex, cette couche externe qui enveloppe nos hémisphères cérébraux, représente près du trois quart de notre volume cérébral total et est impliqué principalement dans les fonctions cognitives dites supérieures telles que les perceptions sensorielles, le language et le raisonnement. Les habiletés cognitives dites supérieures permettent aux sujets de se fixer des objectifs à atteindre, d'être attentif à son environnement et d'adapter son comportement en conséquence.

686 Williams (2020), *supra* note 648.

687 JY Lettvin et al, « What the Frog's Eye Tells the Frog's Brain » (1959) 47:11 *Proceedings of the IRE* 1940, doi : <doi.org/10.1109/JRPROC.1959.287207>.

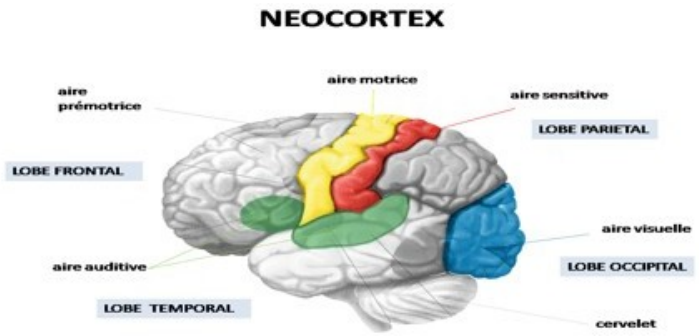
688 F Rosenblatt, *supra* note 110.

689 David E Rumelhart, Geoffrey E Hinton et Ronald J Williams, *supra* note 195. Voir aussi : Paul J Werbos, *The Roots of Backpropagation : From Ordered Derivatives to Neural Networks and Political Forecasting*, New York, John Wiley & Sons, 1994.

690 Kunihiko Fukushima, *supra* note 652.

691 *Cf.* Thomas Pircher et al, « The Structure Dilemma in Biological and Artificial Neural Networks » (2021) 11 *Scientific Reports*, doi : <doi.org/10.1038/s41598-021-84813-6>.

Figure 20
Notre néocortex



Source : <www.msrblog.com/science/biology/neocortex.html>

Outre de simples analogies d'ordre structurel et fonctionnel entre les réseaux de neurones biologiques et artificiels, le processus d'apprentissage profond permettrait de développer des réseaux de représentations qui correspondent davantage aux enregistrements d'activités neuronales de notre néocortex que des modèles existants en neurosciences[692].

En apprentissage profond, les réseaux de neurones apprennent typiquement de leurs erreurs (écarts) de prédiction par la technique de descente du gradient via l'algorithme de rétropropagation (*backpropagation*), laquelle permet d'ajuster, de façon itérative et en rétrospective au regard des résultats attendus, le paramétrage (poids et biais) des neurones individuels[693]; renvoi au Chapitre 2). Malgré son succès avéré, cette rétropropagation ne correspond toutefois pas exactement au fonctionnement de notre système nerveux, qui l'est par impulsions discrètes (seuil d'activation du tout ou rien) plutôt que des fonctions d'activation continues requises pour dériver les taux d'erreurs minimaux. Mais surtout, cette rétro-validation des résultats de la couche de sortie vers la couche d'entrée, nécessite une grande quantité de données étiquetées en plus de faire dépendre l'ajustement d'un seul paramètre de l'ensemble des paramétrages et calculs en aval de cette propagation[694]. De l'autre côté, il n'est pas démontré que notre cerveau fasse aussi appel dans son apprentissage à ce lourd et coûteux mécanisme[695].

---

692 Charles F Cadieu et al, « Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition » (2014) 10 PLoS Comput Biol, e1003963, doi : <doi.org/10.1371/journal.pcbi.1003963>.

693 David E Rumelhart, Geoffrey E Hinton et Ronald J Williams, *supra* note 195.

694 Voir notamment Richa Bhatia, « Back-Propagation : Is It The Achilles Heel Of Today's AI », *Analytics India Magazine* (7 novembre 2017), en ligne : <analyticsindiamag.com/back-propagation-is-it-the-achilles-heel-of-todays-ai>.

695 Yoshua Bengio et al, « Towards Biologically Plausible Deep Learning » (2015), en ligne : <arxiv.org/pdf/1502.04156>. Voir aussi : Alessandro Betti, Marco Gori et Giuseppe Marra, « Backpropagation and Biological Plausibility » (2018), en ligne : <arxiv.org/abs/1808.06934>; Timothy P Lillicrap et al, « Backpropagation and the brain » (2020) 21 Nature Reviews Neuroscience 335, doi : <doi.org/10.1038/s41583-020-0277-3>.

En s'inspirant des différents modèles d'apprentissage neuronaux[696], plusieurs alternatives ne faisant pas appel à la descente du gradient ont été proposées, dont les machines de Boltzmann, la mémoire temporelle hiérarchique, les réseaux de neurones impulsionnels (*Spiking Neural Networks* ou SNNs), les réseaux neuronaux à capsule (CapsNet), voire certains réseaux superficiels adaptés[697].

De leur côté, Russin, O'Reilly et Bengio[698] ont relevé des similitudes intéressantes entre les défis qui se posent présentement dans la recherche en apprentissage profond et certaines fonctions caractéristiques de notre cortex préfrontal (CPF), partie antérieure du lobe frontal de notre cerveau et la partie la plus développée de notre néocortex. En effet, les tâches pour lesquelles les systèmes intelligents demeurent sous-performants constituent des fonctions dites exécutives auxquelles notre cortex préfrontal joue un rôle clé, telles que le transfert de connaissance aux nouveaux domaines, la représentation de structures systématiques ou de composition, la planification efficiente et le raisonnement abstrait. Les auteurs recommandent l'élaboration de méthodes d'apprentissage automatique calquées sur les principes et biais inductifs à l'œuvre dans notre cortex préfrontal.

Structures sous-corticales –Malgré ses grandes capacités d'abstraction, le néocortex n'est pas la seule région pertinente, surtout lorsqu'on a en vue non pas uniquement une intelligence artificielle au sens étroit, mais aussi générale (renvoi au Chapitre 1), notamment dans le développement de la robotique et des systèmes autonomes[699]. Les structures sous-corticales, situées anatomiquement en dessous du cortex cérébral, jouent un rôle essentiel dans la cognition sociale, dont la mémoire (hippocampe), la régulation émotionnelle (complexe amygdalien), les circuits de récompense (ganglions de la base) et la coordination motrice (cervelet).

---

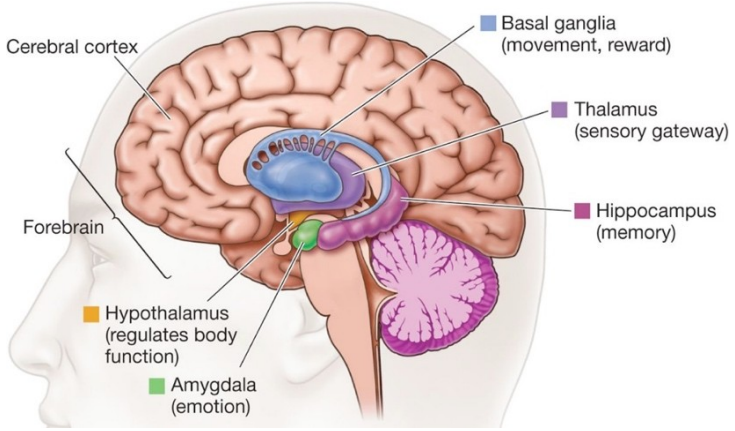
696 James CR Whittington et Rafal Bogacz, « Theories of Error Back-Propagation in the Brain » (2019) 23:3 *Trends in Cognitive Sciences Review* 235, doi : <doi.org/10.1016/j.tics.2018.12.005>.

697 Bernd Illing, Wulfram Gerstner et Johanni Brea, « Biologically plausible deep learning – But how far can we go with shallow networks? » (2019) 118 *Neural Networks* 90, doi : <doi.org/10.1016/j.neunet.2019.06.001>.

698 Jacob Russin, Randall C. O'Reilly et Yoshua Bengio, « Deep Learning Needs a Prefrontal Cortex », *Bridging AI and Cognitive Science*, ICLR 2020, en ligne : <baicsworkshop.github.io/pdf/BAICS_10.pdf>.

699 Terrence J Sejnowski, « The unreasonable effectiveness of deep learning in artificial intelligence » (2020) 117:48 PNAS 30033, doi : <doi.org/10.1073/pnas.1907373117>.

Figure 21
Aperçu des principales structures souscorticales



Source : <quizlet.com/226063347/csd-309-cerebral-connections-and-subcortical-structures-flash-cards/>

Ainsi, en s'inspirant du fonctionnement des neurones miroirs intensément étudiés en neurosciences comme participant à notre processus d'apprentissage par imitation, de décodage et reconnaissance des intentions d'émotions d'autrui, quelques architectures artificielles ont été testées pour entraîner les machines à reconnaître des états émotionnels de base[700]) ou utilisées dans l'apprentissage de la locomotion robotique[701].

De l'apprentissage intégré des différentes fonctions cérébrales – Pearl (2016)[702] assimile les progrès de l'apprentissage automatique fondés sur les modèles probabilistes à la théorie de notre évolution vue sous l'angle de la sélection naturelle. La nature statistique et probabiliste de l'apprentissage, tout comme une sélection naturelle au hasard des mutations génétiques aléatoires, se caractérise par une « extrême lenteur » pour parvenir à un résultat qui soit en adéquation avec les besoins de l'environnement. Selon Pearl (2016), ce qui aurait permis à l'humain d'accélérer cette évolution naturelle aléatoire est notre capacité à se représenter mentalement les différentes hypothèses et issues possibles. L'usage de ces raisonnements dits contrefactuels de type *what if ?* nous permet de valider le bien-fondé de nos choix et les relations de causalité sous-tendant des séquences d'événement sans toujours prendre action ou procéder par essais-erreurs. Pourtant, cette capacité à se former une imagerie mentale sur sa manière d'agir et ses choix est absente de la machine. Les travaux ayant montré que le raisonnement contrefactuel engage différentes structures cérébrales impliquées dans la simulation mentale, le

---

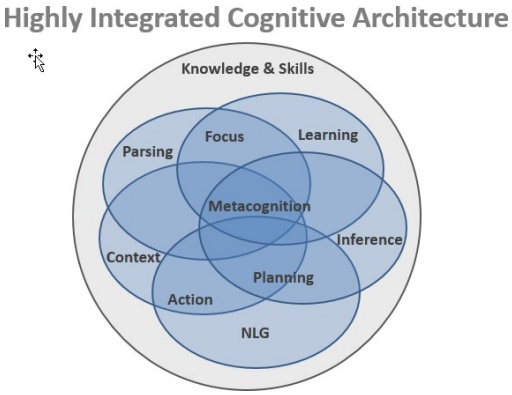
700 Faisal Rehman et al, « Design and Development of AI-based Mirror Neurons Agent towards Emotion and Empathy » (2020) 11:3 *International Journal of Advanced Computer Science and Applications* 386, en ligne : <usir.salford.ac.uk/id/eprint/56970>.

701 Junpei Zhong, Cornelius Weber et Stefan Wermter, « Robot Trajectory Prediction and Recognition Based on a Computational Mirror Neurons Model » dans Timo Honkela et al, dir, *Artificial Neural Networks and Machine Learning – ICANN 2011*, part II, Springer, 2011, 333, doi : <doi.org/10.1007/978-3-642-21738-8_43>.

702 Judea Pearl, *Theoretical Impediments to Machine Learning. A Position Paper*, 2016, en ligne : <web.cs.ucla.edu/~kaoru/theoretical-impediments.pdf>.

contrôle cognitif et le traitement des affects[703], dont le cortex préfrontal, les futures percées en apprentissage automatique pourraient reposer sur un apprentissage intégré du fonctionnement de différentes structures cérébrales. Cette approche sous-tend en effet la troisième vague de l'IA, axée sur le design et l'implémentation d'architectures cognitives devant accommoder l'ensemble des cognitions humaines de haut niveau en coordonnant les différents modules spécifiques aux différentes tâches[704] :

Figure 22
Intelligence artificielle (IA) de troisième vague



Source : Peter Voss, « The Third Wave of AI », *Becoming Human AI* (25 septembre 2017), en ligne : <becominghuman.ai/the-third-wave-of-ai-1579ea97210b>

Alors que l'on s'éloigne de la mathématisation abstraite pour se rapprocher également du cerveau primitif, c'est aussi remettre à l'avant-plan la nécessité de se (con)fondre dans son environnement, d'en survivre et de s'y adapter, militant pour une approche incarnée de l'apprentissage[705] insistant sur le rôle essentiel de l'interaction entre le corps (même artificiel) et son environnement pour le développement de la cognition (artificielle).

---

703 Nicole Van Hoeck, Patrick D Watson et Aron K Barbey, « Cognitive neuroscience of human counterfactual reasoning », (2015) 9 *Front Hum Neurosci* 420, doi : <doi.org/10.3389/fnhum.2015.00420>.

704 Peter Voss, « The Third Wave of AI », *Becoming Human AI* (25 septembre 2017), en ligne : <becominghuman.ai/the-third-wave-of-ai-1579ea97210b>; Gary Marcus, « The Next Decade in AI : Four Steps Towards Robust Artificial Intelligence », (2020), en ligne : <arxiv.org/abs/2002.06177>.

705 Larry Shapiro, « The Embodied Cognition Research Programme », (2007) 2:2 *Philosophy Compass* 338, doi : <doi.org/10.1111/j.1747-9991.2007.00064.x>; Sebastian Schneegans et Gregor Schöner, « 13 – Dynamic Field Theory as a Framework for Understanding Embodied Cognition », dans Paco Calvo et Antoni Gomila, dir, *Handbook of Cognitive Science*, coll « Perspectives on cognitive science », *Elsevier Science*, 2008, 241, doi : <doi.org/10.1016/B978-0-08-046616-3.00013-X>; Pierre de Loor, Alain Mille et Mehdi Réguigne-Khamassi, « Intelligence artificielle : l'apport des paradigmes incarnés », (2015) 64:2 *Revue de l'Association pour la recherche cognitive* 27, doi : <doi.org/10.3406/intel.2015.1011>.

Pour aller plus loin :

Ananthaswamy, A., « Deep Neural Networks Help to Explain Living Brains », Quanta Magazine (28 octobre 2020), en ligne : <www.quantamagazine.org/deep-neural-networks-help-to-explain-living-brains-20201028/>

Chen, L., « AI and biological consciousness – is there a connection ? », Young Scientists Journal (22 juillet 2020), en ligne : <ysjournal.com/ai-and-biological-consciousness-is-there-a-connection/>

D'Avila Garcez, A. et L.C. Lamb, « Neurosymbolic AI : The 3rd Wave » (2020), en ligne : <arXiv:2012.05876>

Dickson, B., « Artificial neural networks are more similar to the brain than they get credit for », TechTalks (22 juin 2020), en ligne : <bdtechtalks.com/2020/06/22/direct-fit-artificial-neural-networks/>

Howard, N. et al., « BrainOS : A Novel Artificial Brain-Alike Automatic Machine Learning Framework » (2020) Front Comput Neurosci, doi : <doi.org/10.3389/fncom.2020.00016>

Savage, N., « How AI and neuroscience drive each other forwards » (2019) 571 Nature S15, doi : <doi.org/10.1038/d41586-019-02212-4>

VanRullen, R., « Perception Science in the Age of Deep Neural Networks » (2017) Front Psychol, doi : <doi.org/10.3389/fpsyg.2017.00142>

5. De la robotique développementale

Les robots cognitifs sont dotés d'une architecture matérielle et logicielle qui leur permettent de dépasser l'automatisme ou la simple réactivité des robots dits de première et de deuxième génération, pour se comporter de manière intelligente dans un monde complexe et dynamique[706], allant de la locomotion bipède et l'évitement dynamique d'obstacles aux robots de service.

Dès 1950, Alan Turing[707] a proposé le concept d'une machine-enfant (*child machine*) qui soit en mesure d'apprendre par conditionnement opérant, en associant, à la manière de jeunes enfants, un (modèle de) comportement particulier à leurs conséquences positives (renforcements) ou indésirables (punitions). Il a fallu toutefois attendre jusqu'à l'orée des années 1980 pour que l'approche incarnée trouve preneur au sein de la communauté : avec les progrès de l'apprentissage « automatique », des chercheurs commençaient à entrevoir l'idée voulant que des comportements complexes puissent résulter non pas nécessairement d'un système autonome complexe, mais de la complexité de son environnement[708].

---

706 Christophe Sabourin, *Systèmes cognitifs artificiels : du concept au développement de comportements intelligents en robotique autonome*, Université Paris Est Crétail, 2016, en ligne : <hal.archives-ouvertes.fr/tel-01352195>; Yanfei Liu et al, « Cognitive Modeling for Robotic Assembly/Maintenance Task in Space Exploration », dans *International Conference on Applied Human Factors and Ergonomics*, Los Angeles (CA), Springer, 143, doi : <doi.org/10.1007/978-3-319-60642-2_13>.

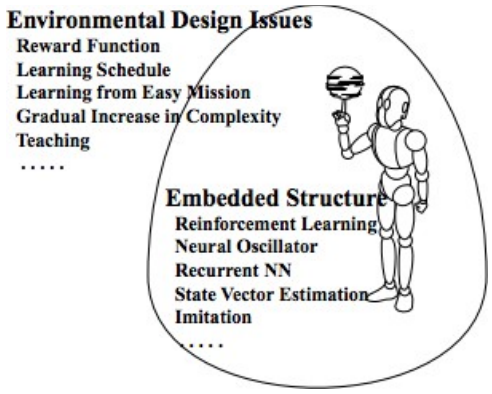
707 Alan M Turing, *supra* note 99.

708 Rodney A Brooks, « A Robust Layered Control System for a Mobile Robot » (1986) 2:1 IEEE *Journal of Robotics and Automation* 14, doi : <doi.org/10.1109/JRA.1986.1087032>.

Plutôt qu'une reproduction exacte de la structure de notre système nerveux, il s'agit de construire une intelligence / cognition artificielle dans l'action, en s'inspirant des théories du développement humain, voire animal. Après tout, la connaissance (vraie) se construit en mettant à l'épreuve les théories apprises dans la pratique; les représentations mentales (théoriques) ne correspondent pas nécessairement à ce que l'on peut rencontrer dans « la vraie vie ».

La robotique développementale (« *DevRob* ») ou épigénétique cherche ainsi non pas tant à programmer un système cognitif artificiel qui soit en mesure d'accomplir des tâches précises, mais plutôt à entretenir un processus de développement évolutif par lequel un système artificiel se construit avec son environnement. Elle adopte une approche constructiviste axée sur une adaptation progressive du robot aux tâches complexes requises des situations dynamiques qui peuvent être difficiles à modéliser d'avance[709]. Dans cet ordre d'idées, le cycle entraînement-apprentissage s'entretient de lui-même en laissant au robot une occasion d'apprendre de ses erreurs et des conséquences de ses actions (choix) :

Figure 23
Apprentissage incarné : Interaction entre un robot et son environnement



Source : Asada et al., 2001

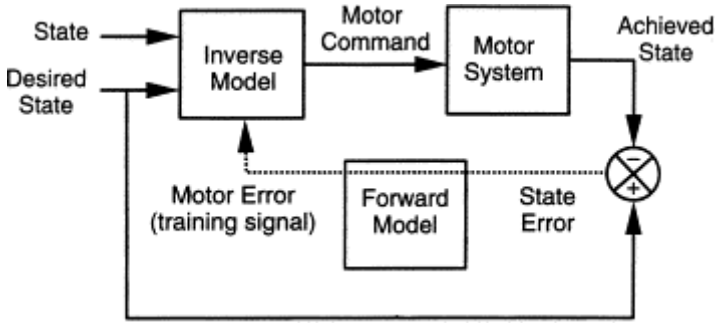
Cette incarnation s'inspire encore une fois des modèles internes relatifs au fonctionnement de notre système propriomoteur. Ce dernier repose sur la complémentarité des modèles prédictif et inverse pour associer l'état présent du système moteur à ses états futurs :

- d'un côté, le modèle prédictif (*forward model*) anticipe la rétroaction sensorielle attendue d'une commande motrice spécifique;

709 Minoru Asada et al, « Cognitive Developmental Robotics As a New Paradigm for the Disgn of Humanoid Robots » (2001) 37:2-3 Robotics and Autonomous System 185, doi : <doi.org/10.1016/S0921-8890(01)00157-9>.

- de l'autre côté, le modèle inverse (*inverse model*) détermine la commande motrice nécessaire pour arriver à cette fin, en raisonnant « à l'inverse » à partir de la connaissance de l'objectif à atteindre (résultat du modèle prédictif).

Figure 24
Complémentarité des modèles prédictif et inverse



Source : Miall et Wolpert (1996)[710]

Un autre aspect essentiel au développement des robots cognitifs est la mise en place d'un système de mémoire perceptive ou sensorielle, afin de leur permettre d'appréhender et d'assimiler les différentes potentialités d'action (*affordances*) présentes dans l'environnement[711], soit l'ensemble des propriétés de différents objets, les interactions possibles entre le robot et les autres agents de son environnement ainsi que leurs relations d'interdépendance[712].

À la différence de la robotique traditionnelle, les tâches qu'un robot développemental devra apprendre sont inconnues lors de la programmation; les représentations spécifiques à la tâche sont générées et mises à jour au fur et à mesure des interactions vécues[713]. Plus particulièrement, la robotique développementale cherche, par l'approche incarnée, à développer des habiletés perceptuelles et sensorimotrices avancées qui restent un maillon faible de la cognition artificielle[714], y compris la conduite autonome et l'apprentissage développemental du langage.

---

710 RC Miall et DM Wolpert, « Forward Models for Physiological Motor Control » (1996) 9:8 *Neural Networks* 1265, doi : <doi.org/10.1016/S0893-6080(96)00035-4>.

711 James J Gibson, « The Theory of Affordances » dans Robert E Shaw et John Bransford, *Perceiving, Acting, and Knowing : Toward an Ecological Psychology*, Hillsdale (NJ), Lawrence Erlbaum Associates, 1977, aux pp 67, 127.

712 Marc Kammer et al, « A Perceptual Memory System for Affordance Learning in Humanoid Robots » dans Timo Honkela et al, dir, *Artificial Neural Networks and Machine Learning – ICANN 2011*, part II, Springer, 2011, 349, doi : <doi.org/10.1007/978-3-642-21738-8_45>.

713 Juyang Weng, « Developmental Robotics : Theory and Experiments » (2004) 1:2 *International Journal of Humanoid Robotics* 199, doi : <doi.org/10.1142/S0219843604000149>.

714 Bruno Lara et al, « Embodied Cognitive Robotics and the Learning of Sensorimotor Schemes » (2018) 26:5 *Adaptive Behavior*, doi : <doi.org/10.1177/1059712318780679>.

Plus récemment, des chercheurs commencent à entraîner des robots à grimacer à l'application d'une charge électrique[715] ou à réagir rapidement à l'application de la force physique en initiant un processus d'auto-réparation des structures endommagées[716]; cet apprentissage des stimuli douloureux constituerait un premier pas vers le développement de l'empathie chez les robots en vue d'une interaction plus affective (et expressive) avec les humains.

L'environnement au sein duquel un robot cognitif évolue pouvant comprendre non seulement des humains, d'autres êtres vivants ou encore des objets inanimés, mais par ailleurs d'autres robots cognitifs, l'interaction collaborative entre les différents robots cognitifs en vue de la réalisation d'un objectif commun, ainsi que des tâches d'une complexité accrue, promet l'émergence de systèmes extrêmement adaptatifs et évolutifs avec éventuellement le développement de nouvelles fonctionnalités que les développeurs humains ne pourraient pas nécessairement anticiper (Parker, Schneider et Schultz, 2005[717]; Levi et Kernbach, 2010[718]; Gautam et Mohan, 2012[719]) : pensons notamment aux missions de sauvetage ou de reconnaissance militaire, au transport et à la livraison parallèles des marchandises, voire à la prestation collaborative de certains services. Les habiletés sociales constituent en effet les premières pierres d'assise de l'apprentissage par les pairs. De l'apprentissage par imitation à l'apprentissage collaboratif (Tomasello, Kruger et Ratner, 1993[720]; MacLean, 2016[721]) émerge peu à peu une culture de groupe, sorte d'épiphénomène qui entretient un *modus vivendi* propice au développement d'une intelligence collective – alchimie des compétences, habiletés et aptitudes complémentaires allant au-delà de la somme des intelligences isolées.

Cela étant, l'approche incarnée connaît ses limites en ce que l'interaction avec l'environnement n'explique pas la genèse de tous nos états mentaux. En effet, plusieurs expériences subjectives internes et états cognitifs (non perceptuels) émergent précisément lors des périodes de sommeil

---

715 Christopher McFadden, « Researchers in Japan to Make Artificial Skin That Can Feel Pain », *Interesting Engineering* (17 février 2020), en ligne : <interestingengineering.com/researchers-in-japan-to-make-artificial-skin-that-can-feel-pain>.

716 Rohit Abraham John et al, « Self healable neuromorphic memtransistor elements for decentralized sensory signal processing in robotics » (2020) 11 *Nature Communications*, doi : <doi.org/10.1038/s41467-020-17870-6>.

717 Lynne E Parker, Frank E Schneider et Alan C Schultz, *Multi-Robot Systems. From Swarms to Intelligent Automata. Volume III*, Proceedings from the 2005 International Workshop on Multi-Robot Systems, Springer, 2005, doi : <doi.org/10.1007/1-4020-3389-3>.

718 Paul Levi et Serge Kernbach, dir, *Symbiotic Multi-Robot Organism. Reliability, Adaptability, Evolution*, Springer, 2010, doi : <10.1007/978-3-642-11692-6>.

719 Avinash Gautam et Sudeept Mohan, « A review of research in multi-robot systems » dans *2012 IEEE 7th International Conference on Industrial and Information Systems (ICIIS)*, Chennai (Inde), 2012, 1, doi : <doi.org/10.1109/ICIInfS.2012.6304778>.

720 Michael Tomasello, Ann Cale Kruger et Hilary Horn Ratner, « Cultural learning » (1993) 16:3 *Behavioral & Brain Sciences* 495.

721 Evan L MacLean, « Unraveling the evolution of uniquely human cognition » (2016) 113:23 PNAS 6348, doi : <doi.org/10.1073/pnas.1521270113>.

caractérisées par une suspension partielle de nos rapports sensitivo-moteurs avec le milieu environnant[722]. Nos états de rêve, de transe et de conscience altérée modulent pourtant (aussi) notre santé cognitive et pourraient bien contribuer à nous définir en tant qu'espèce. La pensée abstraite et notre conscience réflexive – soit la capacité de réfléchir sur ses pensées ainsi que de se rendre compte et d'évaluer (de manière critique) ses propres processus cognitifs – font partie intégrante de notre métacognition et marquerait, dans une perspective évolutive, la spécificité des humains modernes par rapport aux autres espèces (proto-humaines). La créativité et l'innovation.

Pour en savoir plus :

Hoffmann, M. et R. Pfeifer, « Robots as powerful allies for the study of embodied cognition from the bottom up » dans Newen, A., L. De Bruin et S. Gallagher, dir, *The Oxford Handbook of 4E Cognition*, 2018, doi : <doi.org/10.1093/oxfordhb/9780198735410.013.45>

Nguyen, S.M. et al., « Robots Learn Increasingly Complex Tasks with Intrinsic Motivation and Automatic Curriculum Learning » (2021) 35 Künstl Intell 81, doi : <doi.org/10.1007/s13218-021-00708-8>

Pezzulo, G. et al., « The mechanics of embodiment : a dialog on embodiment and computational modeling » (2011) Front Psychol, doi : <doi.org/10.3389/fpsyg.2011.00005>

Pfeifer, R. et J. Bongard, How the Body Shapes the Way We Think. A New View of Intelligence, MIT Press, 2006

Philippsen, A., « Goal-Directed Exploration for Learning Vowels and Syllables : A Computational Model of Speech Acquisition » (2021) 35 Künstl Intell 53, doi : <doi.org/10.1007/s13218-021-00704-y>

Ziemke, T., « The body of knowledge : On the role of the living body in grounding embodied cognition » (2016) 148 Biosystems 4, doi : <doi.org/10.1016/j.biosystems.2016.08.005>

6. Vers une intelligence artificielle (IA) polyvalente et métacognitive ?

En dépit des progrès fulgurants qu'a connus l'apprentissage automatique jusqu'à présent, les systèmes intelligents demeurent sous-performants quand vient le temps de reproduire de manière synthétique et polyvalente plusieurs aspects de l'intelligence humaine.

Le problème n'est pas nouveau. Le fameux test de Turing (1950) avait déjà souligné le défi de faire imiter par la machine un comportement socialement intelligent comme ce qui serait attendu d'un humain. Alors qu'un enfant (humain) apprend presque instinctivement à craindre le feu de sa cheminée comme de la marmite, il faudrait à une intelligence artificielle soi-disant augmentée plusieurs essais-erreurs sous différentes conditions d'éclairage, d'humidité, d'angles d'approche et d'instruments chauffés pour arriver à la même conclusion. Encore qu'appliquer le qualificatif « craindre », comme tous ceux dénotant un ressenti, une émotion, un sentiment, de l'empathie, à ce qui a été artificiellement construit, s'avère inexact. Quelque chose manque à la machine, un quelque chose qui semble échapper obstinément à la formalisation mathématique.

[722] James F Pagel, *Dream Science: Exploring the Forms of Consciousness*, Academic Press, 2014 à la p 22.

Dès 1959, McCarthy[723] a soulevé l'intérêt de faire en sorte que les programmes puissent apprendre de leurs expériences aussi efficacement que font les humains. Pour McCarthy, le sens commun renvoie à certains processus élémentaires de raisonnement verbal que tout humain non faible d'esprit pourrait maîtriser, de sorte qu'il serait possible d'affirmer qu'un programme est doté de sens commun lorsqu'il peut automatiquement déduire pour lui-même une gamme suffisamment large de conséquences immédiates de tout ce qu'on lui dit et de ses connaissances antérieures[724].

En caractérisant le « sens commun » comme une forme d'intelligence sociale dénotant une connaissance approfondie des comportements de ses congénères, exiger d'une intelligence artificielle une connaissance sophistiquée de l'humain paraît à première vue fausser le débat. Après tout, exiger de l'humain une puissance calculatoire semblable à celle de la machine est aussi irréaliste. Il n'en demeure pas moins que l'intérêt d'instiller une dose de « sens commun » à cette puissance calculatoire réside par ailleurs dans le fait qu'il permettrait d' « approximer » la résolution de problèmes en présence de données incomplètes, ambiguës ou imprécises. Appliqué à l'intelligence artificielle, le sens commun s'avère moins une intelligence « sociale » que « relationnelle », soit une connaissance pas nécessairement bien articulée ou définie de ce qui est communément rencontré ou attendu dans un environnement donné, ainsi que la capacité d'orienter ses choix en fonction de ces conditions externes.

Jusqu'à présent, les initiatives entreprises pour résoudre les situations d'ambiguïté restent cantonnées dans des domaines isolés, notamment en matière de traitement du langage naturel[725]. De plus en plus, des chercheurs[726] se penchent sur la vision artificielle pour capter les détails, nuances et associations latentes qui s'articulent difficilement par les mots. Cela étant, une approche unifiée pour entraîner l'algorithme à faire preuve de « sens commun » en toute circonstance fait défaut[727].

De toutes les interactions de la machine avec l'environnement, le comportement humain est sans doute l'un des plus imprévisibles, notamment par la richesse et sophistication de nos états

---

723 John McCarthy, supra note 374 : « Our Ultimate Objective is to Make Programs that Learn from their Experience as Effectively as Humans Do ».

724 *Ibid.* : « … a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows.»

725 Dan Roth, « Learning to resolve natural language ambiguities : a unified approach » (1998) AAAI 806, en ligne : <dl.acm.org/doi/10.5555/295240.295894>.

726 Tom Simonite, « Facebook's AI Chief : Machines Could Learn Common Sense from Video » (2017) MIT Technology Review, en ligne : <www.technologyreview.com/2017/03/09/153343/facebooks-ai-chief-machines-could-learn-common-sense-from-video>.

727 Pour des efforts récents en ce sens, voir Marco F Cusumano-Towner et al, « Gen : a general-purpose probabilistic programming system with programmable inference » (2019) PLDI 221, doi : <doi.org/10.1145/3314221.3314642>.

mentaux[728]. C'est l'une des raisons pour laquelle modéliser une théorie de l'esprit rudimentaire chez la machine suscite dorénavant un intérêt soutenu. La théorie de l'esprit est un concept développé en psychologie[729] qui renvoie à cette aptitude cognitive, constatée dès notre plus jeune âge[730], d'attribuer les états mentaux à autrui et de comprendre leurs intentions, pensées et émotions. En robotique, les travaux en cours consistent à intégrer une simulation du monde dans la machine afin d'y donner un espace de représentations des différents scénarios envisageables avec la possibilité de tester et partant d'anticiper, par l'implémentation d'une logique contrefactuelle, les conséquences prévisibles des différents choix d'actions et de réactions[731]. Une des applications immédiates de cette simulation est de renforcer la sécurité des interactions homme-machine en entraînant les robots à mieux anticiper les réactions des humains en cas d'urgence ou lors de la conduite autonome (p.ex. système anticollision).

De la théorie de l'esprit à une connaissance de soi (« *self-awareness* ») artificielle, développer chez un robot la capacité – limitée – de prendre un recul réflexif sur ses propres processus cognitifs (métacognition)[732] fait ses premiers pas au sein de la communauté technoscientifique de l'IA. La connaissance de soi renvoie à l'aptitude à prendre en compte son propre comportement et ses états mentaux dans l'appréciation d'une situation, et éventuellement les confronter avec les attentes, les actions et les réactions (mentales) d'autrui. L'avantage d'apprendre aux robots « à se connaître » serait, dans un premier temps, de renforcer leur aptitude s'adapter aux changements environnementaux imprévus[733]. Prolongée dans le temps

---

728 Sur cette question, voir notamment : Gary A Cziko, « Unpredictability and Indeterminism in Human Behavior : Arguments and Implications for Educational Research » (1989) 18:3 *Educational Researcher* 17, doi : <doi.org/10.2307/1174887>; Michael Scriven, « An Essential Unpredictabilty in Human Behavior » dans Benjamin B Wolman et Ernest Nagel, dir, *Scientific Psychology : Principes and Approaches*, 1965, 411

729 David Premack et Guy Woodruff, « Does the chimpanzee have a theory of mind? » (1978) 1:4 *Behavioral & Brain Sciences* 515, doi : <doi.org/10.1017/S0140525X00076512>.

730 À ce jour, l'existence d'une théorie de l'esprit chez les animaux reste controversée : Derek C Penn et Daniel J Povinelli, « On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind' » (2007) 362:1480 *Philos Trans R Soc Lond B Biol Sci* 731, doi : <doi.org/10.1098/rstb.2006.2023>; Joseph Call et Michael Tomasello, « Does the chimpanzee have a theory of mind? 30 years later » (2008) 12:5 *Trends in Cognitive Sciences* 187, doi : <doi.org/10.1016/j.tics.2008.02.010>; J David Smith, « Inaugurating the Study of Animal Metacognition » (2010) 23:3 *Int J Comp Psychol* 401.

731 Alan FT Winfield, « Experiments in Artificial Theory of Mind : From Saftey to Story-Telling » (2018) *Front Robot AI*, doi : <doi.org/10.3389/frobt.2018.00075>; Christian Blum, Alan FT Winfield et Verena V Hafner, « Simulation-Based Internal Models for Safer Robots » (2018) *Front Robot AI*, doi : <doi.org/10.3389/frobt.2017.00074>.

732 Stephen M Fleming et Christopher D Frith, dir, *The Cognitive Neuroscience of Metacognition*, Springer, 2014, doi : <doi.org/10.1007/978-3-642-45190-4>.

733 Antonio Chella et al, « Developing Self-Awareness in Robots via Inner Speech » (2020) *Front Robot AI*, doi : <doi.org/10.3389/frobt.2020.00016>.

par la mémoire, cette connaissance de soi permet d'affiner les prédictions, de réévaluer, sur une base rétrospective, ses choix antérieurs et de s'ajuster en conséquence[734].

Pour établir l'existence d'une conscience de soi chez de très jeunes enfants ou des animaux non humains, le test de référence traditionnel consiste à placer le sujet devant un miroir et à tester son aptitude à y reconnaître son propre reflet (plutôt qu'un intrus) en observant ses réactions[735]. Alors que plusieurs espèces animales obtiennent un résultat positif au test du miroir[736], il est également possible d'entraîner un robot à le faire, notamment par le biais d'une correspondance visuo-kinesthésique acquise entre la représentation qu'il se fait de son apparence et la perception de son image reflétée dans un miroir[737]. D'autres ont cherché à développer une voix intérieure (« *inner speech* ») chez des robots pour les aider à renforcer leur conscience situationnelle et à décrire leurs actions, que les robots reconnaissent à travers les capteurs proprioceptifs et perceptifs[738]. Il faudrait également doter les robots d'un système de motivations internes qui les amènent développer leurs propres références ainsi que de nouvelles initiatives[739]. C'est en effet avec le développement d'une conscience réflexive et volontaire qu'une machine pourra, un jour, non seulement modifier (optimiser) ce qu'elle fait, mais aussi (ré)évaluer ce qu'elle doit faire, revoir ses objectifs, déceler de nouveaux problèmes à résoudre, trouver des solutions inédites, déjouer nos prédictions, se forger un monde à leur image, dépasser un anthropomorphisme qui leur a été imposé pour (enfin) surprendre l'humain par une autonomie, « présence d'esprit » et créativité insoupçonnées[740].

---

734 HC Lou, JP Changeux et A Rosenstand, « Towards a cognitive neuroscience of self-awareness » (2017) 83 *Neuroscience & Biobehavioral Reviews* 765, doi : <doi.org/10.1016/j.neubiorev.2016.04.004>.

735 Gordon G Gallup, Jr, « Chimpanzees : Self-Recognition » (1970) 167:3914 *Science* 86, doi : <doi.org/10.1126/science.167.3914.86> et « Self recognition in primates : A comparative approach to the bidirectional properties of consciousness » (1977) 32:5 *American Psychologist* 329, doi : <doi.org/10.1037/0003-066X.32.5.329>.

736 Dont les dauphins, les éléphants, certaines espèces de singes voire d'oiseaux : Michael D Breed et Janice Moore, « Chapter 6 – Cogntion » dans *Animal Behavior*, 2e éd, Elsevier, 2015, 175, doi : <doi.org/10.1016/C2013-0-14008-1>.

737 Matej Hoffmann et al, « Robot in the Mirror : Toward an Embodied Computational Model of Mirror Self-Recognition » (2021) 35 *Künstliche Intelligenz* 37, doi : <doi.org/10.1007/s13218-020-00701-7>.

738 Antonio Chella et Arianna Pipitone, « A cognitive architecture for inner speech » (2020) 59 *Cognitive Systems Research* 287, doi : <doi.org/10.1016/j.cogsys.2019.09.010>.

739 Raja Chatila et al, « Toward Self-Aware Robots » (2018) *Front Robot AI*, doi : <doi.org/10.3389/frobt.2018.00088>.

740 Pagel et Kirshtein, *supra* note 675 aux pp 61, 62 : « A strong artificial consciousness would be able to take the further step of having the capacity to modify not only how it does what it does, but also to modify what it does.[…] In order to consider any system an autonomous entity it must not only be aware of and able to manipulate its environment, it must also be in pursuit of its own agenda and be able to affect what it senses in the future. If desired, it could be within the capacity of such a system to appear as human if that met the requirements of that system. Such an artificial conscious being could appear to be conscious. It would have the capacity to act and behave as if it were a conscious human being. »

Pour aller plus loin :

Buts, M.V., « Towards Strong AI » (2021) 35 Künstl Intell 91, doi : <doi.org/10.1007/s13218-021-00705-x>

Cox, M.T., « Metacognition in computation : A selected research review » (2005) 169:2 Artificial Intelligence 104, doi : <doi.org/10.1016/j.artint.2005.10.009>

Czerwinski, M., J. Hernandez et D. McDuff, « Building an AI That Feels. AI systems with emotional intelligence could learn faster and be more helpful », *IEEE SPECTRUM* (30 avril 2021), en ligne : <spectrum.ieee.org/artificial-intelligence/machine-learning/building-an-ai-that-feels>

Fjelland, R., « Why general artificial intelligence will not be realized » (2020) 7 Humanities & Social Sciences Communications, doi : <doi.org/10.1057/s41599-020-0494-4>

Freed, S., *AI and Human Thought and Emotion*, Boca Raton, 2019, doi : <doi.org/10.1201/9780429001123>

Ng, G.W. et W.C. Leung, « Strong Artificial Intelligence and Consciousness » (2020) 7:1 Journal of Artificial Intelligence & Consciousness 63, doi : <doi.org/10.1142/S2705078520300042>

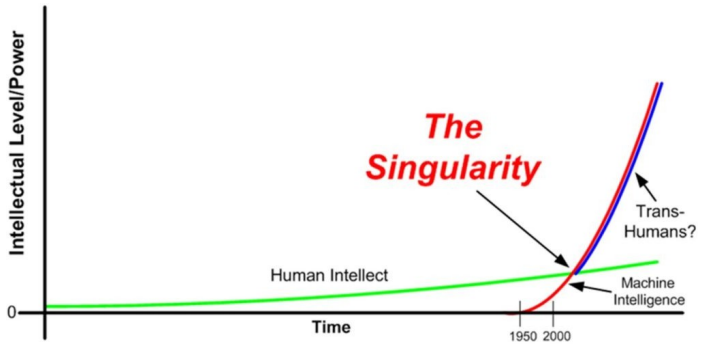
Pagel, J.F. et P. Kirshtein, *Machine Dreaming and Consciousness*, Academic Press, 2017

Siarri, P., « Can you teach humor to an AI ? », *Medium* (4 avril 2019), en ligne : <medium.com/futuresin/can-you-teach-humor-to-an-ai-13ef4cff6cac>

7. De la singularité technologique

Depuis l'avènement de l'intelligence artificielle (IA), le spectre des singularités technologiques a inspiré plusieurs projections surenthousiastes qui chevauchent la science-fiction. Le concept de « singularité technologique » désigne un stade hypothétique du développement technologique où l'intelligence artificielle surpasserait l'intelligence humaine et s'émanciperait de tout contrôle jusqu'à s'entretenir d'elle-même :

Figure 25
L'atteinte de la singularité technologique



Source :Levinson, 2017[741]

741 Frank H Levinson, « Man and Superman : Life Near An Approaching Technology Singularity », *One Million by One Million Blog* (5 juillet 2017), en ligne : <www.sramanamitra.com/2017/07/05/man-and-superman-life-near-an-approaching-technology-singularity/>.

S'inspirant du spectre des singularités initiale et finale en astrophysique – marquant une réelle discontinuité dans notre appréhension du cosmos, la singularité technologique serait théoriquement possible mais pèche par un excès d'imprécision, tant dans la gamme de conséquences qu'elle implique que l'horizon temporel de sa réalisation[742].

Pour en savoir plus sur les mythes et réalités entourant les récits de singularité technologique, voir :

Ganascia, J.-G., Le Mythe de la Singularité. Faut-il craindre l'intelligence artificielle?, Seuil, 2017

Latorre, J. I., « The Singularity », CCCB LAB, 14 janvier 2019, en ligne : <lab.cccb.org/en/the-singularity/>

Pandya, J., « The Troubling Trajectory of Technological Singularity », Forbes (10 février 2019), en ligne : <www.forbes.com/sites/cognitiveworld/2019/02/10/the-troubling-trajectory-of-technological-singularity/>

Tegmark, M., Life 3.0 : Being Human in the Age of Artificial Intelligence, Knopf, 2017

Alors que la littérature fait planer le spectre d'une « machine ultraintelligente »[743], d'une intelligence artificielle générale qui surpasserait celle de l'humain[744] ou encore l'avènement de futures machines qui seront « humaines sans être biologiques »[745], il n'y aurait moins à craindre d'une superintelligence artificielle qui excellerait dans toutes les facettes de l'intelligence humaine, qu'une accélération incontrôlée des progrès technologiques aux conséquences imprévisibles et potentiellement dangereuses. En effet, point n'est besoin d'une superintelligence (générale) artificielle pour faire engager des pronostics d'ordre eschatologique : la menace nucléaire, les désastres environnementaux et les catastrophes naturelles sont autant d'épées de Damoclès suspendues au-dessus de l'humanité que le recours à des systèmes d'armes létaux complètement autonomes, robots tueurs dont un déploiement malavisé changeront au mieux la face de la guerre et au pire, repartiront notre civilisation au temps zéro.

À ce jour, le risque principal que pose le recours à des systèmes d'armes létaux complètement autonomes réside non pas dans la formation – inattendue – d'une « intention » artificielle malveillante de se retourner contre ses créateurs, mais plutôt dans le risque de manipulations (humaines) adverses par l'ennemi – y compris le piratage, les dysfonctionnements / défaillances du système, ou encore des réactions inattendues face à certaines conditions

742 Aux dernières nouvelles, l'échéance, maintes fois repoussée, se situerait aux alentours de 2060 : AI Multiple, *995 experts opinion : AGI/singularity by 2060 [2021 update]*, 2 février 2021, en ligne : <research.aimultiple.com/artificial-general-intelligence-singularity-timing/>.

743 Irving John Good, « Speculations Concerning the First Ultraintelligent Machine » (1966) 6 *Advances in Computers* 31, doi : <doi.org/10.1016/S0065-2458(08)60418-0>.

744 Vernor Vinge, « The Coming Technological Singularity : How to Survive in the Post-Human Era » (1993) *VISION-21 Symposium*, en ligne : <edoras.sdsu.edu/~vinge/misc/singularity.html>.

745 Ray Kurzweil, *The Singularity is Near : When Humans Transcend Biology*, Viking, 2005.

environnementales[746]. Ces facteurs de risque peuvent être atténués à l'aide de contre-vérifications, évaluations et contrôles des systèmes, sans être éliminés complètement. Il nous incombe ultimement (à l'humain) de pondérer les risques découlant du déploiement des systèmes autonomes contre l'utilité militaire attendue de ces opérations.

Pour en savoir plus sur les systèmes d'armes létaux autonomes, voir aussi :

Abaimov, S. et M. Martellini, « Artificial Intelligence in Autonomous Weapon Systems » dans Martellini, M. et R. Trapp, *21st Century Prometheus. Managing CBRN Safety and Security Affected by Cutting-Edge Technologies*, Springer, 2020, 141, doi : <doi.org/10.1007/978-3-030-28285-1_8>

Fernandez, J., « Les systèmes d'armes létaux autonomes : en avoir (peur) ou pas ? » (2016) 6:791 Revue Défense Nationale 133, doi : <doi.org/10.3917/rdna.791.0133>

Pedron, S.M. et J. de Arimateia da Cruz, « The Future of Wars : Artificial Intelligence (AI) and Lethal Autonomous Weapon Systems (LAWS) » (2020) 2:1 International Journal of Security Studies, en ligne : <digitalcommons.northgeorgia.edu/ijoss/vol2/iss1/2>

Beaucoup plus insidieuses et redoutables pourraient devenir ces risques qui, subtilement mais sûrement, transforment notre société humaine jusqu'à la rendre méconnaissable, dont notre dépendance (inconsciente) à la technologie et aux biais algorithmiques, ainsi qu'une foi aveugle dans l'omniscience du *Big Data* et l'objectivité incontestée des mathématiques. Le défi de la régulation technologique se corse avec l'éventail des conséquences sociales (renvoi à la section pertinente de notre document de travail n° 2), économiques (), politiques (renvoi à la section pertinente de notre document de travail n° 2), éthiques (renvoi à la section pertinente de notre document de travail n° 2) et juridiques (renvoi à la section pertinente de notre document de travail n° 2) que la prolifération des systèmes d'AI pose d'ores et déjà à l'humanité, et ce, bien avant l'atteinte de toute singularité technologique.

Considérées sous cet angle, les fameuses lois de la robotique, telles que proposées par Asimov[747], risquent moins d'être enfreintes à la lettre (ne pas porter atteinte à un être humain, ne pas désobéir aux ordres donnés par les êtres humains, protéger son propre existence dans la mesure où cet impératif n'entre pas en contradiction avec les deux précédents) que par une omission déraisonnable des êtres humains à en assurer pleinement le respect, c'est-à-dire dans tous les contextes – évolutifs – pertinents et eu égard à l'objet de ces lois, à savoir assurer une évolution / cohabitation harmonieuse des sociétés humaines avec les progrès technologiques.

746 Paul Scharre, *Autonomous Weapons and Operational Risk*, Ethical Autonomy Project, Center for a New American Security, février 2016, en ligne : <s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf>.

747 Isaac Asimov, « Runaround » (mars 1942) Astouding Science Fiction, en ligne : <web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html>.

8. Conclusion

À mesure que les travaux pluridisciplinaires mettent en évidence le caractère protéiforme, multiple, quasi omniprésent de l'intelligence dans la nature, la maniabilité du concept met aussi en perspective les avancées de notre intelligence artificielle comme étant limitées aux tâches précises faisant appel aux fonctions cognitives supérieures. Plus récemment, le saut des algorithmes classiques à auto-apprenants a permis d'automatiser le processus d'apprentissage de la machine en l'émancipant, dans une certaine mesure, des règles programmées par l'humain. La part de l'humain y reste néanmoins cruciale, tant dans la détermination des objectifs de l'apprentissage (tâches à accomplir) que du choix des données initiales ou des conditions d'apprentissage auxquelles la machine est assujettie. Entre ces deux points de départ et d'arrivée, l'intelligence artificiellement construite perfectionne son apprentissage principalement par diverses méthodes s'inspirant du paradigme probabiliste en vue d'optimiser le traitement de l'information et d'apprivoiser (de quantifier) l'incertitude.

En lui-même, l'apprentissage probabiliste s'implémente de manière lourde et coûteuse (à coup d'essais-erreurs) en même temps qu'il n'est pas nécessairement adapté aux situations nouvelles, d'occurrence rare ou imprévues. Aussi, la machine peine à résoudre les situations d'ambiguïté et à tirer des inférences approximatives à partir des données incomplètes, dégradées ou imprécises. Ainsi, l'apprentissage profond excellerait uniquement dans des environnements contrôlés (prévisibles) et à l'égard des tâches précises.

Afin de dépasser ces limitations, il importerait tout d'abord de percer le plafond de verre des méthodes probabilistes en élargissant l'apprentissage autonome des données sur une base discrète vers un apprentissage autonome du contexte ainsi que des relations latentes entre les données. À cet effet, le fonctionnement de notre système nerveux reste le modèle (d'apprentissage) de référence par excellence, sans négliger son indispensable interrelation avec l'environnement. Après tout, que l'intelligence soit à base de silicone ou de carbone, l'éternel dilemme de la nature et de la culture reste d'actualité.

Sans doute, une question intrigante demeure : outre la possibilité de simplement modéliser notre gamme d'expériences subjectives (p.ex. émotions, hallucinations, rêves) dans la machine, comment permettre à cette dernière d'accéder à une réelle compréhension de celles-ci ? Certes, nuançant le test de Turing, la simple manipulation de symboles dans le respect des règles programmées n'équivaut pas, *a priori*, à une réelle connaissance[748]. De l'autre côté, les jeunes enfants apprennent d'abord par cœur et imitation avant de réellement comprendre. La conscience de ce qui nous motive est l'iceberg des processus neurochimiques et psychiques inconscients qui en balisent le fondement. Il aurait fallu à l'univers même des milliards d'années de nucléosynthèse primordiale pour faire accoucher, des poussières d'étoiles, la conscience cosmique de l'homme[749]. Ne pourrait-on pas éventuellement appliquer à une machine

---

748 John R Searle, *supra* note 58.

749 Hubert Reeves, Patience dans l'azur. L'évolution cosmique, 2e éd, Seuil, 1988.

intelligente cette vieille remarque du poète Riley (1849-1916) : « *When I see a bird that walks like a duck and swims like a duck and quacks like a duck, I call that bird a duck*. » ? Réservons notre jugement tant que le phénomène d'émergence de la conscience (artificielle) demeure un des mystères les plus fascinants à la croisée des sciences cognitives[750], de l'intelligence artificielle[751] et de la biocybernétique[752].

En attendant, indépendamment de l'atteinte éventuelle de la singularité technologique – annonçant l'ère révolutionnaire de la supermachine s'émancipant définitivement de l'humain, il y a lieu dès à présent de faire face à la nécessité d'expliciter le fonctionnement des algorithmes complexes ainsi que de s'atteler dès à présent aux diverses transformations sociétales que pose d'ores et déjà l'intégration de l'intelligence artificielle à notre quotidien.

---

750 Holk Cruse et Malte Schilling, « Mental States as Emergent Properties. From Walking to Consciousness » dans Thomas Metzinger et Jennifer Windt, dir, *Open Mind*, Francfort, PUB, 2015, doi : <doi.org/10.15502/9783958570436>; Todd E Feinberg et Jon Mallatt, « Phenomenal Consciousness and Emergence : Eliminating the Explanatory Gap » (2020) 11 *Front Psychol*, doi : <doi.org/10.3389/fpsyg.2020.01041>; Ramón Guevara, Diego M Mateos et José Luis Pérez Velázquez, « Consciousness as an Emergent Phenomenon : A Tale of Different Levels of Description » (2020) 22:9 *Entropy (Basel)* 921, doi : <doi.org/10.3390/e22090921>.

751 Jan Scheffel, « On the Solvability of the Mind-Body Problem » (2020) 30 *Axiomathes* 289, doi : <doi.org/10.1007/s10516-019-09454-x>; Riccardo Fesce, « Subjectivity as an Emergent Property of Information Processing by Neuronal Networks » (2020) *Front Neurosci*, doi : <doi.org/10.3389/fnins.2020.548071>; Elisabeth Hildt, « Artificial Intelligence : Does Consciousness Matter? » (2019) *Front Psychol*, doi : <doi.org/10.3389/fpsyg.2019.01535>.

752 David Rudrauf et al, « The Role of Consciousness In Biological Cybernetics : Emergent Adaptive and Maladaptive Behaviours in Artificial Agents Governed by the Projective Consnciousness Model » (2020), en ligne : <arxiv.org/abs/2012.12963>.

CHAPITRE 3
L'INTELLIGENCE ARTIFICIELLE (IA) ET LES PRESSIONS DE LA MONDIALISATION

Points saillants

Du commerce électronique à l'économie Internet (intelligent), la course à l'intelligence artificielle (IA) a toujours bénéficié du soutien et de l'intérêt des États à l'innovation. En effet, de par la nature transversale et englobante de l'Internet, le développement de l'IA n'échappe pas aux pressions de la mondialisation. Dans une perspective géo-politique, le présent chapitre donne tout d'abord un panorama sur le rôle des États ainsi que des instances régionales dans l'innovation, en portant ensuite, comme étude de cas, un regard critique sur la politique (locale) de l'intelligence artificielle au Québec.

Du fait de la nature transversale et englobante de l'Internet, le développement de l'intelligence artificielle (IA) n'échappe pas aux pressions de la mondialisation. La course à l'innovation n'est plus qu'une question économique liée à la globalisation des échanges ou l'intégration des marchés, mais également un enjeu géopolitique liée à l'affirmation / consolidation de grandes puissances (régionales). Dans un premier temps, il sera fait état de cet engouement renouvelé (3.1) que la société éprouve pour l'intelligence artificielle (IA) et plus particulièrement à l'égard des nouvelles percées réalisées par l'apprentissage profond [renvoi au Chapitre 2]. Cet engouement sociétal est redevable à des politiques publiques favorables à l'innovation qui en vient à cibler plus particulièrement la discipline de l'intelligence artificielle (3.2), comme l'étude de cas du Québec (3.3). Il convient toutefois de nuancer le modèle de l'intervention publique qui serait soutenu plus par une rhétorique des promesses technoscientifiques qu'une analyse raisonnée quant aux limites méthodologiques et une juste évaluation des coûts sociaux engendrés par le développement des nouvelles technologies (3.4).

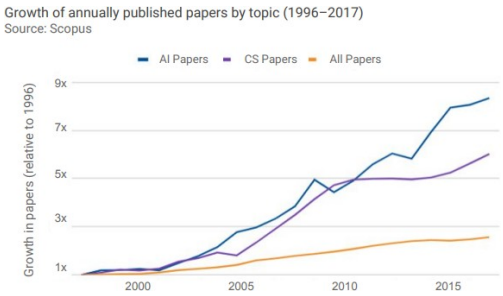
1. Un engouement partagé

Le tournant de l'année 2010 marque le renouveau d'un intérêt sociétal sans précédent pour l'intelligence artificielle (IA). Cet intérêt, qui n'est pas sans coïncider avec la consécration de l'apprentissage profond [renvoi au Chapitre 1], se manifeste tant dans la recherche (3.1.1) que l'enseignement supérieur (3.1.2) et le milieu de l'industrie (3.1.3).

1.1 Par la recherche

Le nombre de publications est l'un des indicateurs principaux de la performance de la recherche. Entre 1998 et 2018, le nombre de publications académiques portant sur l'IA a triplé et connaît une croissance qui se démarque, notablement depuis 2010, des autres domaines, y compris l'informatique (*computer science*) :

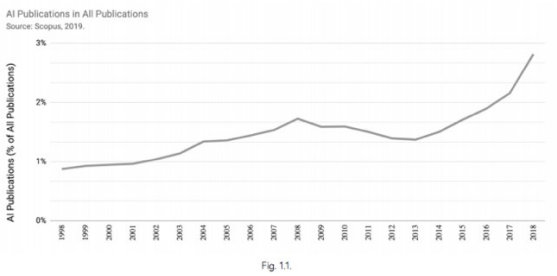
Figure 4
Hausse tendancielle du nombre de publications sur l'IA (1996-2017)



Source : (AI Index 2018 Annual Report[753] à la p 9)

En 2018, près de 3 % des articles publiés dans des revues révisées par les pairs de même que 9 % des articles de conférence ayant fait l'objet d'une publication, touchent au domaine de l'IA :

Figure 27
Publications sur l'IA parmi toutes les publications



Source : (AI Index 2019 Annual Report[754] à la p 14)

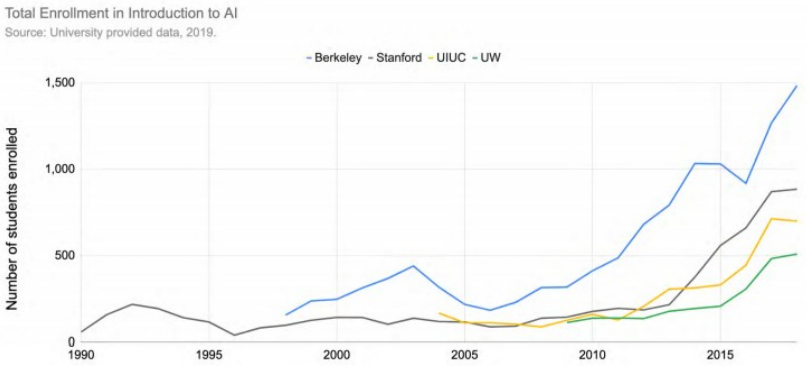
1.2 Par l'enseignement supérieur

L'intérêt que manifeste la relève pour l'IA est un autre axe de comparaison. Depuis 2010, le nombre d'étudiants inscrits à un cours de niveau universitaire touchant à l'intelligence artificielle (IA) ou à l'apprentissage automatique a connu une hausse importante. Cette tendance se constate tant au sein des grandes universités américaines qu'un peu partout dans le monde :

753 Yoav Shoham et al, *The AI Index 2018 Annual Report*, AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford (CA), décembre 2018, en ligne : <hai.stanford.edu/sites/default/files/2020-10/AI_Index_2018_Annual_Report.pdf>.

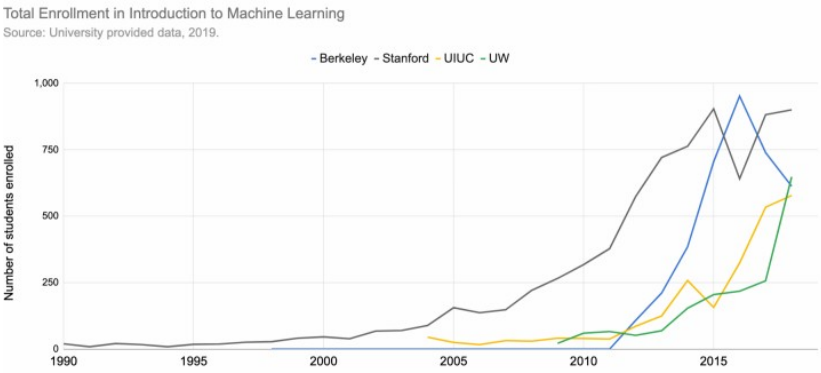
754 Raymond Perrault et al, *The AI Index 2019 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford (CA), décembre 2019, en ligne : <hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf>.

Figure 28
Nombre d'étudiants inscrits à un cours d'introduction à l'IA dans les quatre grandes universités américaines



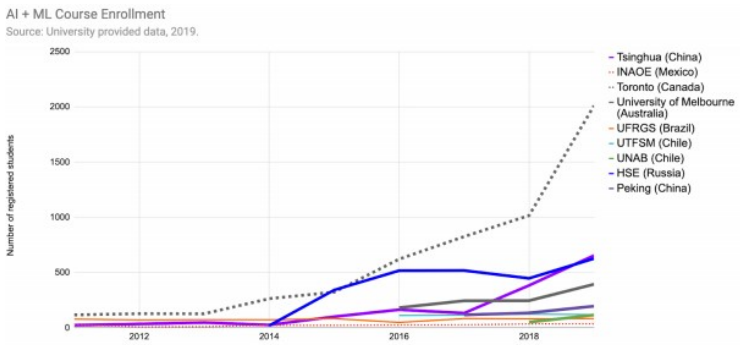
Source : (AI Index 2019 Annual Report à la p 111)

Figure 29
Le nombre d'étudiants inscrits à un cours d'introduction à l'apprentissage automatique dans les quatre universités américaines



Source : (AI Index 2019 Annual Report à la p 111)

Figure 30
Le nombre d'étudiants inscrits à un cours d'IA ou d'apprentissage automatique au sein des plus grandes universités internationales



Source : (AI Index 2019 Annual Report à la p 113)

À l'autre extrémité du parcours universitaire, l'intelligence artificielle et l'apprentissage automatique constituent par ailleurs les domaines de spécialisation les plus populaires chez les

nouveaux doctorants en (génie) informatique ou en information en Amérique du Nord, que ces derniers travaillent dans le milieu universitaire ou dans l'industrie. Encore une fois, c'est une tendance qui se dessine depuis 2009, année à partir de laquelle l'intelligence artificielle commence à se démarquer des autres sous-disciplines connexes comme l'interaction homme-machine, les systèmes d'information, les bases de données et les réseaux[755].

Pour aller plus loin :

Chassignol, M. et al., « Artificial Intelligence trends in education : a narrative overview » (2018) 136 Procedia Computer Science 16, doi : <doi.org/10.1016/j.procs.2018.08.233>

Guan, C., J. Mou et Z. Jiang, « Artificial intelligence innovation in education : A twenty-year data-driven historical analysis » (2020) 4:4 International Journal of Innovation Studies 134, doi : <doi.org/10.1016/j.ijis.2020.09.001>

Luan, H. et al., « Challenges and Future Directions of Big Data and Artificial Intelligence in Education » (2020) 11 Front Psychol, doi : <doi.org/10.3389/fpsyg.2020.580820>

Zawacki-Richter, O. et al., « Systematic review of research on artificial intelligence applications in higher education – where are the educators ? » (2019) 16 Int J Educ Technol High Educ, doi : <doi.org/10.1186/s41239-019-0171-0>

1.3 Par l'industrie

Quoique toutes les méthodes d'intelligence artificielle (IA) ne soient pas brevetables[756], l'impact de l'IA dans (les produits de) l'industrie peut s'apprécier, dans une certaine mesure, à l'aune de la proportion des brevets accordés par les gouvernements aux inventeurs pour bénéficier d'un monopole d'exploitation d'une durée limitée à l'égard de leurs nouvelles inventions.

Selon l'Organisation mondiale de la propriété intellectuelle[757], un total de 339 828 brevets sont liés à l'intelligence artificielle (IA), représentant près de 0,6 % de toutes les inventions brevetées depuis 1960[758]. L'apprentissage automatique (*machine learning*) s'avère de loin la famille de techniques dominante qui a été divulguée et incluse dans près d'un tiers de tous les brevets liés à l'IA. Elle est suivie de près par la programmation logique (*logic programming*) ou l'IA symbolique [renvoi au Chapitre 2] – dont font partie les systèmes experts – et la logique floue (*fuzzy logic*) :

---

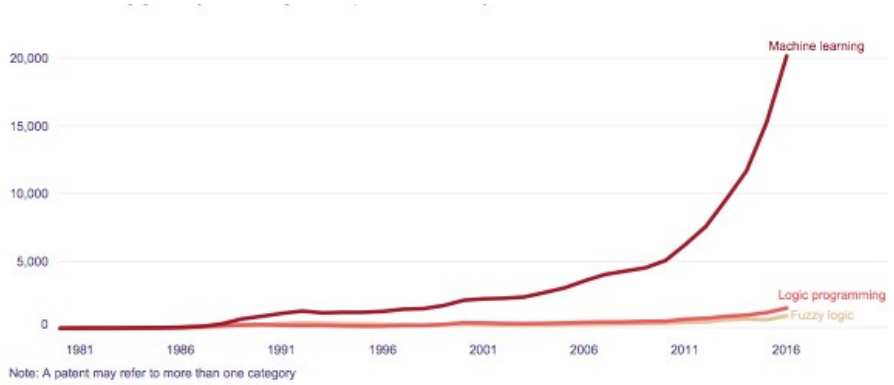
755 Computing Research Association (CRA), *The CRA Taulbee Survey*, 2001–2019, en ligne : <cra.org/resources/taulbee-survey/>.

756 Camille Aubin, « Intelligence artificielle et brevets », (2018) 30:3 *Les Cahiers de la propriété intellectuelle* 947, en ligne : <www.lescpi.ca/articles/v30/n3/intelligence-artificielle-et-brevets/>; Shuijing Hu et Tao Jiang, « Artificial Intelligence Technology Challenges Patent Laws », dans *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Changsha (Chine), IEEE, 2019, 241, DOI : <doi.org/10.1109/ICITBS.2019.00064>; Shlomit Yanisky-Ravid et Regina Jin, « Summoning a New Artificial Intelligence Patent Model : In The Age Of Pandemic » (2020), version préimprimée, DOI : <doi.org/10.2139/ssrn.3619069>.

757 Organisation mondiale de la Propriété intellectuelle (OMPI), *Artificial Intelligence. WIPO Technology Trends 2019*, Genève, OMPI, 2019, en ligne : <www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf>..

758 *Ibid.*, p 23

Figure 31
Évolution des brevets liés aux différentes techniques d'IA



Source : (OMPI, 2019 à la p 42)

Le nombre de demandes de brevets mentionnant l'une ou l'autre des techniques de l'apprentissage automatique a en effet connu une hausse annuelle moyenne de 26 % entre 2011 et 2016. L'une de ces techniques en particulier, l'apprentissage profond (*deep learning*), a connu une hausse annuelle moyenne de 175 % entre 2013 et 2016. La popularité croissante de l'apprentissage profond est suivie de loin par les tendances à la hausse qu'ont connu également, au cours de la même période, l'apprentissage multi-tâche (49 %) et les réseaux de neurones (46 %).

La vision par ordinateur [renvoi au Chapitre 2] (49 %), le traitement du langage naturel [renvoi au Chapitre 2] (14 %) et le traitement vocal [renvoi au Chapitre 2] (13 %) constituent les trois applications fonctionnelles ayant fait l'objet du nombre le plus élevé de brevets liés à l'IA. À l'intérieur de ces domaines d'application principaux, la biométrie (vision par ordinateur), l'extraction de l'information et la sémantique (traitement du langage naturel) ont connu une hausse substantielle depuis 2013, variant de 24 % à 33 %. C'est toutefois dans les applications liées à la robotique et les méthodes de contrôle que la hausse annuelle des demandes de brevets a été la plus élevée, avec une moyenne de 55 % entre 2013 et 2016[759].

Quant aux domaines d'application des brevets liés à l'IA, le transport et les télécommunications se démarquent notablement des autres depuis 2011. Dans le domaine du transport, l'essor est attribuable au développement des véhicules autonomes et aussi de l'aérospatiale / avionique, lesquels ont connu, en l'espace de trois ans (2013–2016), une hausse annuelle moyenne de 42 % et de 67 % respectivement[760].

L'analyse croisée des techniques, des applications fonctionnelles et des domaines d'application permettent par ailleurs de déceler certaines tendances conjointes. Par exemple, la vision par ordinateur mobilise très souvent les techniques de l'apprentissage profond (63,2 %), des machines à vecteurs de support (53,2 %) et de l'apprentissage non supervisé (47,9 %). Des

759 *Ibid.*, p. 47.

760 *Ibid.*, pp. 52-53.

corrélations significatives s'observent pareillement entre l'apprentissage des règles (*rule learning*) et la représentation des connaissances (67,5 %) ainsi, dans une moindre mesure, qu'entre l'apprentissage supervisé et le traitement du langage naturel (22,2 %). L'apprentissage automatique est aussi mentionné dans les brevets le plus souvent en lien avec les sciences médicales et de la vie ainsi que des télécommunications, alors que la vision par ordinateur est la plus fréquemment sollicitée dans les domaines des télécommunications et du transport[761].

Pour aller plus loin :

Gaudry, K. et T. Franklin, « Patent Trends Study Part One : Twelve-Industry Overview », IP Watchdog (1er mai 2019), en ligne : <www.ipwatchdog.com/2019/05/01/patent-trend-study-twelve-industry-overview/id=108739/>

OCDE et Union européenne, World Corporate Top R&D Investors : Shaping the Future of Technologies and of AI, Luxembourg, 2019, en ligne : <www.oecd.org/sti/world-corporate-top-rd-investors-shaping-future-of-technology-and-of-ai.pdf>

Office de la propriété intellectuelle du Canada, Traitement de l'intelligence artificielle : Aperçu du paysage canadien des brevets, Sa Majesté la Reine du chef du Canada, 2020, en ligne : <www.ic.gc.ca/eic/site/cipointernet-Internetopic.nsf/fra/h_wr04776.html>

United States Patent an Trademarok Office (USPTO), Inventing AI. Tracing the diffusion of artificial intelligence with U.S. patents, octobre 2020, en ligne : <www.uspto.gov/sites/default/files/documents/OCE-DH-AI.pdf>

2. Le rôle des États dans l'innovation

De ce qui précède, il est clair que l'intelligence artificielle (IA) a connu un essor massif ces dernières années. Ce regain d'intérêt partagé découle en partie d'un climat politique propice à l'innovation, « parrainé » par les États qui jouent un rôle névralgique dans l'élaboration et la mise en œuvre des politiques de soutien à l'innovation (technologique). En effet, la théorie économique, tout en soulignant « l'importance du progrès et donc de l'innovation pour soutenir la croissance à long terme de l'économie »[762], appuie la pertinence d'une intervention des autorités publiques pour corriger les défaillances du marché attribuables à des efforts d'innovation sous-optimaux lorsque confiés à la seule discrétion des entrepreneurs privés[763] : ces derniers hésiteront à s'y mettre pleinement lorsque les avancées obtenues en recherche et développement (R&D) bénéficieront également à leurs concurrents, alors qu'il revient aux seuls investisseurs d'assumer le risque inhérent à toute activité de recherche, sans compter l'ampleur

761 *Ibid.*, pp. 51-53.

762 Chantal Kegels, « La politique d'innovation dans une économie de la connaissance » (2009) 1-2 Reflets et perspectives de la vie économique 151, doi : <doi.org/10.3917/rpve.481.0151>.

763 Kenneth Arrow, « Economic Welfare and the Allocation of Resources for Invention », dans *Universities-Nationa Bureau Committee for Economic Research*, Committee on Economic Growth of the Social Science Research Council, *The Rate and Direction of Inventive Activity : Economic and Social Factors*, Princeton University Press, 1962, à la p 609, en ligne : <www.nber.org/system/files/chapters/c2144/c2144.pdf>.

de certains projets qui ne pourront être menés à terme que moyennant une collaboration inter-entreprise, multidisciplinaire et de longue haleine.

Aujourd'hui, l'expression « intelligence artificielle (IA) » a fait son chemin jusqu'aux plus hauts énoncés de politique publics. Nous passerons brièvement en revue quelques initiatives multilatérales et intergouvernementales caractéristiques des orientations communes en matière d'innovation, de l'Internet et de l'IA (3.2.1). Plus particulièrement, les stratégies nationales des États-Unis (3.2.2), du Canada (3.2.3), du Japon (3.2.4), de la Chine (3.2.5) et celles régionales de l'Union européenne (3.2.6) seront mises en exergue en tant qu'acteurs clés de cette course à l'innovation. Ensemble, ils témoignent de la prégnance de l'intelligence artificielle (IA) dans les discours publics (3.2.7).

2.1 Des initiatives multilatérales et intergouvernementales

Investie du mandat de contribuer au développement de l'économie mondiale[764], l'Organisation de coopération et de développement économiques (OCDE) a joué un rôle coordonnateur de premier plan dans les orientations publiques en matière d'Internet – pour un temps limité au commerce électronique – et de l'économie d'Internet jusqu'à l'Internet intelligent au sens large. D'autres initiatives multilatérales et intergouvernementales (p.ex. UNESCO, G7, G8, G20) emboîteront le pas pour faire de l'Internet intelligent non seulement un enjeu économique, mais également un outil de gouvernance géopolitique.

2.1.1 De l'Internet …

(1) Le commerce électronique

En 1998, un Plan d'action pour le commerce électronique a été adopté lors d'une conférence ministérielle de l'OCDE tenue à Ottawa (Canada). Reconnaissant que le commerce électronique est « par nature transfrontière », les ministres présents à la conférence conviennent du rôle clé que peut jouer l'OCDE à cet égard « en tant qu'enceinte privilégiée pour le dialogue entre gouvernements nationaux, organisations internationales et secteur privé ». Plus spécifiquement, le Plan d'action identifie à cet égard quatre domaines d'intervention prioritaires pour l'OCDE, soit :

- renforcer la confiance des utilisateurs et des consommateurs dans l'environnement numérique, en particulier à l'égard des mesures prises pour :
  - protéger la vie privée et les données de caractère personnel; et
  - assurer la sécurité des infrastructures et des technologies;

[764] *Convention relative à l'Organisation de Coopération et de Développement Économiques*, Paris, 14 décembre 1960, art. 1, en ligne : <www.oecd.org/fr/general/conventionrelativealorganisationdecooperationetdedeveloppementeconomiques.htm>.

- adopter les cadres juridiques et commerciaux aux réalités du marché numérique, notamment dans le domaine de la fiscalité;

- améliorer l'infrastructure de l'information pour le commerce électronique et son accès, avec entre autres l'élaboration de politiques de télécommunications adaptées;

- optimiser les avantages du commerce électronique et de son utilisation par les entreprises, les consommateurs et les institutions, notamment en analysant son impact économique et social.

Ce Plan d'action de l'OCDE ne fait pas explicitement référence au recours à l'intelligence artificielle (IA) ou d'agents intelligents pour faciliter le commerce électronique. Cela étant, l'OCDE a depuis toujours entretenu une vision large du commerce électronique, qui signifie « avoir une activité économique sur Internet, vendre des biens et des services qui sont livrés par les canaux traditionnels aussi bien que des produits pouvant être "numérisés" et diffusés en ligne, comme les logiciels informatiques ». Ainsi, dès 2000, l'OCDE a évoqué la possibilité que les consommateurs puissent, dans un avenir rapproché, bénéficier du recours aux « agents intelligents » qui « parcourent Internet et automatisent, par exemple, la recherche et les comparaisons de prix sur différents sites de commerce électronique »[765].

(2) L'économie Internet

Les promesses de l'Internet excéderont bientôt le déploiement d'un « simple » commerce électronique, mais s'affirmeront en tant qu'une composante indispensable qui s'est non seulement intégrée dans plusieurs aspects de notre quotidien, mais encore les sculpte à sa guise. À l'occasion de la réunion ministérielle de l'OCDE de 2008 (Séoul), les participants y évoquent les profondes mutations et d'importants défis relatifs à l'avenir de « l'économie Internet », qui couvre dorénavant « tout l'éventail de nos activités économiques, sociales et culturelles rendues possibles par l'Internet et par les technologies de l'information et des communications (TIC) »[766].

Encore une fois, un environnement réglementaire qui conforte « le caractère ouvert, décentralisé et dynamique de l'Internet »[767] et qui facilite la libre circulation de l'information ainsi que la recherche et le développement (R&D) s'avère essentiel pour contribuer au succès de l'économie Internet.

La Déclaration de Cancún, adoptée lors la Réunion ministérielle de 2016 de l'OCDE sur l'économie numérique, atteste de la volonté des pays signataires à stimuler l'innovation et la créativité numériques, à tirer parti du potentiel des plateformes numériques ainsi que des opportunités

765 *Ibid.*, p. 228.

766 OCDE, « Déclaration de Séoul sur le futur de l'économie internet », OECD Ministerial Meeting on the Future of the Internet Economy, Seoul, Korea, 17-18 juin 2008, p. 6, en ligne : https://www.oecd.org/fr/sti/40839567.pdf.

767 *Ibid.*, p. 7.

offertes par les applications et les technologies émergentes – dont l'analytique de données, et à réaliser les promesses de l'économie numérique sur les plans de l'emploi et de l'entrepreneuriat (commerce électronique). Ces priorités appellent à la nécessité de préserver « le caractère ouvert de l'Internet » et de favoriser la libre circulation de l'information, sans oublier la collecte de données probantes pour apprécier les effets de la transformation numérique sur la société, les défis socio-économiques qu'elle pose et les meilleures pratiques pour y faire face.

2.1.2 … à l'Internet intelligent

Lors de la réunion ministérielle du G7 qui s'est tenue à Montréal les 27 et 28 mars 2018, les ministres de l'Innovation du G7 ont jugé opportun d'adopter une Déclaration commune au sujet de l'intelligence artificielle[768]. Considérant les effets positifs attendus des innovations de l'IA dans tous les pays du G7 ainsi que la nécessité d'augmenter la confiance envers l'IA et d'en promouvoir le déploiement, cette Déclaration présente une « vision commune d'une IA centrée sur l'humain » dictant les engagements suivants :

- investir dans la recherche et développement (R&D) pour soutenir l'entrepreneuriat en IA et préparer la population active à l'automatisation;

- encourager la recherche et un dialogue multipartite – notamment entre les gouvernements, les acteurs du marché, la société civile et divers groupes communautaires – pour relever les défis sociétaux et aborder les enjeux éthiques liés à l'IA;

- promouvoir « des démarches neutres sur le plan technologique et appropriées sur les plans technique et éthique », dont la protection de la vie privée, l'investissement dans la cybersécurité et l'utilisation de technologies transformatrices pour protéger la vie privée et la transparence ainsi qu'améliorer la qualité et la sécurité des données;

- appuyer les efforts pour sensibiliser le public relatif aux avantages et potentiels de l'IA;

- soutenir la libre circulation de l'information grâce au recensement de pratiques exemplaires et d'études de cas sur les données gouvernementales ouvertes, ainsi que la coopération internationale en matière de partage et de protection de données;

- diffuser la Déclaration du G7 à l'échelle mondiale pour promouvoir la collaboration sur la scène internationale.

La Déclaration commune des ministres de l'Innovation du G7 considère qu'encourager la recherche et le développement (R&D) en matière d'intelligence artificielle s'inscrit en opposition « à des exigences de localisation des données qui sont injustifiables, en tenant compte d'objectifs légitimes de politiques publiques, ainsi qu'à des politiques d'application générale qui exigent

768 Déclaration des ministres de l'Innovation du G7 au sujet de l'intelligence artificielle, 28 mars 2018, en ligne : <http://www.g8.utoronto.ca/employment/2018-labour-annex-b-fr.html>.

l'accès au code source de logiciels de grande diffusion ou le transfert de code source comme condition d'accès au marché, tout en reconnaissant l'intérêt légitime des gouvernements à évaluer la sécurité de ces produits ».

Le 22 mai 2019, le Conseil de l'OCDE énonce cinq (5) principes complémentaires pour « une approche responsable en appui d'une IA digne de confiance »[769]. L'intelligence artificielle (IA) y est envisagée en tant qu'« une technologie générique qui promet d'améliorer le bien-être des individus, de contribuer à une activité économique mondiale dynamique et durable, de stimuler l'innovation et la productivité, et d'aider à affronter les grands défis planétaires ». Tirer parti du plein potentiel de cette promesse technologique « générique » s'accompagne néanmoins de défis importants au chapitre notamment des mutations économiques et sociales (p.ex. inégalités), du libre-échange (p.ex. concurrence), de l'adaptation des marchés du travail et des risques d'atteinte aux droits fondamentaux que pose l'utilisation croissante, si non balisée, des systèmes d'IA. Les parties prenantes sont ainsi invitées :

- à adopter une approche responsable de manière proactive pour optimiser les bénéfices de l'IA pour les individus et la collectivité (Principe 1 – « Croissance inclusive, développement durable et bien-être »);

- à instituer des mécanismes de garantie pour respecter les droits humains et les valeurs démocratiques tout au long du cycle de vie des systèmes d'IA (Principe 2 – « Valeurs centrées sur l'humain et équité »);

- à favoriser la transparence et une divulgation responsable des données liées aux systèmes d'IA (Principe 3 – « Transparence et explicabilité »);

- à assurer la sécurité et la robustesse des systèmes d'IA tout au long de leur cycle de vie ainsi que la traçabilité des données pour favoriser l'analyse des résultats produits (Principe 4 – « Robustesse, sûreté et sécurité »);

- à assurer le bon fonctionnement des systèmes d'IA et des principes énoncés ci-dessus (Principe 5 – « Responsabilité »).

Outre l'énoncé des cinq (5) principes, le Conseil de l'OCDE recommande à cette occasion cinq (5) domaines d'intervention à mettre en œuvre dans le cadre des politiques nationales et de la coopération internationale, soit :

- la recherche et le développement (interdisciplinaire) en matière d'IA, dont des investissements publics à long terme et des investissements privés pour stimuler l'innovation ainsi que relever les défis techniques, sociojuridiques et éthiques posés par l'IA;

---

769 OCDE, « Recommandation du Conseil sur l'intelligence artificielle », adoptée le 21 mai 2019, en ligne : <https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449>.

- l'instauration d'un « écosystème numérique pour l'IA », comprenant « notamment des technologies et infrastructures numériques et des mécanismes de partage des connaissances en matière d'IA, en fonction des besoins »;

- l'instauration d'un cadre d'action balisé au développement de l'IA, depuis le stade de recherche au déploiement des systèmes d'IA, ce qui implique notamment la mise en place d'un environnement contrôlé (*cf.* bac à sable) pour tester la performance des systèmes d'IA;

- la préparation des parties prenantes à la transformation du marché du travail et de la société, pour leur permettre d'interagir efficacement avec les systèmes d'IA et les doter des compétences nécessaires pour utiliser ces systèmes;

- la coopération active des pouvoirs publics avec les instances mondiales et régionales pour promouvoir une utilisation responsable des systèmes d'IA.

Les cinq (5) principes recommandés par le Conseil de l'OCDE ont été peu après repris par le G20 au sommet d'Osaka[770].

Le 11 mars 2020, l'UNESCO[771] a nommé un groupe de 24 experts internationaux pour rédiger une recommandation mondiale de haut niveau sur l'éthique de l'intelligence artificielle (IA). Il s'agit d'élaborer un « instrument normatif mondial visant à doter l'IA d'une base éthique solide »[772]. L'initiative fait suite à une étude préliminaire[773] ayant conclu à une grande hétérogénéité tant dans les énoncés de principe que la mise en œuvre des différentes déclarations et instruments normatifs existants. Outre la (ré)affirmation des principes fondamentaux liés à l'utilisation de l'IA, la recommandation devrait également comprendre des propositions spécifiques en vue d'aider les États à la mettre en œuvre et à réglementer leurs usages de l'IA dans les domaines relevant du mandat de l'UNESCO, y compris l'éducation, la diversité culturelle et l'inclusivité, l'égalité des genres, l'accès universel à l'information, le maintien de la paix, l'introduction responsable de l'IA dans la pratique scientifique ainsi que la prévention et la gestion des crises environnementales. Il est attendu de cette recommandation qu'elle réunisse « à la fois les pays développés et les pays en développement, [concilie] différents points de vue culturels et moraux ainsi qu'un éventail varié de parties prenantes des sphères publiques et privées au sein d'un processus réellement

---

770 Gouvernement du Canada, *Déclaration des dirigeants du G20*, 28 et 29 juin 2019, Osaka, Japon, en ligne : <https://www.international.gc.ca/world-monde/international_relations-relations_internationales/g20/2019-06-29-g20_leaders-dirigeants_g20.aspx?lang=fra>.

771 UNESCO, *Composition du groupe d'experts ad hoc (GEAH) pour la recommandation sur l'éthique de l'intelligence artificielle*, SHS/BIO/AHEG-AI/2020/INF.1 REV, Paris, 11 mars 2020, en ligne : <unesdoc.unesco.org/ark:/48223/pf0000372991>.

772 UNESCO, Élaboration d'une Recommandation sur l'éthique de l'intelligence artificielle, en ligne : <fr.unesco.org/artificial-intelligence/ethics>.

773 UNESCO, Commission mondiale d'éthique des connaissances scientifiques et des technologies (COMEST), *Étude préliminaire sur l'éthique de l'intelligence artificielle*, SHS/COMEST/EXTWG-ETHICS-AI/2019/1, Paris, 26 février 2019, en ligne : <unesdoc.unesco.org/ark:/48223/pf0000367823_fre>.

international d'élaboration d'un ensemble complet de principes et de propositions concernant l'éthique de l'IA »[774].

Soulignons enfin le Partenariat mondial sur l'intelligence artificielle (PMIA), lancé le 15 juin 2020, représentant sans doute l'initiative multipartite la plus aboutie pour le partage de la recherche multidisciplinaire sur l'IA et la coopération internationale. Le PMIA rassemble tant les acteurs de l'industrie et du monde universitaire que de la société civile, des gouvernements et des organisations internationales, pour « favoriser un développement responsable de l'IA fondé sur ces principes de droits de l'homme, d'inclusion, de diversité, d'innovation et de croissance économique »[775]. Les États-Unis, la France, l'Allemagne, l'Union européenne, le Canada et le Japon comptent parmi ses membres fondateurs. Un secrétariat hébergé par l'OCDE à Paris sera mis à la disposition du PMIA, ainsi que deux centres d'expertise sis respectivement à Montréal et à Paris[776].

Ces initiatives intergouvernementales, pilotées notamment par l'OCDE, tissent une toile de fond propice à l'innovation sur laquelle différents États ont laissé leur marque.

2.2 Les États-Unis

De l'Arpanet à l'Internet, les États-Unis ont été les pionniers de l'innovation et les bailleurs de fonds principaux, à l'échelle mondiale, des infrastructures technologiques tissant un cyberespace sans frontières[777]. Le rapport[778] à l'attention du Président a tracé la voie à une politique nationale scientifique pour le bien-être de la nation. Le rapport Bush (1945) avait identifié certains domaines stratégiques qui nécessitent un financement public, dont la recherche militaire, l'agriculture, le logement, la santé publique, la recherche médicale ainsi que des projets de recherche d'envergure nécessitant un investissement en capital au-delà de la capacité des bailleurs privés. En particulier, le rapport Bush met l'emphase sur la nécessité d'un financement

774 *Ibid.,* à la p 27.

775 Le Partenariat mondial sur l'intelligence artificielle (PMIA), *À propos*, en ligne : <gpai.ai/fr/a-propos/>.

776 OCDE, *L'OCDE hébergera le Secrétariat du nouveau Partenariat mondial sur l'intelligence artificielle*, 15 juin 2020, en ligne : <www.oecd.org/fr/presse/l-ocde-hebergera-le-secretariat-du-nouveau-partenariat-mondial-sur-l-intelligence-artificielle.htm>.

777 Laurent Bloch, *L'Internet, vecteur de puissance des États-Unis?,* Diploweb, 2017, en ligne : <www.diploweb.com/-L-Internet-vecteur-de-puissance-des-Etats-Unis-.html>; National Research Council, *Funding a Revolution. Government Support for Computing Research*, Washington, DC, National Academies Press, 1999, doi : <doi.org/10.17226/6323>.

778 Vannevar Bush, *Science : The Endless Frontier*, Washington (DC), United States Government Printing Office, 1945, en ligne : <www.nsf.gov/od/lpa/nsf50/vbush1945.htm#ch1.3> : "We have no national policy for science. The Government has only begun to utilize science in the nation's welfare. There is no body within the Government charged with formulating or executing a national science policy. There are no standing committees of the Congress devoted to this important subject. Science has been in the wings. It should be brought to the center of the stage – for in it lies much of our hope for the future."

public soutenu à la recherche fondamentale, qui doit par nature viser un horizon à long terme de cinq ans et plus; elle cesserait d'être fondamentale si des résultats immédiats étaient attendus[779].

Dans la foulée du rapport *Science : The Endless Frontier* (1945), l'histoire de l'Internet est étroitement entrelacée avec celle du *Defense Advanced Research Projects Agency* (DARPA), du département de la Défense américaine. Le DARPA a piloté la révolution de l'Internet[780], y compris un investissement massif dans le développement de l'intelligence artificielle (IA), tant symbolique (décennie 1960) que connexionniste (décennie 1990) depuis plus de cinquante ans. Le 7 septembre 2018, à l'occasion du Symposium D60 soulignant le 60e anniversaire du DARPA[781] annonce un investissement pluriannuel de plus de 2 milliards de dollars américains pour repousser les frontières de la connaissance dans le développement des technologies de l'IA d'apprentissage profond dites de troisième vague, qui soient capables de raisonnements contextuels pour notamment s'adapter de façon autonome aux situations changeantes. Dans le cade de ce programme[782], les principaux domaines d'intérêt comprennent :

- l'automatisation des processus métier du département de la Défense (p.ex. vérification des autorisations de sécurité, accréditation des systèmes logiciels aux fins de déploiement opérationnel);
- l'amélioration de la robustesse et de la fiabilité des systèmes d'IA;
- le renforcement de la sécurité et de la résilience des technologies d'apprentissage automatique et de l'IA;
- la réduction des inefficiences computationnelles, de traitement de données et de performance;
- le développement d'algorithmes et d'applications intelligentes de « nouvelle génération » dotées de sens commun et capables d'expliquer leur propre fonctionnement.

Cette plus récente initiative du Pentagone s'inscrit dans le cadre du Plan stratégique pour le recherche et développement en intelligence artificielle (2016)[783] rendu public en octobre 2016

---

779 *Ibid. :* "Basic research is a long-term process – it ceases to be basic if immediate results are expected on short-term support. Methods should therefore be found which will permit the agency to make commitments of funds from current appropriations for programs of five years duration or longer. […]."

780 Mitch Waldrop, *DARPA and the Internet Revolution*, DARPA, 2015, en ligne : <www.darpa.mil/attachments/(2O15)%20Global%20Nav%20-%20About%20Us%20-%20History%20-%20Resources%20-%2050th%20-%20Internet%20(Approved).pdf>.

781 Defense Advanced Research Projects Agency (DARPA), *DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies*, 7 septembre 2018, en ligne : <www.darpa.mil/news-events/2018-09-07>.

782 Defense Advanced Research Projects Agency (DARPA), *AI Next Campaign*, en ligne : <www.darpa.mil/work-with-us/ai-next-campaign>.

783 National Science and Technology Council (Networking and Information Technology Research and Development Subcommittee), *The National Artificial Intelligence Research and Development Strategic Plan*, Executive Office of the President of the United States, octobre 2016, en ligne : <www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf>.

par le National Science and Technology Council (NSTC). Au nom du gouvernement fédéral, ce plan propose les sept (7) priorités de financement pour la recherche en intelligence artificielle (IA), soit :

1. ***Prioriser les investissements à long terme (de 5 à 10 ans, voire plus) en intelligence artificielle pour permettre aux États-Unis de rester un chef de file mondial en intelligence artificielle –***

   À cet égard, le Plan stratégique fait référence à l'évolution lente du World Wide Web à l'apprentissage profond au terme de plus de trente (30) ans de développement assidu mais avec des gains significatifs. Dans cette perspective, un financement à long terme permet d'éponger la longue période d'incubation requise pour la maturation des technologies intelligentes, notamment dans les domaines suivants liés aux technologies de nouvelle génération :

   - les analyses de données avancées pour l'extraction de connaissances nouvelles à partir des *Big Data*;
   - les avancées en matière de capacités de perception des systèmes d'IA;
   - une compréhension approfondie du potentiel et des limitations théoriques de l'IA;
   - la recherche continue en intelligence artificielle générale ou forte par opposition à l'intelligence artificielle étroite ou faible [renvoi au Chapitre 1];
   - le développement des multi-systèmes d'IA évolutifs;
   - la recherche en IA explicable, notamment dans le cadre de tutorat ou d'assistance intelligent;
   - le développement de la robotique, notamment pour affiner l'extraction de l'information par les robots, améliorer leur interaction avec les humains et développer une conscience situationnelle en temps réel (*real-time situational awareness*);
   - la recherche et le développement des infrastructures au soutien des technologies d'IA avancées.

2. ***Développer des méthodes de collaboration / d'interaction efficaces entre l'humain et les systèmes d'IA –***

   Il s'agit d'optimiser, dans les domaines qui le justifient, une division des tâches fonctionnelle entre l'humain et les systèmes d'IA. Selon le NSTC, cette division fonctionnelle des tâches s'articule autour des trois (3) modalités suivantes :

   - l'IA au soutien d'un décideur humain (p.ex. analyses prédictives);
   - l'IA intervenant à la demande de l'humain (p.ex. systèmes d'alarme, d'avertissement ou de diagnostic);

- l'IA se substituant à l'humain dans la réalisation de certaines tâches qui s'y prêtent, comme celles à haute technicité, requérant une réactivité immédiate ou menées dans des environnements dangereux / toxiques.

3. ***Étudier les implications sociales, juridiques et éthiques liées à l'utilisation des technologies de l'IA et le design des systèmes d'IA qui s'alignent avec les objectifs sociétaux –***

4. ***Assurer la sécurité et la fiabilité des systèmes d'IA –***

5. ***Développer des ensembles de données publiques partagées pour tester et entraîner les outils intelligents –***

6. ***Évaluer les progrès d'IA à l'aide de méthodes balisées et de normes reconnues notamment en matière de performance, de sécurité, d'aptitude à l'utilisation, d'interopérabilité et de traçabilité –***

7. ***Mieux comprendre les demandes en effectifs actuels et futurs en R&D de l'IA pour garantir la disponibilité d'un bassin d'experts capables de travailler sur les autres priorités présentées plus haut –.***

2.3 Le Canada

À l'instar de l'exemple américain [renvoi à la sous-section 3.2.2], le gouvernement canadien avait voulu développer son propre réseau de transfert et de partage de données, sans grand succès[784]. Le projet Télidon, développé au courant des années 1970 par le ministère fédéral des Communications, voit son financement coupé court au printemps 1983 en raison d'un déploiement limité et d'un manque d'intérêt des usagers[785]. La décennie 1990, tout en marquant la naissance de l'Internet moderne, coïncide avec le deuxième hiver de l'IA [renvoi au Chapitre 1] dont il a fallu attendre jusqu'à la consécration de l'apprentissage profond pour marquer un renouveau, tant du courant connexionniste qu'un regain d'intérêt des subventionnaires et du gouvernement.

Une *Stratégie pancanadienne en matière d'intelligence artificielle pour la recherche et le talent* (ci-après « Stratégie pancanadienne » ou « Stratégie ») est ainsi annoncée dans le budget fédéral de 2017. Les « industries numériques » y sont présentées comme l'un des six principaux

---

784 Nick Moreau, « Internet in Canada », *The Canadian Encyclopedia* (3 décembre 2012), en ligne : <www.thecanadianencyclopedia.ca/en/article/internet>.

785 Donald J. Gillies, « Technological Determinism in Canadian Telecommunications: Telidon Technology, Industry and Government » (1990) 15:2 Canadian Journal of Communication 1, en ligne : <digital.library.ryerson.ca/islandora/object/RULA%3A4805>; William Richards, Sasha Yusufali et Roy Marsh, « Télidon », *L'Encyclopédie canadienne* (28 janvier 2007), en ligne : <www.thecanadianencyclopedia.ca/fr/article/telidon>.

domaines à forte croissance de l'économie[786] et comme offrant un grand potentiel de création d'emploi. L'intelligence artificielle, plus particulièrement, se démarque par une application polyvalente dans de nombreux secteurs de l'économie, profitant aux entreprises de toutes tailles et générant une croissance économique solide. Sur cette toile de fond, le gouvernement fédéral propose d'affecter une enveloppe de 125 millions de dollars au lancement d'une Stratégie pancanadienne « [p]our maintenir en poste et attirer la crème du talent universitaire, et pour accroître le nombre de stagiaires et de chercheurs de deuxième cycle qui étudient l'intelligence artificielle et l'apprentissage profond » (p. 117). La Stratégie devrait favoriser la collaboration, d'un océan à l'autre, des principaux centres canadiens d'expertise, et positionner le Canada « en tant que destination de calibre mondial pour les entreprises désirant investir dans l'intelligence artificielle et l'innovation » (p. 117). L'Institut canadien de recherches avancées (ICRA), reconnu comme « un chef de file dans le domaine de l'intelligence artificielle » (p. 117), a été investi de la mission d'administrer le financement de cette Stratégie.

Aux dernières nouvelles, le gouvernement du Québec devrait annoncer prochainement une « Stratégie d'adoption de l'intelligence artificielle » pour soutenir son utilisation et baliser son usage par les organisations publiques[787].

2.4 Le Japon

Le 31 mars 2017, le Conseil stratégique de la technologie de l'intelligence artificielle (IA) a rendu public sa stratégie en matière d'IA[788]. Considérant l'immense potentiel que présente cette dernière pour la société, un plan de développement en trois phases y est prévu, passant (1) de l'utilisation et de l'application de l'IA axée sur les données, à (2) l'utilisation publique de l'IA et des données, et enfin à (3) la création d'écosystèmes de connexions multidomaines. Pour ce faire, le partenariat université-industrie est encouragé, de même que le développement de systèmes intelligents par des start-ups et leur appariement avec les grandes entreprises et institutions financières.

---

786 À côté de la fabrication de pointe, de l'agroalimentaire, des technologies propres, des sciences biologiques et de la santé ainsi que des ressources propres.

787 Secrétariat du Conseil du Trésor, *Stratégie de transformation numérique gouvernementale 2019-2023*, gouvernement du Québec, 2019, à la p 23, en ligne : <cdn-contenu.quebec.ca/cdn-contenu/adm/min/secretariat-du-conseil-du-tresor/publications-adm/strategie/StrategieTNG.pdf?1559512998>.

788 Strategic Council for AI Technology, *Artificial Intelligence Technology Strategy*, gouvernement du Japon, 31 mars 2017, en ligne : <ai-japan.s3-ap-northeast-1.amazonaws.com/7116/0377/5269/Artificial_Intelligence_Technology_StrategyMarch2017.pdf>.

2.5 La Chine

Le 8 juillet 2017, le Conseil d'État de Chine rend public son *Plan de développement de l'intelligence artificielle de nouvelle génération*[789]. Ce plan prévoit trois séries d'objectifs stratégiques échelonnés sur un horizon décennal, dont une première phase de mise à niveau et de développement d'une industrie excédant 150 milliards de yuans d'ici 2020, suivie de percées et d'avancées majeures à être réalisées d'ici 2025, ainsi que de la maturation de systèmes intelligents pour 2030, lesquels permettront de positionner la Chine comme le leader mondial de l'IA et le principal centre d'innovation.

Le plan de développement promeut une vision à long terme mettant l'emphase sur le rôle de l'État tant dans les subventions directes à la recherche et au développement (par le biais d'incitatifs fiscaux aux entreprises), que l'implémentation d'infrastructures de recherche intelligentes pour soutenir le développement industriel, le traitement de données massives et le calcul intensif (*supercomputing*). Le capital social est aussi mobilisé pour financer le développement de l'IA, par le biais notamment d'investisseurs providentiels (*angel investment*) et le financement de marché. L'État se reconnaît également un rôle de coordonnateur supervisant l'avancement des différents projets majeurs en matière d'IA et de technologies. Il convient également d'encourager la recherche pluridisciplinaire et la collaboration avec les entreprises et les centres de recherche à l'étranger.

La mise en œuvre du plan de développement est confiée à un comité spécial du ministère de la Science et de la Technologie, qui a été constitué expressément à cette fin[790].

2.6 L'Union européenne

Si les politiques de l'innovation revêtent une importance amplement reconnue tant pour les Communautés européennes que l'Union européenne (UE), ce n'est que tout récemment que les programmes d'action font explicitement référence à l'intelligence artificielle (IA). La Stratégie européenne pour l'IA[791], présentée par la Commission européenne le 25 avril 2018, propose une vision évolutive de l'intelligence artificielle (IA) qui, au-delà du commerce électronique et de

---

789 Conseil d'État de Chine, 新一代人工智能发展规划 *[Plan de développement de l'intelligence artificielle de nouvelle génération]*, 8 juillet 2017, en ligne : <www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm>; traduit en anglais par Graham Webster, Rogier Creemers, Paul Triolo et Elsa Kania, *Full Translation : China's 'New Generation Artificial Intelligence Development Plan'*, 1er août 2017, en ligne : <www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>.

790 Pour une analyse critique de la politique nationale chinoise en matière d'IA, voir Huw Roberts et al, « The Chinese Approach to Artificial Intelligence : An Analysis of Policy, Ethics, and Regulation » (2021) 36 AI & Society 59, doi : <doi.org/10.1007/s00146-020-00992-2>.

791 Commission européenne, *L'intelligence artificielle pour l'Europe*, communication de la Commission au Parlement européen, au Conseil européen, au Conseil, au Comité économique et social européen et au Comité des régions, COM(2018) 237 final, Bruxelles, 25 avril 2018, en ligne : <ec.europa.eu/transparency/regdoc/rep/1/2018/FR/COM-2018-237-F1-FR-MAIN-PART-1.PDF>.

l'économie Internet, s'avère prometteuse pour participer à la résolution d'enjeux mondiaux (émergents) comme la prise en charge des maladies chroniques, la lutte contre le changement climatique, la cybersécurité et la sécurité routière. D'où l'importance pour l'Union européenne (UE) d'adopter une approche coordonnée sur le développement et l'utilisation de l'IA. La Stratégie définit l'intelligence artificielle de façon large comme embrassant « les systèmes qui font preuve d'un comportement intelligent en analysant leur environnement et en prenant des mesures – avec un certain degré d'autonomie – pour atteindre des objectifs spécifiques » (Stratégie européenne pour l'IA, 2018, p. 1). Soulignant la nécessité pour l'UE de bénéficier d'un soutien fort dans ce domaine, la stratégie présente les grandes lignes d'une initiative européenne sur l'IA, axée sur les objectifs suivants :

- Encourager le recours à l'IA dans tous les secteurs de l'économie, tant le secteur public que le secteur privé, y compris un soutien aux petites et moyennes entreprises représentant 99 % des entreprises de l'UE –

  Alors que la Commission européenne estime à 4-5 milliards d'euros le montant total des investissements – tant publics et privés – effectués pour la R&D de l'IA dans l'Union européenne en 2017, il est prévu que ce montant sera porté à au moins 20 milliards d'euros d'ici la fin de 2020 et même excédera les 20 milliards d'euros par an au cours de la prochaine décennie.

  Tant la recherche fondamentale que la recherche industrielle sont encouragées, du laboratoire jusqu'au marché. Certains des domaines d'application clés identifiés par la Commission européenne comprennent la santé, les véhicules intelligents, l'agriculture, les technologies de nouvelle génération, la sécurité et les administrations publiques, dont la justice.

- Encourager le recours à l'IA dans tous les secteurs de l'économie, y compris les entreprises des secteurs *a priori* non technologiques et les administrations publiques, implique la mise en place d'infrastructures d'essais et d'expérimentation ainsi qu'un accès facilité aux ressources utiles à l'IA.

  Outre les investissements publics, le Fonds européen pour les investissements stratégiques sera mis en place pour mobiliser les investisseurs privés.

- Enfin, promouvoir la R&D dans le domaine de l'IA, notamment l'apprentissage automatique, requiert une quantité importante de données dont l'accès à des fins de partage de connaissances, de réutilisation ainsi que d'entraînement des systèmes d'IA, peut constituer un avantage compétitif. C'est la raison pour laquelle l'UE travaille non seulement pour l'ouverture des données détenues par le secteur public, mais encourage également une plus grande disponibilité des données non personnelles détenues par le secteur privé (p.ex. données industrielles).

- Soutenir la transition numérique sur le marché de l'emploi, comprenant :

- la modernisation de l'enseignement et de la formation (continue) pour doter les travailleurs des compétences technologiques requises pour faire face aux mutations dans le monde du travail;
- la création de nouveaux profils d'emplois spécialisés dans les technologies numériques; et
- l'acquisition de compétences ne pouvant pas être remplacées par l'automatisation, comme la pensée critique et la créativité.

- Baliser le développement de l'IA à l'aide d'un cadre éthique et juridique fondé sur les valeurs de l'UE et les droits fondamentaux, dont :
  - la protection des données à caractère personnel;
  - la circulation des données à caractère non personnel; et
  - l'élaboration des lignes directrices en matière d'éthique de l'IA.

Le 19 février 2020, la Commission européenne a adopté un Livre blanc sur l'intelligence artificielle[792] prônant encore une fois une approche axée sur la régulation et l'investissement. Y est mise en exergue la nécessité de coopérer avec les États membres pour favoriser le développement et l'utilisation de l'IA en Europe, de déployer des efforts concertés pour la recherche et l'innovation, de mettre à niveau les compétences des régulateurs sectoriels pour comprendre et travailler avec les systèmes d'IA, ainsi que d'encourager le recours à l'IA par le secteur public et les petites et moyennes entreprises (PME). Le rôle des investissements privés ou des partenariats publics-privés y est de nouveau souligné.

En date du 25 février 2020, la Commission européenne[793] a pu recenser l'existence d'une stratégie en matière d'intelligence artificielle (IA) dans la plupart des États membres de l'Union européenne (UE). Les stratégies nationales sont à différents stades de leur développement, et plus de la moitié des États membres (16) ont déjà publié la leur.

792 Commission européenne, *LIVRE BLANC : Intelligence artificielle. Une approche européenne axée sur l'excellence et la confiance*, COM(2020) 65 final, Bruxelles, 19 février 2020, en ligne : <www.eesc.europa.eu/fr/our-work/opinions-information-reports/opinions/livre-blanc-sur-lintelligence-artificielle>.

793 Commission européenne, *AI Watch. National Strategies on Artificial Intelligence. An European perspective in 2019*, JRC Technical Report, Union européenne, 2020, en ligne : <publications.jrc.ec.europa.eu/repository/bitstream/JRC119974/national_strategies_on_artificial_intelligence_final_1.pdf>.

Figure 32
État des stratégies nationales en matière d'IA au sein de l'UE (2019)

Country	Status	Date
Austria	Final draft	June 2019
Belgium	In progress	
Bulgaria	In progress	
Croatia	Final draft	Nov. 2019
Cyprus	Published	Jan. 2020
Czech Republic	Published	May 2019
Denmark	Published	March 2019
Estonia	Published	July 2019
Finland	Published	Oct. 2017
France	Published	March 2018
Germany	Published	Nov. 2018
Greece	In progress	
Hungary	Action plan	Nov. 2019
Ireland	In progress	
Italy	Final draft	July 2019
Latvia	Published	Febr. 2020
Lithuania	Published	April 2019
Luxembourg	Published	May 2019
Malta	Published	Oct. 2019
Netherlands	Published	Oct. 2019
Poland	Final draft	Aug. 2019
Portugal	Published	June 2019
Romania	In progress	
Slovakia	Published	Oct. 2019
Slovenia	In progress	
Spain	Final draft	Nov. 2019
Sweden	Published	May 2019
United Kingdom	Published	April 2018

Source : Commission européenne (2020) à la p 6

Les différentes stratégies nationales s'articulent autour des pôles suivants :

- le développement du capital humain, y compris la formation (continue) de la main-d'œuvre en matière d'IA, l'intégration de la technologie en milieu de travail et l'adaptation du marché du travail aux changements technologiques induits par l'IA;

- la recherche et développement, encourageant les initiatives en IA dans le secteur privé ainsi qu'une meilleure efficience dans la prestation des services publics;

- la mise en réseau, l'attraction et la rétention des talents en IA dans les secteurs public et privé;

- le développement d'un cadre réglementaire et éthique pour l'utilisation de l'IA;

- la mise en place d'une infrastructure numérique et de télécommunications pour permettre et faciliter la cueillette, l'analyse et le traitement des données massives.

***

Avec le lancement de l'Observatoire des politiques de l'IA de l'OCDE[794], il est dorénavant facile de suivre (en temps réel) l'évolution des politiques de l'IA dans différentes juridictions. Le Conseil de l'Europe[795] tient également à jour les initiatives en matière d'IA émanant de différentes

794 OECD.AI Policy Observatory, *National Ai Policies & Strategies*, en ligne : <oecd.ai/dashboards>.

795 Conseil de l'Europe, *Initiatives sur l'IA*, en ligne : <www.coe.int/fr/web/artificial-intelligence/national-initiatives>.

juridictions, *think tank* et institutions à travers le monde, qu'il s'agisse de projets, d'énoncés de position, de plans d'action ou de développement. Ces mises à jour se font aussi – il vaut la peine de le souligner – avec le concours de l'IA.

Pour aller plus loin :

Atkinson, R.D., « Understanding the U.S. National Innovation System, 2020 », Information Technology & Innovation Foundation (ITIF), 2020, en ligne : <itif.org/publications/2020/11/02/understanding-us-national-innovation-system-2020>

Block, F.L. et M.R. Keller, *State of Innovation. The U.S. Government's Role in Technology Development*, Routledge, 2011, en ligne : <www.routledge.com/State-of-Innovation-The-US-Governments-Role-in-Technology-Development/Block-Keller/p/book/9781594518249>

Breznitz, D., *Innovation and the State : Political Choice and Strategies for Growth in Insrael, Taiwan, and Ireland*, Yale University Press, 2007, en ligne : <www.jstor.org/stable/j.ctt1nppt9>

Groth, O.J. et al., *Comparison of National Strategies to Promote Artificial Inteligence. Part 1*, Berlin, Konrad-Adenauer-Stiftung, 2019, en ligne : <www.kas.de/documents/252038/4521287/Comparison+of+National+Strategies+to+Promote+Artificial+Intelligence+Part+1.pdf/397fb700-0c6f-88b6-46be-2d50d7942b83?version=1.1&t=1560500570070>

International Institute of Communications (IIC) et TRPC, *Artificial Intelligence in the Asia-Pacific Region. Examining policies and strategies to maximise AI readiness and adoption*, février 2020, en ligne : <www.iicom.org/wp-content/uploads/IIC-AI-Report-2020.pdf>

Meltzer, J.P. et C.F. Kerry, *Strengthening international cooperation on artificial intelligence*, Brookings Institute, 17 février 2021, en ligne : <www.brookings.edu/research/strengthening-international-cooperation-on-artificial-intelligence/>

National Research Council, *Best Practices in State and Regional Innovation Initiatives : Competing in the 21st Century*, Washington, DC, National Academies Press, 2013, doi : <doi.org/10.17226/18364>

National Research Council, *Rising to the Challenge : U.S. Innovation Policy for the Global Economy*, Washington, DC, National Academies Press, 2012, doi : <doi.org/10.17226/13386>

Papaioannou, T., « Reflections on the entrepreneurial state, innovation and social justice » (2020) 1 Review of Evolutionary Political Economy 199, doi : <doi.org/10.1007/s43253-020-00018-z>

3. La prégnance de l'intelligence artificielle (IA) dans les discours publics

Certes, il n'y a pas que les stratégies portant explicitement sur « l'intelligence artificielle » ou « l'apprentissage profond » qui influenceront le développement et le déploiement de l'IA. Les plans d'action, stratégies et énoncés de position sur « la croissance ou l'économie numérique », « les technologies numériques », « l'innovation », sont tout aussi susceptibles d'orienter (in)directement le développement des technologies intelligentes. C'est sans compter les (autres) stratégies, plans d'action ou de développement ainsi que divers énoncés de position à l'initiative des subdivisions / collectivités territoriales, communautés autonomes, voire municipalités d'un même pays. Il n'empêche que l'emploi explicite de l'expression « intelligence artificielle » dans

les différentes stratégies nationales et régionales marque l'ancrage de cette discipline dans les discours publics.

Au reste, les stratégies nationales et intergouvernementales tournent pour l'essentiel autour des mêmes thèmes que sont le développement du capital humain ainsi que le soutien à l'industrie, à la recherche et au partenariat université-industrie. Au-delà des énoncés de principe et du cadre institutionnel général assurant le développement du libre marché (p.ex. droit de propriété, brevet, droit de la concurrence), les États peuvent encourager l'innovation technologique par le biais d'une panoplie d'instruments, comprenant :

- les soutiens fiscaux à l'innovation;
- les subventions directes à la recherche, à la formation d'une main-d'œuvre qualifiée et à l'emploi;
- l'implantation d'infrastructures technologiques de recherche (para-)publique;
- l'attraction et la rétention de talents hautement qualifiés sur le territoire national;
- le contrôle des investissements étrangers dans des domaines stratégiques.

Nous les détaillerons plus bas dans notre étude de cas sur la politique de l'IA au Québec.

3.1 Perspective locale : la politique de l'intelligence artificielle (IA) au Québec

Depuis les années 1960, un changement radical s'est opéré au sein des États dans leur relation avec la science et la technologie. De la recherche scientifique jusqu'alors promue en tant qu'une fin en soi, celle-ci s'arrime de plus en plus avec les impératifs du marché jusqu'à former un nœud devenu inextricable. Dans un rapport publié en 1963, l'OCDE (1963)[796] recommande aux gouvernements d'adopter une approche plus interventionniste en matière de « politique scientifique ». Cette recommandation a inspiré, au courant des années 1960 et 1970, diverses stratégies nationales que les pays industrialisés ont adoptées, chacun à leur rythme, dans le but d'instrumentaliser la recherche scientifique à des fins socio-économiques. Le Canada[797] et le Québec ne font pas exception.

---

796 Organisation de coopération et de développement économique (OCDE), La Science et la politique des gouvernements. L'influence de la science et de la technique sur la politique nationale et internationale, 1963; Voir aussi Muriel Le Roux et Girolamo Ramunni, « L'OCDE et les politiques scientifiques » (2000) 3 Revue pour l'histoire du CNRS, doi : <doi.org/10.4000/histoire-cnrs.2952>.

797 Nous n'aborderons pas en détail l'évolution des politiques scientifiques fédérales dans le cadre de ce rapport. Pour un regard rétrospectif sur la politique scientifique et technologique du gouvernement du Canada, voir Paul Dufour et Yves Gingras, « La politique scientifique et technologique du gouvernement du Canada » dans Robert Dalpé et Réjean Landry, dir, *La politique technologique au Québec*, Montréal, Presses de l'Université de Montréal, 1993, 129, en ligne : <archipel.uqam.ca/557/1/Politique_scientifique_technologique_Canada.pdf>.

Jusqu'aux années 1970, la recherche scientifique québécoise était conduite principalement au sein des universités et d'instituts de recherche[798]. Cet accent, mis exclusivement sur l'enseignement supérieur et la recherche comme fin en soi, s'était fait au détriment d'un arrimage adéquat avec les besoins de l'industrie. À cette époque, le Québec investissait peu dans la recherche et le développement (technologique); les chiffres sont éloquents :

> […] en 1981, 1 % seulement de son PIB [du Québec] était investi dans la R&D; en 1979, le Québec se classait au douzième rang, après le Canada et avant l'Italie, parmi les principaux pays de l'OCDE au titre de l'investissement en R-D. […] le Québec, pays peu peuplé, y consacrait 93 $ per capita en 1979, alors qu'en moyenne les douze principaux pays de l'OCDE investisseurs en R&D dépensaient 183 $.
>
> […] en 1979, [l'industrie québécoise] ne dépensait que 0,50 % du PIB québécois dans la recherche, loin derrière la performance suédoise (1,80 %), suisse (1,31 %), japonaise (1,22 %), proche de celle de l'Italie (0,49%), de l'Australie (0,24 %) et du Canada dans son ensemble (0,48 %)[799].

À ce *statu quo* peu propice à l'innovation, le Livre vert du gouvernement sur la *Politique québécoise du développement culturel* (1978)[800], la phase II de l'énoncé de politique économique *Bâtir le Québec*, substitue une vision de la recherche scientifique appliquée envisagée en tant qu'un « instrument du développement économique et social »[801]. La manière d'instrumentaliser la recherche en tant que « force de production »[802] s'avère toutefois moins évidente du fait de « l'impossibilité de mesurer l'utilité de la science et encore plus de la recherche scientifique »[803].

À cette question difficile, notre *Virage technologique* (1982) propose une intervention de l'État axée essentiellement sur trois priorités stratégiques :

- la prestation des programmes d'aide directe pour augmenter le bassin des ressources humaines (hautement qualifiées) et des ressources en capital;

---

798 Yves Gingras, Benoît Godin et Michel Trépanier, « La place des universités dans les politiques scientifiques et technologiques canadiennes et québécoises » dans Denis Bertrand et Paul Beaulieu, dir, *L'État québécois et les universités : Acteurs et enjeux*, Sainte-Foy, Presses de l'Université du Québec, 1999, 69, en ligne : <archipel.uqam.ca/537/1/Place_universite_dans_politiques_scientifique.pdf>.

799 Louise-E Fortin, « La politique technologique québécoise » (1985) 8 Politique 23, doi : <doi.org/10.7202/040496ar>.

800 Ministre d'État au développement culturel du Québec, *La politique québécoise du développement culturel*, vol 2 « Les trois dimensions d'une politique : genres de vie, création, éducation », 1978, en ligne : <classiques.uqac.ca/contemporains/Quebec_gouvernement_du/Politique_qc_devel_culturel_t2/Politique_qc_devel_culturel_t2.pdf>.

801 *Ibid.,* à la p 276.

802 *Ibid.,* à la p 276.

803 *Ibid.,* à la p 276.

- l'amélioration des ressources externes à l'industrie; et
- l'établissement des liens privilégiés avec l'université.

Cette vision radicalement nouvelle était assortie de priorités sectorielles comprenant l'informatique-électronique et les biotechnologies. Elle sera mise en œuvre par le biais d'incitatifs fiscaux à l'innovation (3.3.1), d'aides financières directes (3.3.2) et des politiques favorables à l'attraction des talents étrangers (3.3.3).

3.1.1 Des soutiens fiscaux en recherche et développement (R&D)

À la fois source de financement, levier d'innovation et mécanisme de redistribution des richesses, la fiscalité se présente comme un instrument de politique économique des plus polyvalents. Au-delà du financement des services publics, les dépenses fiscales permettent en effet aux gouvernements de poursuivre des objectifs stratégiques en accordant des allègements fiscaux à des groupes déterminés de contribuables ou à l'égard de certaines activités. Ces dépenses fiscales peuvent prendre diverses formes, qu'il s'agisse de la non-imposition de certains revenus, des taux réduits d'imposition, de déductions dans le calcul du revenu imposable, de crédits ou de reports d'impôt.

Au Canada, tant les gouvernement fédéral que des provinces canadiennes[804] offrent des mesures d'aide fiscale aux entreprises pour la recherche scientifique et le développement expérimental (RS&DE), sous forme de crédits d'impôt remboursables relatif aux salaires versés aux employés qui travaillent directement dans le domaine de RS&DE ainsi que les employés de soutien au RS&DE[805]. Le crédit peut également porter sur la moitié du montant d'un contrat de recherche relatif à des travaux de RS&DE effectués par un sous-traitant n'ayant pas de lieu de dépendance avec l'entreprise. Le taux de crédit d'impôt, établi aujourd'hui à 14 %, peut atteindre 30 % dépendamment de l'actif de la société privée si elle est sous contrôle canadien, et ce, indépendamment de la réussite ou de la rentabilité des projets. Le taux progressif est établi de telle sorte que les petites et moyennes entreprises (PME) sous contrôle canadien bénéficient des crédits d'impôt les plus avantageux.

---

804 À l'exception notable de l'Île-du-Prince-Édouard. Le gouvernement de l'Alberta, de son côté, a éliminé plusieurs crédits d'impôt provinciaux, dont le crédit provincial en RS&DE, à compter du 1er janvier 2020, y préférant une approche plus inclusive pour encourager la création d'emplois avec l'abaissement du taux d'imposition des sociétés albertaines : Alberta Treasury Board and Finance, *Fiscal Plan : A Plan for Jobs and the Economy 2020-23*, Budget 2020, Edmonton, février 2020 à la p 171, en ligne : <open.alberta.ca/dataset/05bd4008-c8e3-4c84-949e-cc18170bc7f7/resource/79caa22e-e417-44bd-8cac-64d7bb045509/download/budget-2020-fiscal-plan-2020-23.pdf>.

805 *Loi de l'impôt sur le revenu* (Canada), LRC 1985, c 1 (5e suppl), art 37; *Loi sur les impôts* (Québec), LRQ c I-3, art 1029.6.1–1029.8.0.2, 1029.8.9.1 – 1029.8.16.1.

Dans leur forme actuelle[806], les crédits d'impôt RS&DE relatifs aux salaires ont été institués au courant des années 1980 « pour tenir compte de l'importance de la R&D dans la croissance et la compétitivité du Canada »[807] ainsi que « développer les conditions nécessaires à l'éclosion du potentiel innovateur de l'économie »[808].

La RS&DE comporte des activités impliquant une « investigation ou recherche systématique d'ordre scientifique ou technologique, effectuée par voie d'expérimentation ou d'analyse »[809], qui correspondent aux recherches pure et appliquée ainsi qu'au développement expérimental, y compris les travaux de génie, la conception, la recherche opérationnelle, l'analyse mathématique, la programmation informatique; la collecte de données – à l'exclusion de la collecte « normale » ou « courante » de données, les essais et la recherche psychologique. Les travaux de RS&DE admissibles comprennent des travaux directement à la recherche pure ou appliquée ou au développement expérimental, de même que des travaux de soutien effectués en fonction des besoins liés au RS&DE. Ils doivent avoir été entrepris au Canada.

Le Québec offre par ailleurs trois autres crédits d'impôt spécifiques aux activités de R&D qui n'ont pas d'équivalent au fédéral, soit :

- depuis le 1er mai 1987, un crédit d'impôt remboursable pour la recherche universitaire, effectuée par un centre de recherche public ou par un consortium de recherche[810], afin d'encourager la synergie entreprise-universitaire initialement envisagée dans les domaines de l'aérospatiale, des biotechnologies, de l'informatisation des entreprises, de la micro-électronique et des nouveaux matériaux[811]). Ce crédit porte sur 80 % du montant d'un contrat de recherche, lorsque les travaux de R&D sont confiés en sous-traitance à une entité universitaire, centre de recherche public ou consortium de recherche admissible qui n'est pas lié à l'entrepreneur;

---

806 Pour un regard historique sur l'évolution des incitatifs fiscaux en matière de RS&DE au Canada, voir Gouvernement du Canada, *Évolution du Programme de la RS&DE – une perspective historique*, 2015, en ligne : <www.canada.ca/fr/agence-revenu/services/recherche-scientifique-developpement-experimental-programme-encouragements-fiscaux/evolution-programme-perspective-historique.html>.

807 Ministre des Finances du Canada, *Pour assurer le renouveau économique. Documents budgétaires*, déposés à la Chambre des Communes, 23 mai 1985 à la p 16, en ligne : <www.budget.gc.ca/pdfarch/1985-pap-fra.pdf>.

808 Ministre des Finances du Québec, *Budget 1987-1988. Discours sur le budget et Renseignements supplémentaires*, 30 avril 1987, Annexe A à la p 5, en ligne : <www.budget.finances.gouv.qc.ca/budget/archives/fr/documents/1987-88_fine.pdf>.

809 *Loi de l'impôt sur le revenu* (Canada), *supra* note 805, art 248(1) « activités de recherche scientifique et de développement expérimental »; *Loi sur les impôts* (Québec), *supra* note 805, art 222.

810 *Loi sur les impôts* (Québec), *supra* note 805, art 1029.8.1–1029.8.7.2.

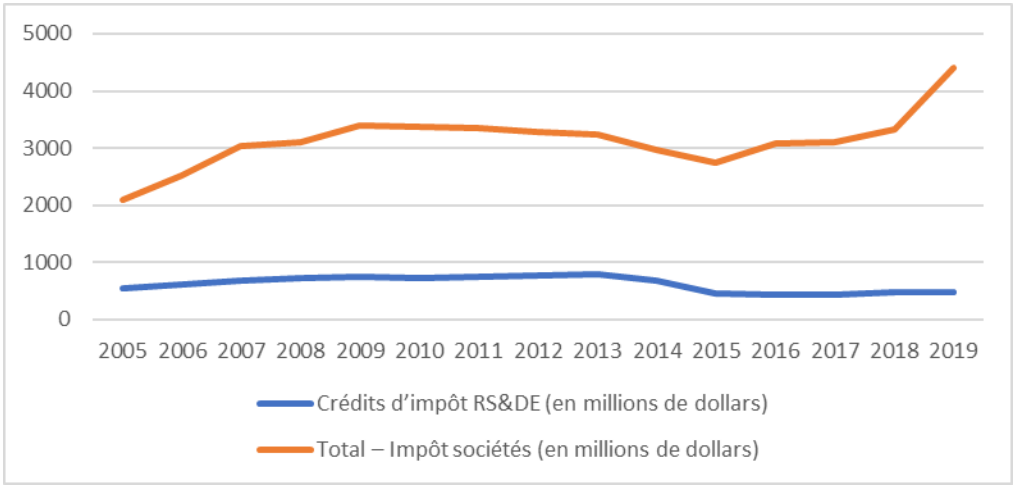
811 Ministre des Finances du Québec, *Budget 1988-1989. Discours sur le budget et Renseignements supplémentaires,* 12 mai 1988, à la p. 23, en ligne : <www.budget.finances.gouv.qc.ca/budget/archives/fr/documents/1988-89_fine.pdf>.

- depuis le 23 mars 2006, un crédit d'impôt remboursable pour la recherche précompétitive en partenariat privé[812]. Ce crédit porte sur un pourcentage des dépenses admissibles se rapportant aux travaux de RS&DE que plusieurs personnes conviennent d'effectuer au Québec ou de faire effectuer pour leur bénéfice au Québec dans le cadre d'une entente de partenariat de recherche;

- un crédit d'impôt remboursable relatif aux cotisations et aux droits versés à un consortium de recherche, afin d'encourager la collaboration dans la recherche[813]. Ce crédit s'applique aux cotisations et droits qu'une personne verse à un consortium de recherche admissible et qui ont trait aux travaux de RS&DE effectués par le consortium en rapport avec une entreprise de cette personne.

Les quatre crédits d'impôt remboursables (R&D salaire, R&D universitaire, R&D partenariat privé, R&D consortium) constituent le point névralgique des mesures de soutien fiscales relatives aux activités de R&D. Ils visent à inciter les entreprises exploitées au Québec à effectuer et à accroître leurs travaux de R&D. D'année en année, les dépenses fiscales ciblant la recherche scientifique et le développement expérimental (RS&DE) figurent parmi les plus importantes dans le régime d'imposition des sociétés Québec (Dépenses fiscales, 2005, p. III; 2006, p. VI; 2007, p. VIII; 2008, p. VIII; 2009, p. VIII; 2010, p. VIII; 2011, p. VIII; 2012, p. viii; 2013, p. viii; 2014, p. A.23; 2015, p. vii; 2016, p. vii; 017, p. vi; 2018, p. vi; 2019, p. B6).

Graphique 1
Évolution de la part de dépenses fiscales attribuables aux crédits d'impôt RS&DE (2005-2019)



Les montants de crédits d'impôt RS&DE sont restés stables d'une année à l'autre; les points d'infléchissement que l'on peut observer après 2005 et 2014 respectivement, sont attribuables

812 *Loi sur les impôts* (Québec), *supra* note 805, art 1029.8.16.1.1–1029.8.16.1.9.

813 *Ibid.*,, art 1029.8.9.0.2–1029.8.9.0.4.

à une bonification du taux de crédit applicable aux PME après 2005 ainsi qu'à une réduction des taux de crédit et à l'introduction d'un seuil de dépenses minimales admissibles après 2014[814].

Dans son rapport final au gouvernement du Québec pour rendre la fiscalité québécoise plus compétitive, plus efficace et plus équitable eu égard aux cibles fixées pour le retour à l'équilibre budgétaire dans le budget 2014-2015, la Commission d'examen sur la fiscalité québécoise[815] recommande le maintien sans modifications du crédit d'impôt pour la recherche scientifique et le développement expérimental (RS&DE) vu l'impact direct des mesures fiscales « sur la compétitivité des entreprises investissant dans la recherche et le développement » (p. 104). Notamment, il est d'avis de la Commission qu'« [e]n l'absence de soutien gouvernemental, plusieurs de ces investissements seraient menacés » (p. 104).

S'il est difficile de départager parmi les dépenses fiscales consacrées au RS&DE celles qui relèvent spécifiquement de l'intelligence artificielle, il ne fait pas de doute que les travaux de l'IA qui relèvent par ailleurs du RS&DE y sont admissibles.

Pour une perspective comparative, voir :

Castellacci, F. et C.M. Lie, « Do the effects of R&D tax credits vary across industries ? A meta-regression analysis » (2015) 44:4 Research Policy 819, doi : <doi.org/10.1016/j.respol.2015.01.010>

Fjaerli, E. et al., « Evaluation of the Norwegian R&D Tax Credit Scheme » (2010) 5:3 Journal of Technology, Management & Innovation 96, doi : <doi.org/10.4067/S0718-27242010000300007>

OCDE, *OECD Compendium of Information on R&D tax incentives*, 2020, en ligne : <www.oecd.org/sti/rd-tax-stats-compendium.pdf>

3.1.2 Des subventions directes à la recherche, à la formation et à l'emploi

Dans la foulée du *Virage technologique* (1982) [renvoi à la section introductive 3.3], le soutien à l'innovation demeure depuis une préoccupation constante du gouvernement québécois. On pense notamment à la mise en place du Fonds de développement technologique (1989) ainsi que d'autres initiatives comme Innovation Québec et Valorisation-Recherche Québec (1999). C'est toutefois à partir du budget 2017-2018 que l'expression « intelligence artificielle » fait son apparition récurrente dans les discours du gouvernement.

814 Dépenses fiscales, 2019, à la p. B13, en ligne : <https://numerique.banq.qc.ca/patrimoine/details/52327/17108?docref=oIkVd6SLeXhb0T5hmDCYxQ>.

815 Commission d'examen sur la fiscalité québécoise, *Compétitivité, efficacité, équité. Se tourner vers l'avenir du Québec*, vol 1 « Une réforme de la fiscalité québécoise », rapport final, gouvernement du Québec, mars 2015, en ligne : <www.groupes.finances.gouv.qc.ca/examenfiscalite/uploads/media/Volume1_RapportCEFQ_01.pdf>.

Le Plan économique du Québec, annoncé lors du Budget 2017-2018, endosse la proposition du Conseil consultatif sur l'économie et l'innovation (CCEI)[816], de « faire émerger un écosystème de l'intelligence artificielle pour stimuler le développement et l'adoption de ses applications » (p. B.17). Plus particulièrement, une contribution de 100 millions de dollars, échelonnée sur six ans (2016 à 2022), sera versée au ministère de l'Économie, de la Science et de l'Innovation pour « la création d'une super-grappe en intelligence artificielle » (p. B92). Cette initiative, s'inscrivant dans le cadre de la Stratégie québécoise de la recherche et de l'innovation (2017-2022)[817], s'articule autour de cinq volets d'intervention prioritaires qui sont :

- l'attraction et la rétention des talents dans un contexte où la demande pour l'expertise en apprentissage profond est très forte;

- la consolidation d'une masse critique de chercheurs de haut calibre en intelligence artificielle au Québec pour attirer les jeunes chercheurs à y faire carrière et les industriels à s'y intéresser;

- la création d'un environnement d'affaires favorable à la valorisation et à la commercialisation des progrès scientifiques en produits et en solutions;

- le démarrage d'entreprises et l'accès au capital de risque dans le domaine de l'intelligence artificielle au Québec ainsi que la sensibilisation des étudiants des cycles supérieurs au démarrage d'entreprises dans ce domaine;

- l'acceptabilité et l'impact social de l'intelligence artificielle afin de s'assurer que les questions que celle-ci soulève, notamment à l'égard de la confidentialité, sont discutées avec non seulement les experts, mais également les citoyens[818].

Cette super-grappe fera « de Montréal un pôle économique et scientifique de premier plan pour la recherche, la formation, le transfert technologique et la création de produits et de solutions à valeur ajoutée ainsi que d'emplois et d'entreprises spécialisés dans l'exploitation et l'analyse de mégadonnées pour faciliter la prise de décision »[819].

---

816 Le CCEI, formé en octobre 2016 par le gouvernement, regroupe une trentaine de membres entrepreneurs, industriels, investisseurs et dirigeants de grandes institutions du Québec. Il est présidé par Mme Monique Leroux, qui est également présidente du conseil d'administration d'Investissement Québec.

817 Ministère de l'Économie, de la Science et de l'Innovation, *Stratégie québécoise de la recherche et de l'innovation 2017-2022*, gouvernement du Québec, 2017, en ligne : <www.economie.gouv.qc.ca/fileadmin/contenu/documents_soutien/strategies/recherche_innovation/SQRI/sqri_complet_fr.pdf>.

818 Plan économique du Québec, Budget 2017-2018, à la p B.106, en ligne : <http://www.budget.finances.gouv.qc.ca/budget/2017-2018/fr/documents/PlanEconomique_Mars2017.pdf>.

819 *Ibid.*, p. B.104.

S'y ajoutent 15 millions de dollars supplémentaires à être versés au fonds RV Orbite Montréal, qui est un fonds de soutien au démarrage d'entreprises technologiques dans les secteurs des technologies de l'information, des technologies avancées, des technologies liées à l'intelligence artificielle (p.ex. Internet des objets, données massives, robotique, voiture connectée, réalité virtuelle), des technologies financières et des villes intelligentes.

Le Plan économique de mars 2018[820] prévoit « le renforcement du leadership du Québec dans des domaines porteurs tels que l'intelligence artificielle » (p. D.19). Le gouvernement y annonce l'octroi d'une aide de 10 millions de dollars sur cinq ans pour appuyer spécifiquement deux initiatives dans le domaine de l'IA, soit le Creative Destruction Lab (CDL) de Montréal[821] et l'accélérateur NextAI[822] (p. D.23) :

- le Creative Destruction Lab (CDL) de Montréal est un programme d'accompagnement d'une durée de neuf mois offert aux startups technologiques spécialisées en intelligence artificielle;

- le programme NextAI est un programme de formation d'une durée totale de 225 heures et s'échelonnant sur une période de huit mois à l'attention des étudiants, des professionnels et des entrepreneurs intéressés par l'intelligence artificielle. La formation comprend des ateliers d'essais industriels et des ateliers sur les compétences indispensables en affaires.

Le plan budgétaire 2019-2020 du Gouvernement du Québec[823] prévoit une enveloppe de 329,3 millions de dollars qui seront investis en cinq ans pour accélérer spécifiquement l'adoption de l'intelligence artificielle (p. D.33). L'aide vise les volets suivants :

- l'élargissement de l'offre de formation en intelligence artificielle pour les étudiants de tous niveaux et les travailleurs spécialisés (12,5 millions);

- l'attraction au Québec des chercheurs en intelligence artificielle dans les universités québécoises et soutenir la formation doctorale et postdoctorale (38 millions);

- l'adoption de l'intelligence artificielle dans les entreprises et les organisations publiques (65 millions);

- l'augmentation de la puissance de calcul du Québec, comprenant l'aménagement des locaux pour héberger des calculateurs de haute performance et l'acquisition d'épquipements spécialisés (34,5 millions);

---

820 Plan économique du Québec, Budget 2018-2018, en ligne : <http://www.budget.finances.gouv.qc.ca/budget/2018-2019/fr/documents/PlanEconomique_18-19.pdf>.

821 Voir en ligne : <https://creativedestructionlab.com/locations/montreal/>.

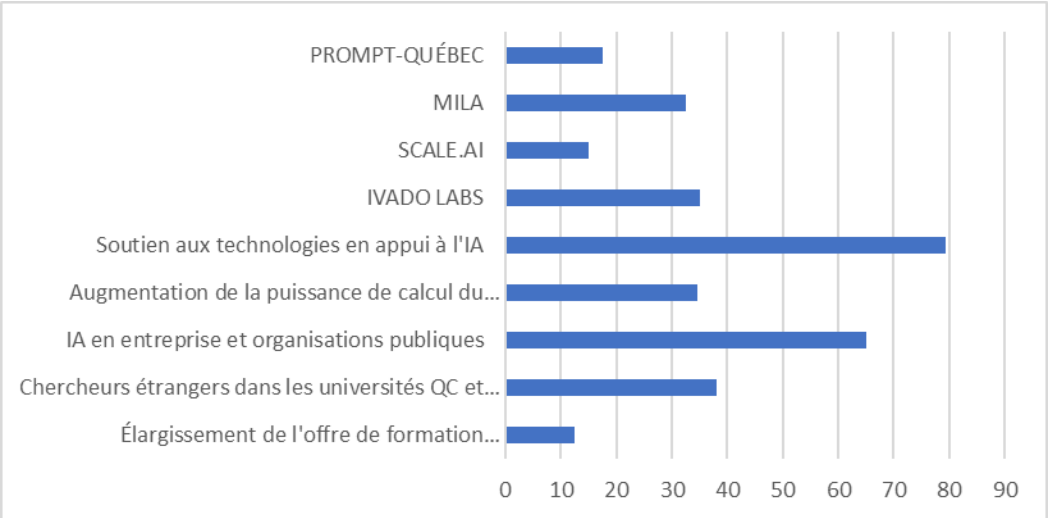
822 Voir en ligne : < https://www.nextcanada.com/next-ai/>.

823 Gouvernement du Québec, *Vos priorités, votre budget : Plan budgétaire*, budget 2019-2020, mars 2019, en ligne : <www.budget.finances.gouv.qc.ca/budget/2019-2020/fr/documents/PlanBudgetaire_1920.pdf>.

- le soutien aux technologies en appui à l'intelligence artificielle, notamment dans les domaines du design électronique, de l'optique-photonique et des semi-conducteurs (79,3 millions); et

- le soutien aux activités de recherche en intelligence artificielle, et plus particulièrement aux quatre initiatives suivantes (100 millions) :
  - IVADO LABS (35 millions);
  - SCALE.AI (15 millions);
  - MILA – Institut québécois d'intelligence artificielle (32,5 millions);
  - PROMPT-QUÉBEC (17,5 millions).

Graphique 2
Enveloppe budgétaire prévue sur 5 ans pour accélérer l'adoption de l'IA (en millions de dollars)



En date de la rédaction des présentes, le ministère de l'Économie et de l'innovation du Québec finance 142 projets en cours dans le domaine de l'intelligence artificielle à hauteur de 23 717 524 $ pour des coûts totaux de 59 749 754,17 $, représentant un pourcentage de financement moyen de 43 % réparti comme suit :

- 54 projets d'innovation en entreprises solos à hauteur de 3 193 316 $ pour des coûts totaux de 10 485 780 $, représentant un pourcentage de financement moyen de 34 %;

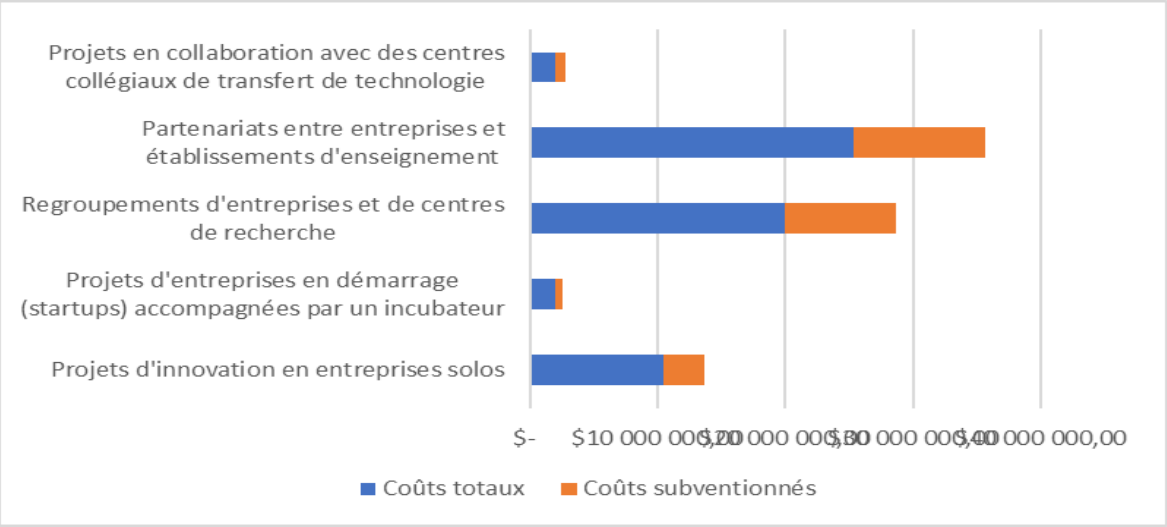
- 23 projets d'entreprises en démarrage (*startups*) accompagnées par un incubateur, à hauteur de 572 000 $ pour des coûts totaux de 2 013 271 $, représentant un pourcentage de financement moyen de 36 %;

- 31 regroupements d'entreprises et de centres de recherche, à hauteur de 8 727 307 $ pour des coûts totaux de 19 967 281,17 $, représentant un pourcentage de financement moyen de 47 %;

- 30 partenariats entre entreprises et établissements d'enseignement, à hauteur de 10 357 301 $ pour des coûts totaux de 25 359 922 $, représentant un pourcentage de financement moyen de 48 %;

- 5 projets en collaboration avec des centres collégiaux de transfert de technologie (CCTT), à hauteur de 867 600 $ pour des coûts totaux de 1 923 500 $, représentant un pourcentage de financement moyen de 50 %.

Graphique 3
Projets IA en cours financés par le Ministère QC de l'Économie et de l'Innovation



Pour une perspective comparative, voir :

Afcha, S. et G.L. López, « Public funding of R&D and its effect on the composition of business R&D expenditure » (2014) 17:1 BRQ 22, doi : <doi.org/10.1016/j.cede.2013.01.001>

Antonelli, C. et F. Crespi, « The Matthew effect in R&D public subsidies : The Italian evidence » (2013) 80:8 Technological Forecasting & Social Change 1523, doi : <doi.org/10.1016/j.techfore.2013.03.008>

Commission européenne, *The Economic Rationale for Public R&I Funding and its Impact*, Luxembourg, mars 2017, en ligne : <ri-links2ua.eu/object/document/326/attach/KI0117050ENN_002.pdf>

Grilli, L. et S. Murtinu, « De public subsidies affect the performance of new technology-based firms ? The importance of evaluation schemes and agency goals » (2012) 30:1 Critical Studies in Innovation 97, doi : <doi.org/10.1080/08109028.2012.676836>

Malahov, V., « Public Support Programs for Applied Research Conducted by Small and Medium-Sized Businesses : a Review of International Experience » (2017) 12:2 Science Governance & Scientometrics Journal 5, en ligne : <ideas.repec.org/a/akt/journl/v12y2017i2p5-29.html>

Mewes, L. et T. Broekel, « Subsidized to change ? The impact of R&D policy on regional technological diversification » (2020) 65 The Annals of Regional Science 221, doi : <doi.org/10.1007/s00168-020-00981-9>

Zúñiga-Vicente, J. et al., « Assessing the Effect of Public Subsidies on Firm R&D Investment : A Survey » (2014) 28:1 Journal of Economic Surveys 36, doi : <doi.org/10.1111/j.1467-6419.2012.00738.x>

3.1.3 Des politiques favorables à l'attraction des talents étrangers

En complément aux aides directes et soutiens fiscaux accordés pour les activités de R&D entreprises au Canada et encourager la création d'emplois, des mesures spécifiques ont été adoptées pour faciliter le recrutement de l'expertise étrangère dans la recherche et le développement (R&D), notamment en intelligence artificielle (IA). Encore une fois, le gouvernement québécois a opté pour le mécanisme des incitatifs fiscaux pour encourager la mise en œuvre de projets d'innovation sur le territoire du Québec. Plus récemment, un projet pilote d'immigration a été lancé à l'attention des travailleurs de l'IA.

Congé fiscal au bénéfice des chercheurs, experts et stagiaires postdoctoraux étrangers – Depuis le 1er janvier 1987, dans le but de faciliter le recrutement du personnel spécialisé dans le domaine de la recherche scientifique et le développement expérimental[824], une personne qui ne réside pas au Canada et qui vient travailler au Québec dans le cadre d'un projet lié à la recherche scientifique et au développement expérimental (RS&DE) peut bénéficier d'une exemption d'impôt sur le salaire qui lui est versé à titre de chercheur étranger[825]. Cette exemption prend la forme d'une déduction dans le calcul du revenu imposable du chercheur étranger. Depuis le 31

824 Ministre des Finances du Québec, *supra* note 808, aux pp 6, 7.

825 *Loi sur les impôts* (Québec), *supra* note 809, art 737.19–737.22.

mars 1998[826], cette mesure a été élargie pour viser également certains stagiaires postdoctoraux étrangers recrutés par les entités universitaires et centres de recherche publics admissibles[827]. L'année suivante[828], cette mesure a été derechef bonifiée pour étendre la période d'exemption de deux à cinq ans et vise désormais également les experts étrangers[829].

Programme pilote d'immigration à l'attention des travailleurs de l'IA – Le 28 octobre 2020, le gouvernement du Québec publie un projet de règlement lançant notamment un programme pilote quinquennal d'immigration permanente à l'attention « des travailleurs des secteurs de l'intelligence artificielle, des technologies de l'information et des effets visuels »[830]. Ce projet pilote comporte deux volets, « Intelligence artificielle » et « Technologies de l'information et effets visuels », s'adressant tant aux francophones qu'aux non-francophones, pour un nombre maximal de 550 places par année réparties également entre les deux profils linguistiques.

Parmi les conditions de sélection générales du volet « Intelligence artificielle », un ressortissant étranger même non francophone est admissible s'il occupe ou a accepté un emploi à temps plein au Québec dans le secteur de l'intelligence artificielle et dont le salaire annuel brut est d'au moins 75 000 $ ou de 100 000 $ si le lieu habituel d'emploi se situe sur le territoire de la Communauté métropolitaine de Montréal (*sous-volet « Travailleur étranger »*). Pour les titulaires d'un diplôme d'études supérieures délivré par un établissement d'enseignement au Québec, ils pourront être admissibles s'ils occupent ou ont accepté un emploi à temps plein au Québec dans le secteur de l'intelligence artificielle (*sous-volet « Étudiant étranger diplômé du Québec »*).

En ce qui concerne le volet « Technologies de l'information et effets visuels », comptent parmi les emplois admissibles ceux d'analyste et consultant en informatique, de gestionnaire des systèmes informatiques, d'ingénieur et concepteur en logiciel, de programmeur et développeur en médias interactifs, ainsi que de technicien de réseau informatique.

---

826 Ministre des Finances du Québec, *Budget 1998-1999. Renseignements supplémentaires sur les mesures du budget*, 31 mars 1998, à la p 14, en ligne : <www.budget.finances.gouv.qc.ca/budget/1998-1999/fr/PDF/rensupfr.pdf>.

827 *Loi sur les impôts* (Québec), *supra* note 809, art 737.22.0.0.1–737.22.0.0.4.

828 Ministre des Finances du Québec, *Budget 1999-2000. Renseignements supplémentaires sur les mesures du budget*, 9 mars 1999, à la p 24, en ligne : <www.budget.finances.gouv.qc.ca/budget/1999-2000/fr/PDF/disc-fr.pdf >.

829 *Loi sur les impôts* (Québec), *supra* note 809, art 737.22.0.0.5–737.22.0.0.8.

830 *Règlement édictant trois programmes pilotes d'immigration permanente* (projet), (2020) 152 G.O. II, 4592.

Pour aller plus loin (dans une perspective comparée) :

Arnold, Z. et al., *Immigration Policy and the U.S. AI Sector*, rapport préliminaire, Center for Security and Emerging Technology (CSET), septembre 2019, en ligne : <cset.georgetown.edu/publication/immigration-policy-and-the-u-s-ai-sector/>

Cerna, L. et M. Czaika, « European Policies to Attract Talent : The Crisis and Highly Skilled Migration Policy Changes » dans Triandafyllidou, A. et I. Isaakyan, *High-Skill Migration and Recession. Gendered Perspectives*, Springer, 2016, 22, doi : <doi.org/10.1057/9781137467119_2>

Hercog, M. et A. Wiesbrock, « Highly Skilled Migration to the European Union and the United States » dans Besharov D.J. et M.H. Lopez, *Adjusting to a World in Motion : Trends in Global Migration and Migration Policy*, Oxford Scholarship Online, 2016, doi : <doi.org/10.1093/acprof:oso/9780190211394.001.0001>

Cameron, A. et S. Faisal, *Digital Economy Talent Supply : Immigration Stream*, Information and Communications Technology Council (ICTC), Ottawa, 2016, en ligne : <www.ictc-ctic.ca/wp-content/uploads/2016/09/Digital-Economy-Supply_The-Immigration-Stream.pdf>

Mosbah, A. et al., « Migrants in the High-Tech and Engineering Sectors : An Emerging Research Area » dans *2018 IEEE Conference on Systems, Process and Control (ICSPC)*, 2018, 234, doi : <doi.org/10.1109/SPC.2018.8704139>

Rand, D. et L. Milliken, « Winning the Global Race for Artificial Intelligence Expertise. How the Executive Branch Can Streamline U.S. Immigration Options for AI Talent », *NYU Journal of Legislation & Public Policy*, 9 avril 2021, en ligne : <nyujlpp.org/quorum/rand-milliken-winning-global-race-artificial-intelligence/>

Zwetsloot, R. et al., « Skilled and Mobile : Survey Evidence of AI Researchers' Immigration Preferences » (2021), en ligne : <arxiv.org/abs/2104.07237>

***

Les différents instruments d'intervention privilégiés par le gouvernement du Québec pour encourager l'innovation poursuivent ainsi un objectif commun, soit l'attraction et la rétention des investissements tant en capital technique qu'humain sur le territoire québécois. Dans la mesure où les mêmes modalités sont par ailleurs adoptées par d'autres États, une véritable course à l'innovation et à l'hégémonie par l'intelligence artificielle (IA) s'entreprend non plus comme un choix politique, mais s'impose surtout de nécessité pour prévenir un exode des cerveaux et conserver la mainmise nationale sur un écosystème numérique qui transcende les frontières. Nous détachant, pour un temps, de cette course et de l'engouement ambiant, la vue plus panoramique qui s'offre à nous de cette mosaïque géopolitique est-elle à la hauteur d'autant d'ambitions ?

3.2 Regard critique sur les politiques de l'intelligence artificielle (IA)

À n'en pas douter, l'énoncé d'autant de visions stratégiques quant au développement de l'intelligence artificielle (IA) marque une étape névralgique dans la planification d'entreprises et la formulation de politiques publiques à plus ou moins long terme. Or, les progrès technoscientifiques s'accompagnent bien souvent d'attentes et de promesses emphatiques qui

s'entretiennent en partie de la difficulté d'appréhender les conditions de leur réalisation. Chaque nouvelle percée est applaudie comme annonçant d'autres sauts quantiques à l'échéancier indéterminé, cependant qu'ils suscitent dans l'intervalle un climat sociologique plein d'effervescence ainsi qu'un imaginaire populaire séduit par autant d'intrigues dignes de science-fiction.

À telles enseignes que les économies de la promesse (*sociology of expectations*) s'imposent dorénavant comme une sous-discipline à part entière, s'intéressant à la manière dont l'innovation (techno-scientifique) est propulsée par l'énoncé d'attentes et d'anticipations qui deviennent autant de promesses dont les parties prenantes se doivent de rendre compte, en termes d'investissements, d'allocation de ressources ou de mesure de progrès[831].

Pour justifier l'intervention des autorités publiques dans les politiques de soutien à l'innovation technologique, l'argument traditionnel qu'apporte la théorie économique consiste à postuler une meilleure efficience de la main « visible » pour orienter l'allocation des ressources dans l'intérêt social et éponger le risque inhérent aux activités de recherche et développement (R&D). Ce postulat suppose un État omniscient disposant d'une information complète sur la pertinence de ses interventions; ce qui, en pratique, n'est pas toujours le cas[832].

Or, le bénéfice social découlant de l'innovation est incertain. L'horizon temporel sur lequel peut s'étaler la pondération des coûts et bénéfices a été bien souvent omis dans l'équation des économistes; le long terme est en effet une variable qui, en raison de la multitude de facteurs susceptibles d'intervenir entretemps, peut difficilement être pris en compte sans une grande marge d'incertitude. Or, les différentes stratégies nationales tablent précisément sur ce pari temporel en minimisant l'importance des coûts encourus contre un bénéfice escompté à une plus longue échéance.

Donc, le vecteur temporel met en perspective des résultats présents insuffisants à travers le prisme d'avancées révolutionnaires attendues dans un futur plus ou moins lointain. Comme nous le rapporte le National Science and Technology Council (NSTC) dans son Plan stratégique relatif à l'intelligence artificielle (2016) [833]:

> AI research investments are needed in areas with potential long-term payoffs. (…) These payoffs can be seen in 5 years, 10 years, or more. A recent National Research Council report emphasizes the critical role of Federal investments in long-term research, noting "the long, unpredictable incubation period – requiring steady work and funding –

831 Harro Van Lente, « Navigating foresight in a sea of expectations : Lessons from the sociology of expectations » (2012) 24:8 Technology Analysis & Strategic Management 769, doi : <doi.org/10.1080/09537325.2012.715478>.

832 Dominique Guellec, « Les politiques de soutien à l'innovation technologique à l'aune de la théorie économique » (2001) 4-5: 150-151 Économie & Prévision 95, doi : <doi.org/10.3917/ecop.150.0095>.

833 The National Artificial IntelligVoir en ligne : <https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf>.

> between initial exploration and commercial deployment." It further notes that "the time from first concept to successful market is often measured in decades". Well-documented examples of sustained fundamental research efforts that led to high-reward payoffs include the World Wide Web and deep learning. In both cases, the basic foundations began in the 1960s; it was only after 30+ years of continued research efforts that these ideas materialized into the transformative technologies witnessed today in may categories of AI.[834]

À noter que cette perspective de gains élevés (*high payoffs*) résultant d'investissements durables en IA est énoncée en termes non pas d'éventualités ou de probabilités, mais plutôt de certitudes, quoique futures. Contrairement aux entreprises privées, l'État n'est pas pressé par le temps, c'est le « maître des horloges » qui anime une « synergie productive entre l'État et les différents agents de l'économie »[835]. Un peu paradoxalement, ces attentes qui ne sont pas nécessairement réalistes, peuvent être exacerbées par l'incertitude inhérente à tout développement technologique et recherche expérimentale. Aux yeux des non-initiés moins aptes à apprécier les difficultés de parcours ou les limites théoriques des progrès technoscientifiques, incertitude plus (+) efforts investis donne (=) aisément une promesse de résultat. Les longs interludes, comme les deux hivers qui ont balisé, il n'y a pas si longtemps, le développement de l'intelligence artificielle (IA) [renvoi au Chapitre 1], sont trop facilement relégués dans l'oubli. C'est comme si ce marché parallèle des promesses alimente un marché « des *futures* » qui spécule sur l'ensemble de nos attentes dérivées de l'actif technologique sous-jacent. Après tout, les réseaux de neurones multi-couches à la base de l'apprentissage profond [renvoi au Chapitre 1] n'ont-ils pas connu un succès enviable en multipliant de façon exponentielle leurs couches (variables) intermédiaires dont il devient difficile de pénétrer le fonctionnement ?

L'autre variable de l'équation économique, soit les coûts de production et de diffusion de l'innovation, a aussi été sous-estimée. La connaissance est loin d'être « un bien public pur, circulant parfaitement et sans coût »[836]. Certaines recherches, par leur nature même (p.ex. longitudinale, fondamentale), nécessitent un investissement soutenu à long terme pour avoir une chance de porter fruit. Certaines, mais pas toutes. Ainsi, à la différence de la recherche fondamentale, le soutien public à la recherche appliquée est traditionnellement plus

---

834 *The National Artificial Intelligence Research and Development Strategic Plan*, October 2016, aux pp 16, 17, en ligne : <https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf>.

835 Ministre des Finances du Québec, *Budget 1999-2000. Discours sur le budget*, 9 mars 1999 à la p 18, en ligne : <www.budget.finances.gouv.qc.ca/budget/1999-2000/fr/PDF/disc-fr.pdf>.

836 Dominique Guellec, *supra* note 827.

controversé[837], quoique, de plus en plus, des auteurs militent pour une combinaison optiale des deux (recherches fondamentale et appliquée)[838].

Dans le pire des cas, une intervention inadéquate de l'État peut contribuer à créer une (autre) défaillance du marché plutôt que d'y remédier. Par exemple, l'octroi d'aides financières directes confie à l'État l'épineuse responsabilité de choisir ses bénéficiaires en amont. Or, en raison du risque inhérent à tout à projet se voulant innovateur, il peut être difficile de séparer le bon grain de l'ivraie au stade initial des appels d'offres. À cet égard, l'État ne dispose pas nécessairement d'une prescience supérieure aux bailleurs de fonds privés. Discriminer, dans ce contexte, emporterait aussi une prise de risque s'ajoutant aux externalités négatives que l'intervention publique cherche justement à éviter.

Au Québec, les incitatifs fiscaux en R&D constituent le pilier autour duquel s'est structurée notre politique de soutien à l'innovation. Il s'agit d'une tendance partagée par la plupart des pays industrialisés qui privilégient, de plus en plus, les incitatifs fiscaux aux aides financières directes[839]. Les soutiens fiscaux peuvent être vus comme plus englobants en ce qu'ils visent tout projet répondant aux critères d'admissibilité et sont octroyés sur demande au prorata des dépenses engagées en R&D sans égard aux résultats ou à la rentabilité des projets. Ces attributs rendent la politique fiscale particulièrement apte à soutenir l'innovation à long terme sans exiger de reddition de compte immédiate. En ne discriminant pas ses candidats sur la base des critères de rendement, l'État assume donc par sa politique fiscale une partie importante du risque financier inhérent à toute activité de R&D. Cela étant, il ne faudrait pas perdre de vue que l'État joue ainsi en quelque sorte le rôle d'un assureur en remboursant, sans remise en question, une partie des coûts engendrés par les activités de R&D. Cette approche assurantielle doit cependant être adéquatement dosée : si le fait d'offrir un tremplin de sécurité peut constituer un incitatif à innover, il risque par ailleurs de diminuer l'efficacité des projets entrepris sans la pression de rentabiliser ou de rendre compte à ses subventionnaires.

Afin de jauger de l'efficacité des incitatifs fiscaux à la R&D, l'analyse coût-bénéfice, sur le plan théorique, nécessiterait une comparaison entre le coût social résultant de l'octroi des crédits d'impôt ou les sacrifices sociaux résultant de la privation des mêmes fonds qui auraient pu être alloués à d'autres fins d'utilité sociale – et le bénéfice social découlant des travaux de R&D qui

---

837 Martin LaMonica, « Should the Government Support Applied Research? », *MIT Technology Review* (10 septembre 2012), en ligne : <www.technologyreview.com/2012/09/10/183924/should-the-government-support-applied-research/>.

838 Gianni De Fraja, « Optimal public funding for research : a theoretical analysis » (2016) 47:3 The RAND Journal of Econoics 498, en ligne : <www.jstor.org/stable/43895655>; National Research Council, *Furthering America's Research Enterprise*, Washington, DC, The National Academies Press, 2014, doi : <doi.org/10.17226/18804>; André Oosterlinck, Koen Debackere et Gerard Cielen, « Balancing basic and applied research » (2002) 3:1 EMBO Rep 2, doi : <doi.org/10.1093/embo-reports/kvf016>.

839 Bruno Van Pottelsberghe De La Potterie, « Les politiques de science et technologie et l'objectif de Lisbonne » (2004) XLIII:1 Reflets et perspectives de la vie économique 69, doi : <doi.org/10.3917/rpve.431.0069>.

n'auraient pas été entrepris n'eussent été les incitatifs[840]. À défaut de cette information, les chercheurs ont opté pour l'alternative d'estimer les dépenses de R&D supplémentaires qui ont été engendrées par les dépenses fiscales. Les résultats obtenus à date sur l'efficacité des incitatifs fiscaux demeurent controversés (Mansfield et Switzer, 1985[841]; Dagenais, Mohnen et Therrien, 2004). Au Canada, il semblerait qu'une grande partie des dépenses fiscales ont de fait servi à financer des travaux de R&D qui auraient été entrepris de toute façon (*Ibid.*).

Plus spécifiquement en ce qui concerne le développement de l'intelligence artificielle (IA), un rapport de l'Institut de recherche et d'informations socio-économiques (IRIS), rendu public en mars 2019, remet en question les retombées sociales et économiques attendues des investissements massifs octroyés par les gouvernements canadien et québécois. Selon les chercheurs, ces investissements massifs ne vont pas nécessairement dans le sens de l'intérêt collectif, en ce qu'ils favorisent ultimement « la concentration des richesses dans les mains de quelques acteurs »[842]. En particulier, le phénomène d'acquisition des startups financés par les fonds publics permet aux géants du Web de consolider leur intégration verticale et leur position oligopolistique tant sur le marché des données que l'expertise dans le développement de l'intelligence artificielle (IA). Non seulement l'acquisition des startups leur donne accès aux bases de données des entreprises acquises pour se constituer une masse critique d'utilisateurs, mais le rachat des brevets entrave la circulation des connaissances et des innovations dans l'intérêt public. Ces constats rejoignent les conclusions de Colleret et Gingras (2020)[843] qui relèvent la concentration « de l'ensemble des décisions touchant l'IA au Québec dans les mains d'une dizaine d'acteurs multipositionnels » (p. 37).

Si, du point de vue de la théorie économique, l'impératif d'efficacité commande de ne pas « financer avec de l'argent public des recherches privées qui auraient de toute façon eu lieu »[844], un autre vecteur sustentateur d'attentes technoscientifiques est l'instrumentalisation des progrès en intelligence artificielle (IA) en tant qu'un outil géopolitique de puissance, à telles enseignes que certains n'hésitent pas à présager un « retour des empires »[845]. En effet, une véritable course à l'hégémonie par l'intelligence artificielle (IA) est entamée non seulement entre

---

840 Marcel Dagenais, Pierre Mohnen et Pierre Therrien, « Les firmes canadiennes répondent-elles aux incitations fiscales à la recherche-développement? » (2004) 80: 2-3 *L'Actualité économique* 175, doi : <doi.org/10.7202/011385ar>.

841 Edwin Mansfield et Lorne Switzer, « The effects of R&D tax credits and allowances in Canada » (1985) 14:2 *Research Policy* 97, doi : <doi.org/10.1016/0048-7333(85)90017-4>.

842 Lisiane Lomazzi, Myriam Lavoie-Moore et Joëlle Gélinas, *Financer l'intelligence artificielle, quelles retombées économiques et sociales pour le Québec?* Institut de recherche et d'informations socio-économiques (IRIS), mars 2019 à la p 7, en ligne : <cdn.iris-recherche.qc.ca/uploads/publication/file/Intelligence_artificielle_IRIS_WEB4.pdf>.

843 Maxime Colleret et Yves Gingras, *supra* note 329.

844 Chantal Kegels, *supra* note 768.

845 Nicolas Miailhe, « Géopolitique de l'intelligence artificielle : le retour des empires? » (2018) 3 Politique étrangère 105, en ligne : <www.ifri.org/sites/default/files/atoms/files/geopolitique_de_lintelligence_artificielle.pdf>.

entreprises, mais également entre États voire régions géopolitiques pour se positionner en « chef de file » mondial dans ce domaine à hautes promesses technoscientifiques. Outre la nécessité d'investissements durables à long terme considérés comme un prérequis à d'éventuelles percées révolutionnaires, un deuxième argument, très souvent invoqué, a trait aux investissements réalisés par d'autres juridictions, justifiant de son côté un montant à investir encore plus conséquent et accroissant d'autant la valeur perçue des technologies intelligentes. C'est ainsi que la Commission européenne, dans sa Stratégie européenne pour l'IA (2018) justifie en ces termes la nécessité pour l'union européenne d'investir (encore plus) massivement dans l'IA :

> Le gouvernement américain a présenté une stratégie en matière d'IA et a investi quelque 970 millions d'EUR dans la recherche non confidentielle en matière d'IA en 2016. Avec son « plan de développement de l'intelligence artificielle de nouvelle génération », la Chine ambitionne de devenir le leader mondial d'ici à 2030, procédant à des investissements massifs. D'autres pays tels que le Japon et le Canada ont aussi adopté des stratégies en matière d'IA.
>
> Aux États-Unis et en Chine, les grandes entreprises investissent massivement dans l'IA et exploitent de grandes quantités de données.
>
> De manière générale, **l'Europe accuse un retard en matière d'investissements privés** dans l'IA; ceux-ci se sont élevés à environ 2,4-3,2 milliards d'EUR en 2016, contre 6,5-9,7 milliards EUR en Asie et 12,1-18,6 milliards d'EUR en Amérique du Nord.
>
> Il est donc essentiel que l'UE continue à œuvrer à la **création d'un environnement qui stimule les investissements** et utilise des fonds publics pour attirer des investissements privés. Pour ce faire, L'UE doit **préserver ses atouts et en tirer parti**.[846]

La course – universelle – vers l'économie du savoir n'admet pas de retardataires. De part et d'autre de l'Atlantique, « nous pouvons et devons la gagner »[847]. Des quatre coins du globe, l'on reste rivé sur le Global AI Index, le Government AI Readiness Index et l'AI Index. Cette prolifération d'indicateurs relatifs à l'IA est aussi en elle-même un produit dérivé de cette course, laquelle s'entretient d'elle-même en véritable cycle autopoïétique. Il est donc couramment reconnu que l'intelligence artificielle procure un avantage compétitif non seulement aux entreprises, mais également aux États qui savent en tirer parti tant de son pouvoir de coercition (p.ex. défense nationale) que de son pouvoir de convaincre (*soft power*) par l'étendue de ses implications sociales, économiques et politiques. Cette course des grandes puissances a eu pour effet d'accroître la valeur perçue des domaines d'intelligence artificielle. C'est aussi une course quelque peu obligée en ce qu'agir à contre-courant des soutiens publics accordés à l'innovation un peu partout dans le monde industrialisé risque d'entraîner une délocalisation des entreprises

---

846 Commission européenne, 2018, p. 4–5, références omises et emphase dans l'original.

847 Ministre des Finances du Québec, *supra* note 829.

et des talents vers l'étranger. Alors qu'en 1981, 1 % du PIB du Québec était investi dans la R&D[848], après deux décennies de soutien public à l'innovation, ce pourcentage est passé à 2,48 % en 2000 puis à 2,30 % en 2013 (Scientifique en chef du Québec, 2016). Ces pourcentages correspondent respectivement au douzième (12^{e}), sixième (6^{e}) et quatorzième (14^{e}) rang des pays de l'OCDE, soit toujours juste en dessous de la moyenne. Rappelons que le gouvernement du Québec prévoyait augmenter les dépenses de R&D jusqu'à 3 % du PIB québécois d'ici 2010[849]. Comme l'illustre cet exemple québécois, l'enjeu pour les « grandes puissances » est de maîtriser le plein potentiel de l'intelligence artificielle (IA) plutôt que d'en rester de simples consommateurs.

Cela étant, en rétrospective, les cycles des engouements technoscientifiques n'ont rien de nouveau. L'arrivée d'une nouvelle technologie suscite immanquablement une première phase d'excitation et d'attentes exagérées alimentée par des modèles de réussite (*success stories*) élevés au rang de légendes. L'intelligence artificielle (IA) ne fait pas exception. Ainsi, le succès du programme AlphaGo à certains jeux de stratégie face aux meilleurs joueurs humains a alimenté des titres médiatiques hyperboliques comme « L'intelligence artificielle ne peut être vaincue »[850] ou « AlphaGo Zero Shows Machines Can Become Superhuman Without Any Help »[851]. Et *quid* des véhicules autonomes que certains considèrent comme « une révolution qui pourrait sauver des vies »[852] et qui « Will Change Everything »[853] ?

À l'instar des bulles spéculatives[854], les discours (médiatiques) hyperboliques finissent inévitablement par décevoir à mesure qu'on se rend compte tant des limites méthodologiques que des défis liés à la mise en œuvre des nouvelles technologiques. AlphaGo ressemble davantage à une calculatrice hyper performante qu'à une véritable « intelligence »[855]. Le

---

848 Louise-E Fortin, *supra* note 799.

849 Ministre des Finances du Québec, *Budget 2006-2007. Renseignements additionnels sur les mesures du budget*, mars 2006, à la p 61, en ligne : <www.budget.finances.gouv.qc.ca/budget/2006-2007/fr/pdf/RenseignementsAdd.pdf>.

850 « L'intelligence artificielle ne peut être vaincue : un maître de go abandonne », *Le Point* (27 novembre 2019), en ligne : <www.lepoint.fr/high-tech-internet/l-intelligence-artificielle-ne-peut-etre-vaincue-un-maitre-de-go-abandonne-27-11-2019-2350129_47.php>.

851 Will Knight, « AlphaGo Zero Shows Machines Can Become Superhuman Without Any Help », *MIT Technology Review* (18 octobre 2017), en ligne : <www.technologyreview.com/2017/10/18/148511/alphago-zero-shows-machines-can-become-superhuman-without-any-help/>.

852 Florence Sara G. Ferraris, « Les voitures autonomes, une révolution qui pourrait sauver des vies », *Le Devoir* (9 janvier 2017), en ligne : <www.ledevoir.com/societe/transports-urbanisme/488715/voitures-autonomes-une-revolution-qui-pourrait-sauver-des-vies>.

853 Joe D'Allegro, « How Google's Self-Driving Car Will Change Everything », *Investopedia* (20 décembre 2020), en ligne : <www.investopedia.com/articles/investing/052014/how-googles-selfdriving-car-will-change-everything.asp>.

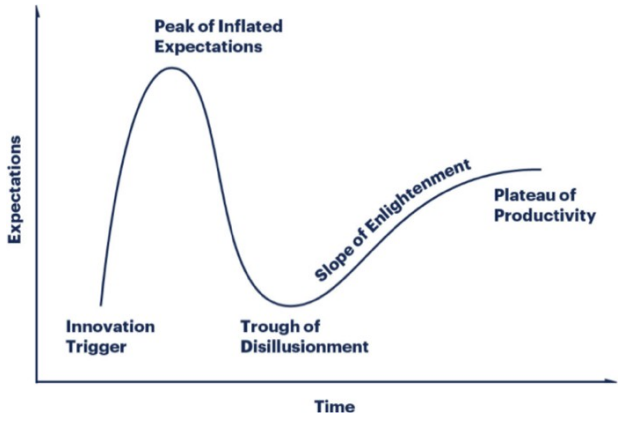
854 Pierre-Benoît Joly, « Le régime des promesses technoscientifiques » dans Marc Audétat, *Sciences et technologies émergentes : Pourquoi tant de promesses*, Paris, Hermann, 2015, 31, en ligne : <www.researchgate.net/publication/297622208_Le_regime_des_promesses_technoscientifique>.

855 Luc Julia, *supra* note 2.

développement des voitures autonomes soulève des difficultés techniques liées notamment à l'entraînement des algorithmes aux interactions sociales complexes et statistiquement peu fréquentes[856] ainsi que des dilemmes éthiques en situation de collision imminente[857].

Cette deuxième phase de découragement ou de désillusion sera éventuellement surmontée avec la maturation des nouvelles technologies qui atteindront, à terme, un plateau de productivité :

Figure 33
Cycle de Gartner décrivant l'évolution typique de l'intérêt éprouvé pour les nouvelles technologies



Source : Gartner Hype Cycle

Outre l'intelligence artificielle, la génomique, la fusion nucléaire et la nanotechnologie[858] figurent parmi les exemples récents d'une dynamique suivant le cycle d'excitation décrit par Gartner, une firme de consultants d'envergure internationale (Gartner Hype Cycle). Le développement de l'intelligence artificielle n'y échapperait guère après deux hivers qui ont freiné, au cours des années 1970 et 1980, tant les investissements que l'engouement publics face à ses promesses technoscientifiques (renvoi au Chapitre 1; voir aussi Leith, 2016[859]). À l'approche d'un « troisième hiver » imminent du fait des limites de l'approche quantitative appliquée[860] et de notre désillusion des grands nombres[861], une considération raisonnée des limites méthodologiques

---

856 Riti Dass, « 5 Key Challenges faced by Self-driving cars », *Medium* (14 septembre 2018), en ligne : <medium.com/@ritidass29/5-key-challenges-faced-by-self-driving-cars-ed04e969301e>.

857 Amy Maxmen, « Self-driving car dilemmas reveal that moral choices are not universal » (2018) 562 *Nature* 469, doi : <doi.org/10.1038/d41586-018-07135-0>.

858 Maxime Colleret et Yves Gingras, *supra* note 329.

859 Philip Leith, « The rise and fall of the legal expert system » (2016) 30:3 *Int Rev Law, Comp & Tech* 94, doi : <doi.org/10.1080/13600869.2016.1232465>.

860 Gary Marcus, *Deep Learning : A Critical Appraisal*, 2018, en ligne : <arxiv.org/pdf/1801.00631.pdf>.

861 Alain Supiot, *La Gouvernance par les nombres. Cours au Collège de France 2012-2015*, Fayard, 2015, en ligne : <www.college-de-france.fr/site/alain-supiot/La-gouvernance-par-les-nombres-film.htm>.

[renvoi au Chapitre 2] et des coûts sociaux liés au développement de l'intelligence artificielle (IA) est de mise, tel qu'il sera plus amplement traité au volume 2 de notre rapport.

Pour en savoir plus :

Accenture et CIFAR, *Pan-Canadian AI Strategy Impact Assessment Report*, octobre 2020, en ligne : <cifar.ca/wp-content/uploads/2020/11/Pan-Canadian-AI-Strategy-Impact-Assessment-Report.pdf>

Bradley, C., R. Wingfield et M. Metzger, *National Artificial Intelligence Strategies and Human Rights : A Review*, Global Partners Digital, Stanford, avril 2020, en ligne : <www.gp-digital.org/wp-content/uploads/2020/04/National-Artifical-Intelligence-Strategies-and-Human-Rights%E2%80%94A-Review_.pdf>

Loucks, J. et al., *Future in the balance ? How countries are pursuing an AI advantage. Insights from Deloitte's State of AI in the Enterprise, 2nd Edition survey*, Deloitte Insights, 1er mai 2019, en ligne : <www2.deloitte.com/us/en/insights/focus/cognitive-technologies/ai-investment-by-country.html>

U.S. National Security Commission on Artificial Intelligence, *Final Report*, 2021, en ligne : <www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf>

Conclusion

L'intelligence artificielle (IA) regroupe de nos jours un ensemble de méthodes et de techniques permettant d'optimiser le traitement de l'information par la machine et son interaction avec l'environnement. Son intérêt réside dans les applications novatrices auxquelles conduit cette extraction intelligente de savoir(-faire) à partir de données brutes. Le caractère révolutionnaire de ces applications dans des domaines aussi divers que l'agroalimentaire, l'astrophysique, la médecine, le marketing, la bourse et les transports, fascine les chercheurs et nourrit l'imaginaire populaire de promesses dignes de science-fiction. Comme nous avons vu, l'engouement atteint jusqu'aux plus hautes sphères (inter-)étatiques qui ne lésinent pas en subventions (salariales) aux entreprises et institutions de recherche, incitatifs fiscaux et politiques d'immigration favorables à l'attraction des talents mondiaux en IA. Forts des promesses technoscientifiques dont il semble difficile de fixer une date d'échéance, les investissements – tant publics que privés – ne misent pas tant sur une reddition de compte orientée vers l'atteinte des résultats qu'une concurrence féroce sur le quantum des montants investis en recherche et développement (R&D). Sans doute l'imprévisibilité, à horizon rapproché, de plusieurs des avancées techniques entretient cette « hystérie » collective vers une IA forte qui non seulement simulera en tout point l'intelligence humaine, mais aussi et surtout la dépassera en acuité, rapidité et efficacité. En attendant, des États au grand public, le vecteur d'innovation est propulsé par un imaginaire collectif que façonnent les différents traitements médiatiques enthousiastes de l'IA et la science-fiction. Contrebalançant cet emportement contagieux, il convient plutôt d'approfondir nos réflexions sur les enjeux sociétaux que posent cette discipline émergente et notre écosystème de l'IA [renvoi au volume 2 du rapport].

CONCLUSION GÉNÉRALE

Le présent document de travail n°1 sur l'épistémologie de l'intelligence artificielle ou augmentée (IA), se veut un tour d'horizon sur l'origine, le développement et l'avenir prévisible de l'intelligence artificielle ou augmentée (IA), comme discipline émergente, technologie, outil d'innovation et sujet d'intérêt politique.

Nous fondant sur le triptyque de l'épistémologie sur « le statut, la méthode et la valeur de la connaissance »[862], trois questions fondamentales ont été soulevées au regard :

- de l'objet de l'IA comme discipline (le « quoi »);
- de sa méthode (le « comment »); et
- de la valeur que l'IA représente pour notre société (le « pourquoi »).

Notre **chapitre 1** porte un regard d'abord tourné vers le passé, retraçant les origines de l'intelligence artificielle (IA) comme discipline émergente dont l'odyssée chevauche avec plusieurs des disciplines connexes et dont le développement se trouvera teinté par les progrès de ces dernières. À la philosophie (de l'esprit) l'IA a emprunté les méthodes de raisonnement éprouvées en logique formelle pour attester de la chaîne de vérité de diverses propositions. Des sciences biologiques et du cerveau, l'IA s'est inspirée pour reproduire, à l'échelle algorithmique et dans la mesure du possible, la manière dont des organismes biologiques traitent et communiquent l'information pour résoudre des problèmes complexes et optimiser leurs chances d'adaptation à l'environnement. Par sa plus grande polyvalence et sophistication, le fonctionnement du cerveau humain s'est avéré une référence de premier plan. Aux théories sur la linguistique, l'IA a prêté une attention soutenue pour raffiner l'articulation et le partage des connaissances humain-machine. De l'arithmétique simple aux modélisations mathématiques, le « langage » mathématique aide à reconstituer les divers états et opérations mentaux sous-jacents à l'analyse des phénomènes complexes (probabilités et statistiques) et coordonner les interactions avec d'autres agents dans l'environnement (*cf.* théorie des jeux et dilemme du prisonnier). À l'ingénierie (informatique) enfin, l'IA est redevable de l'implémentation sur la machine des différentes idées, théories, modalités de raisonnement et modélisations tirées des autres disciplines scientifiques. Sur un plan sociologique, il convient également de souligner que le rêve d'une IA est aussi profondément ancré dans nos cultures et trouve ses racines dans l'une de nos aspirations les plus profondes – rapportées notamment dans les mythes du progrès, légendes salvatrices et récits de science-fiction – de nous élever au-dessus de l'humaine condition, de transcender sa finitude et de vaincre son destin jusqu'à devenir l'égal de son créateur.

À un horizon plus rapproché, l'histoire de l'IA s'est tissée dans le prolongement de deux courants qui tantôt se divisent, tantôt se rejoignent. Dans un premier temps, l'approche symbolique a

862 Jean-Louis Le Moigne, *supra* note 11.

longtemps guidé la recherche de l'IA en postulant le raisonnement humain comme autant d'étapes discrètes marquées par la représentation de concepts abstraits réductibles aux symboles. Des premiers balbutiements de l'informatique aux systèmes experts, l'IA symbolique a participé à la maturation de la discipline en aidant à résoudre, suivant des données fournies et règles prédéfinies, des problèmes typiques du monde réel, tels que la pose de diagnostics selon l'état des connaissances scientifiques de l'époque. Il est rapidement apparu, cependant, que l'intelligence ne consiste pas uniquement en la manipulation de symboles sous la forme d'objets et d'opérateurs mathématiques discrets. De même que l'expérience de la conscience n'est que l'iceberg de processus neurologiques se jouant en sourdine de notre cognition, que la complexité et la richesse du monde réel ne soient pas réductibles aux modèles mathématiques abstraits, que la connaissance objective ne semble pas subsister indépendamment de la perception de ses sujets, l'intelligence aussi s'avère être un phénomène émergeant de la manière dont ses différentes unités sont disposées et d'un traitement intégré et continu – plutôt que successif – de l'information. L'approche connexionniste de l'IA – qui nous rappelle la grammaire générative de Chomsky, le constructivisme social de la sociologie contemporaine ou le structuralisme de Lévi-Strauss – table plutôt sur l'interconnexion des réseaux de neurones simulant le cerveau humain pour faire émerger une connaissance expérientielle qui s'alimente non seulement de règles symboliques, mais également du contexte situationnel ainsi que de l'écosystème dans et avec lequel le système artificiel (co)-évolue. En raison principalement des limites liées au développement de l'infrastructure informatique, l'approche connexionniste, dont la mise en œuvre nécessite une puissance de calcul très élevée ainsi que la disponibilité d'une grande quantité de données, a longtemps été reléguée à l'arrière-plan. La difficulté de l'implémenter, conjuguée aux doutes entretenus quant au potentiel de l'approche symbolique, a marqué deux hivers dans le développement de l'IA au courant de la décennie 1970 puis de la fin des années 1980 au milieu des années 1990. À l'orée des années 2010, le connexionnisme a trouvé enfin un (re)gain de popularité dans l'apprentissage profond automatisant non seulement le processus du raisonnement, mais également l'apprentissage de règles adéquates pour arriver aux résultats attendus. Des premiers systèmes experts aux systèmes d'apprentissage automatique, des auteurs n'hésitent pas à comparer cette évolution comme passant de l'artisanat à l'industrialisation pour automatiser la production de connaissances par la machine. Aujourd'hui, les deux approches – symbolique et connexionniste – tendent à se rejoindre dans l'implémentation d'algorithmes hybrides pour en optimiser la performance.

Depuis la conférence symptomatique de Dartmouth de l'été 1956 où l'expression « intelligence artificielle » (IA) est apparue pour la première fois, plusieurs définitions de l'IA ont été suggérées par les secteurs politiques et institutionnels, de la recherche et de l'industrie. Dans le cadre de nos travaux et sous réserve de nouvelles percées dans l'évolution de la technologie, nous adoptons la définition suivante de l'IA telle que proposée par le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle (GEHN IA) de la Commission européenne :

> Les systèmes d'intelligence artificielle (IA) sont des **systèmes logiciels** (et **éventuellement matériels**) conçus **par des êtres humains** et qui, ayant **reçu un objectif** complexe, agissent dans le **monde réel ou numérique** en **percevant leur environnement** par **l'acquisition de données**, en interprétant les données structurées ou non structurées collectées, en

> appliquant un **raisonnement aux connaissances**, ou en **traitant les informations**, dérivées de ces données et en décidant de **la/des meilleure(s) action(s) à prendre pour atteindre l'objectif donné**. Les systèmes d'IA peuvent soit utiliser des **règles symboliques, soit apprendre un modèle numérique**. Ils peuvent également **adapter leur comportement** en analysant la manière dont l'environnement est affecté par leurs actions antérieure.[863] [Nos soulignements]

Quoique le Groupe d'experts emploie l'expression « intelligence artificielle », nous nous alignons avec la suggestion de l'ingénieur Luc Julia[864] pour y préférer l'expression « intelligence augmentée ». L'« intelligence artificielle » est plus une métaphore, à connotations hyperboliques tendant à alimenter une vision surréaliste d'une super-intelligence (forte) qu'elle ne traduit ce que les systèmes d'IA sont réellement. Il s'agit avant tout de programmes informatiques qui, loin de reproduire ou de surpasser, dans toute sa versatilité, l'intelligence telle que nous la retrouvons dans la nature, permet d'augmenter, de complémenter notre intelligence en optimisant (en temps, en précisions et en possibilités) ce que les humains font.

Le **chapitre 2** plonge au cœur des mécanismes sous-jacents au traitement de l'information, à la résolution des problèmes et à l'atteinte d'objectifs prédéfinis par les différents systèmes d'intelligence augmentée (IA). Une description de haut niveau des principales techniques et modèles algorithmiques utilisés a été présentée, complétée par quelques pistes de réflexion générales sur les défis qui se posent actuellement sur la voie vers l'intelligence artificielle forte.

Les premiers systèmes d'IA dits symboliques, conçus dans les années 1950, raisonnent à partir de symboles (p.ex. objets, images, chiffres) et des différentes règles d'opération encodées dans l'ordinateur. Dans un premier temps, les chercheurs étaient optimistes quant à la possibilité de simuler l'ensemble de nos facultés cognitives par le biais d'un raisonnement opératoire de cause à effet reliant les données d'entrée aux résultats sortants, consistant strictement en la manipulation de représentations symboliques abstraites du monde réel. Point n'est nécessaire pour la machine de comprendre la signification des symboles pourvu que les règles d'opération en dictent une relation inférentielle nécessaire pour arriver aux résultats logiques attendus. Ce qui a l'avantage d'être clair, facile d'application et de garantir un processus d'analyse transparent. Parmi les premières applications de l'approche symbolique, citons le *Logic Theorist* présenté lors de la conférence de Darmouth de 1956 – le tout premier programme qui a donné l'impulsion à la discipline, le programme informatique développé en 1959 par Herbert et ses collaborateurs qui a été surnommé éloquemment le « *General Problem Solver* », ou encore le tout premier agent conversationnel ELIZA conçu en 1966. Durant les années 1970, l'engouement pour les systèmes experts a pris le relais des premiers systèmes symboliques en permettant l'incorporation de connaissances scientifiques susceptibles de résoudre des problèmes de la vie réelle. Il s'agit d'alimenter la machine avec les connaissances scientifiques acquises dans un

---

863 Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle (GEHN IA), *Lignes directrices en matière d'éthique pour une IA digne de confiance*, Commission européenne, 2018 au para 143, en ligne : <ec.europa.eu/newsroom/dae/document.cfm?doc_id=60427>.

864 Luc Julia, *supra* note 2.

domaine donné puis de définir les règles d'inférence nécessaires du type « si …donc » pour relier les différentes propositions et arriver à un résultat donné, par exemple pour poser un diagnostic d'après les symptômes constatés chez un patient (p.ex. MYCIN), identifier les différentes moléculaires à partir de l'analyse des propriétés obtenues par la spectrométrie de masse (p.ex. DENDRAL), apparier les différentes commandes avec les exigences du client (p.ex. XCON ou eXpert CONfigurer), assister les parties dans l'élaboration d'arguments juridiques d'après les règles applicables (p.ex. HYPO, JusticeBot).

Des premiers systèmes symboliques aux systèmes experts, l'approche symbolique suppose que l'ensemble des connaissances puisse être explicitée par le biais de relations simples et formalisables, ce qui n'est pas toujours le cas. Au-delà de la démonstration de théorèmes mathématiques ou des problèmes logiques « simples », la complexité du monde réel n'est pas réductible aux représentations symboliques. Comment formaliser le processus d'apprentissage du vélo, la reconnaissance d'images ou l'interprétation différente des mêmes mots employés dans des contextes différents ? De plus, les systèmes experts s'avèrent d'une maintenance ardue, notamment par la nécessité de mettre à jour, manuellement, minutieusement et laborieusement, les connaissances scientifiques acquises d'après la plus récente revue de littérature qui peut comporter plusieurs milliers de pages de publications. Un manque d'adaptabilité a également été observé chez les systèmes experts qui suivent scrupuleusement les règles prédéfinies mais ne sont pas en mesure de généraliser au-delà ou de traiter des problèmes nouveaux qui ne relèvent pas des règles prédéfinies. Ces limitations ont donné l'impulsion vers le développement de systèmes dits sub-symboliques reposant sur l'approche connexionniste. À la différence de systèmes (experts) symboliques, l'apprentissage automatique passe du calcul formel aux modélisations mathématiques pour approximer certaines tendances, schémas du monde réel et utiliser ces modélisations pour prédire les évènements futurs avec un degré acceptable de précision, qu'il s'agisse des préférences clients, de l'évolution du cours d'une action en bourse, voire des prochaines tendances électorales. Au lieu de définir – comme dans les systèmes symboliques – les règles d'opération ou d'inférence que l'algorithme doit suivre pour relier les données d'entrée à l'étiquette résultante, l'approche connexionniste compte justement sur la configuration globale d'un système pour faire émerger la connaissance implicite nécessaire à la saisie de relations subtiles qui ne sont pas toujours formalisables à l'aide de règles explicites.

Trois grandes familles de techniques algorithmiques sont communément employées pour cette modélisation : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. L'apprentissage supervisé consiste à entraîner l'algorithme à reconnaître, par lui-même, l'association entre les données d'entrée étiquetées et le résultat attendu, pour éventuellement assortir cette même étiquette à de nouvelles données du monde réel. Toutefois, la collecte, la sélection et l'étiquetage de données d'entrée pour les rendre utilisables par l'algorithme prennent du temps et peuvent s'avérer ardus. L'apprentissage non supervisé laisse encore plus de flexibilité à l'algorithme pour repérer par lui-même les schémas et régularités implicites dans les données non structurées et non étiquetées. L'apprentissage par renforcement consiste à apprendre à la machine à effectuer certaines tâches dans un environnement donné, comme la conduite autonome ou la maîtrise de jeux stratégiques. L'état initial de la machine peut

être une vue de caméra, la position initiale des pièces sur un échiquier ou le début d'une quête de jeu vidéo. À partir de là, la machine apprend à interagir avec son environnement, par un processus d'essais et erreurs orienté vers la maximisation de son score de récompenses pour avoir « deviné » correctement. Entre les techniques d'apprentissage supervisé, non supervisé et par renforcement, le choix ne dépend pas des préférences idiosyncratiques des développeurs, mais de la nature des tâches à accomplir.

Souvent considéré comme une sous-catégorie de l'apprentissage automatique, l'apprentissage profond se démarque des autres techniques par une autonomie encore plus grande de l'algorithme pour traiter les données complexes et non structurées de la vie réelle. Qu'il soit supervisé, non supervisé ou par renforcement, l'apprentissage profond fait appel à une architecture novatrice de réseaux de neurones multicouches pour approximer, étape par étape, couche par couche, la complexité des fonctions (relations) non linéaires entre différents variables et paramètres. À partir de cette architecture générale, des réseaux de neurones plus spécialisés ont été conçus pour optimiser le traitement de certaines tâches, dont :

- les réseaux de neurones convolutifs (CNN : *Convolutional Neural Networks*) qui excellent notamment dans diverses tâches de reconnaissance vocale et d'images;

- les réseaux de neurones récurrents (RNN : *Recurrent Neural Networks*) qui permettent un traitement itératif des données donnant naissance à une sorte de mémoire de travail optimisant par exemple l'apprentissage des différents contextes antérieurs dans lesquels un même mot est apparu et son sens interprété;

- les réseaux antagonistes ou adverses génératifs qui reposent sur l'entraînement parallèle de deux réseaux de neurones dans un scénario de théorie des jeux à somme nulle : l'un des réseaux doit apprendre à générer des images réalistes alors que l'autre réseau a pour tâche de discriminer les images réelles de celles générées artificiellement; cette compétition a pour but d'augmenter la qualité (le degré de réalisme) des images créées par le réseau générateur;

- l'apprentissage profond par renforcement qui permet l'apprentissage (par renforcement) à partir des données non structurées.

Malgré des résultats prometteurs surpassant parfois la performance d'experts humains, l'apprentissage profond connaît certaines limitations liées premièrement à l'opacité des processus sous-jacents reliant les données d'entrée aux résultats obtenus, ce qui risque de miner la confiance envers ces systèmes. Les réseaux profonds peuvent également rencontrer des problèmes d'alignement entre, d'une part, les objectifs des développeurs et, d'autre part, les données et paramètres effectivement pris en compte par les algorithmes pour arriver aux solutions, ce qui peut conduire à des problèmes d'inefficiences ou encore de biais algorithmiques préjudiciables à la société.

Des premiers systèmes symboliques à l'apprentissage profond, la voie est-elle pavée pour une intelligence forte dotée de sens commun, d'états mentaux subjectifs et d'une conscience ? Pour y parvenir, une approche axée sur l'apprentissage des tâches discrètes se doit d'être complétée

par une vision plus large orientée vers une plus grande adaptabilité des capacités d'apprentissage et de prédictions ainsi que le transfert transversal des acquis d'un domaine d'applications à l'autre. Outre le raffinement et la sophistication de l'architecture sous-jacente à la cognition artificielle, il y a lieu de compléter l'approche connexionniste axée sur le développement des réseaux de neurones avec d'autres approches « expérientielles » insistant sur l'importance de faire interagir la machine avec l'environnement (p.ex. la robotique développementale) pour mieux assimiler le contexte au sens large et faciliter l'acquisition de connaissances implicites difficilement formalisables. À l'instar des stades du développement humain de l'enfance à l'âge adulte, il s'agit de développer une machine à part entière qui ne se limite pas à l'arithmétique simple ou aux modélisations mathématiques abstraites. Une intelligence artificielle forte, qui se caractérise par sa polyvalence et la capacité de réfléchir sur sa propre cognition, est tout autant un produit de l'environnement que de son architecture. Une machine doit « s'incarner » dans son environnement, interagir avec d'autres agents pour pouvoir un jour, au détour d'un plan d'eau, porter un regard rétrospectif sur son propre reflet et son expérience en même temps que de s'émerveiller devant le monde qui l'entoure, lequel pourrait bien être le reflet d'une autre réalité, plus vraie, celle d'en haut de la caverne.

Le **chapitre 3** s'intéresse au développement de l'IA qui a été notamment propulsé par l'engouement des États et des institutions politiques régionales dans la course à l'innovation technologique. Parallèlement à l'intérêt éprouvé par les secteurs de la recherche, de l'enseignement supérieur et de l'industrie pour cette discipline émergente, la pertinence d'une intervention publique pour soutenir les efforts d'innovation a été soulignée par la théorie économique. Ce souhait a été exaucé. Du commerce électronique à l'économie d'Internet, les initiatives multilatérales (p.ex. OCDE, le G7, l'UNESCO) se sont succédé au rôle névralgique des États-Unis qui, par le biais du *Defense Advanced Research Projects Agency* (DARPA), ont été les bailleurs de fonds principaux, à l'échelle mondiale, des infrastructures technologiques tissant un cyberespace sans frontières. Les autres grandes puissances ne sont pas en reste, dont les Communautés européennes puis l'Union européenne, la Chine et le Japon. D'un peu partout dans le monde se multiplient plans d'action, stratégies et énoncés de position sur « la croissance ou l'économie numérique », « les technologies numériques », « l'innovation » et plus récemment « l'intelligence artificielle ». Au Canada et au Québec en particulier, la politique de l'IA s'inscrit dans le prolongement de priorités sectorielles touchant l'informatique et les biotechnologies, pour faire arrimer la recherche scientifique avec les impératifs du marché et les besoins de la société. Elle est mise en œuvre par le biais d'incitatifs fiscaux en recherche et développement (R&D), de subventions directes à la recherche, à la formation et à l'emploi ainsi que des politiques favorables à l'attraction de talents étrangers.

Ici comme ailleurs, les politiques publiques sur l'IA et l'innovation contribuent à alimenter un climat plein d'effervescence qui tend à la fois à surestimer le bénéfice social dont on peut en tirer à long terme qu'à sous-évaluer le montant d'investissements massifs nécessaires pour tirer parti d'une course à l'hégémonie politique que se disputent les grandes puissances tant pour le pouvoir de coercition de l'IA (p.ex. défense nationale) que son pouvoir de convaincre (*soft power*) avec la ramification de ses implications sociales, économiques et politiques. Il convient toutefois de remettre en perspective ce cycle d'engouements technoscientifiques pour l'IA qui n'en est

qu'un parmi d'autres, alimentés par autant de promesses suivies d'une phase de découragement ou de désillusion avant d'atteindre la phase de plateau avec la maturation d'une nouvelle technologie.

***

Dans notre introduction générale, nous avons comparé l'intelligence artificielle (IA) au Prométhée moderne doté d'un potentiel immense. Immense, mais aussi troublant. Le Prométhée moderne, c'est par ailleurs Frankenstein rappelant les risques d'une poursuite immodérée d'un savoir technique spécialisé aux conséquences pas toujours prévisibles. L'on pourrait penser d'emblée aux armes létales autonomes dont le déploiement dans le cadre de conflits armés nous fait craindre des scénarios apocalyptiques dignes de science-fiction. Pourtant, le Prométhée moderne – comme il sied à une expression métaphorique – n'a pas (nécessairement) l'allure d'homme. Plus que l'assaut caricatural des robots tueurs partant à la (re)conquête de l'humanité, c'est aussi et surtout cette toile invisible que tissent les algorithmes autour de nous à laquelle il nous conviendrait de nous attarder.

En effet, les applications transversales de cette discipline émergente qu'est l'IA nous donneraient presque l'envie de nourrir les algorithmes de toutes sortes de données qu'il soit possible de collecter, ou d'appliquer aux données disponibles toutes sortes de recettes algorithmiques, pour en expérimenter les résultats.

À n'en pas douter, l'ambigüité se rattachant au concept de l'IA n'est pas que sémantique. L'incertitude se situe aussi au plan de ce que cette cognition augmentée pourrait apporter dans ce nouvel âge « augmenté » de l'information. Quels seraient les impacts sociétaux, politiques et juridiques de cette « explosion d'intelligence » ou « cognition augmentée » ?

Épée à double tranchant, récits en miroir, quête du Graal ou boîte de Pandore, l'intelligence artificielle ou augmentée (AI) compte autant d'adeptes que de détracteurs. Certains la voient d'un œil expectatif, comme annonçant une ère sans précédent de collaboration humain-robot : tandis que le cerveau artificiel va en se complexifiant avec le cerveau humain et participe de manière toujours plus intégrée à l'évolution de nos civilisations, une fusion de la technologie avec la nature et l'humanité promet un avenir radieux, plein d'adaptabilité, de complexité et de beauté. D'autres, plus sceptiques, craignent un écart toujours plus grand entre l'évolution du cerveau artificiel et celle du cerveau humain, jusqu'au jour où les mécanismes irrésistibles de la sélection « naturelle » achèvent de destituer l'humain de sa place de conquérant de la nature. Plutôt que la malveillance spontanée d'un *Terminator* déclarant ouvertement la guerre aux créateurs qui l'ont précédé, le danger peut résider dans une conquête plus subtile, par assimilation, d'un prédateur qui se fond dans le paysage et qui, se jouant du clair-obscur environnant, laisse sa proie venir à lui, librement et de plein gré, jusqu'à un point de non-retour.

Notre document de travail n° 2 sur l'épistémologie de l'intelligence artificielle traitera, par collection d'articles et par thèmes, des enjeux tant à court qu'à long terme que pose l'intelligence artificielle ou augmentée (AI) sur notre société à l'ère de l'information.

BIBLIOGRAPHIE

Chapitre 1

Monographies et ouvrages collectifs

Buchanan, B. G. et E. H. Shortliffe (dir.), *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project, coll. The Addison-Wesley series in artificial intelligence, Reading, Mass, Addison-Wesley, 1984.*

Englewood Cliffs (N.J), Prentice Hall, 1995.

Firdaus, M., S. E. Pratiwi, D. Kowanda et A. Kowanda, « Literature review on Artificial Neural Networks Techniques Application for Stock Market Prediction and as Decision Support Tools », dans *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018.

Gardner, H.*, Frames of mind: the theory of multiple intelligences*, Basic Books éd, New York.

Haugeland, J., *Artificial Intelligence: The Very Idea*, Cambridge: MIT Press, 1985.

Hubert, D., *What Computers Can't Do, A Critique of Artificial Reason*, Harper & Row, 1972.

Julia, L. et K. Ondine, *L'intelligence artificielle n'existe pas*, First Forum éd, 2019.

LeCun,Y., *Quand la machine apprend*, Odile Jacob éd, Paris, 2019

Lina, L. T. Ming et L. S. Kar, « A Hybrid Connectionist-Symbolic Approach for Real-Valued Pattern Classification », dans Daoliang Li et Baoji Wang (dir.), *Artificial Intelligence Applications and Innovations*, coll. IFIP, Boston, MA, Springer US, 2005.

Russell, Stuart J. Rus et P. Norvig, *Artificial intelligence: a modern approach*, Englewood Cliffs (N.J), Prentice Hall, 1995.

Shi-Nash, A et D. Hardoon, *Data Analytics and Predictive Analytics in the Era of Big Data*, Internet of Things and Data Analytics Handbooks, John Wiley & Sons, 2017.

Articles de revues

Abiodun, O. I. et al., « Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition », (2019) 7 *IEEE Access*.

Cardon, D., J-P. Cointet et A. Mazières, « La revanche des neurones: l'invention des machines inductives et la controverse de l'intelligence artificielle » (2018) 5:21

Domingos, P., « A few useful things to know about machine learning » (2012) 55:10 *Commun ACM* 78.

Dumouchel, P., « Intelligence, Artificial and Otherwise » (2019) 24:2 *Forum Phisosophicum* 241.

Feigenbaum, E. A., « Knowledge Engineering.: The Applied Side of Artificial Intelligence », (1984) 426:1 Computer Cult *Ann NY Acad Sci* 91.

Grekousis, G., « Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis », (2019) 74 *Computers, Environment and Urban Systems* 244.

Honavar, V. et L. Uhr, « Symbolic Artificial Intelligence, Connectionist Networks & Beyond. », (1994) 76 *Computer Science Technical Reports* 45.

Kaplan, A., M., Haenlein, « Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence » (2019) 62:1 *Business Horizon* 15.

Kitchin, R., « Thinking critically about and researching algorithms » (2016) 20 *Information, Communication & Society* 1-16.

Lee, I., « Big data: Dimensions, evolution, impacts, and challenges » (2017) 60:3 *Business Horizons* 293.

Metaxiotis, K. et J. Psarras, « Expert systems in business: applications and future directions for the operations researcher », (2003) 103-5 *Industrial Management & Data Systems* 361.

Rosenblatt, F., « The Perceptron : A probabilistic Model for Information Storage and Organization in the Brain », (1958) 65-6 *Psychological Review*, en ligne : <http://homepages.math.uic.edu/~lreyzin/papers/rosenblatt58.pdf>.

Rumelhart, D., G. Hinton, et R. J. Williams, « Learning representations by back-propagating errors », (1986) 323 *Nature*, en ligne : <https://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf>.

Stefik, M. et al., « The organization of expert systems, a tutorial », (1982) 18-2 *Artificial Intelligence* 135.

Turing, A., « Computing Machinery and Intelligence » (1950) LIX:236 Mind 433, doi: <doi.org/10.1093/mind/LIX.236.433>

Vasant, H., L., Uhr, « Symbolic Artificial Intelligence, Connectionist Networks & Beyond. » (1994) 76 *Computer Science Technical Reports* 45.

Articles de journaux

Dave Lee, « Tay: Microsoft issues apology over racist chatbot fiasco », *BBC News* (25 mars 2016), en ligne : <https://www.bbc.com/news/technology-35902104>.

« Electronic "Brain" Teaches Itself », *The New York Times*, en ligne : <http://timesmachine.nytimes.com/timesmachine/1958/07/13/91396361.html>.

« John McCarthy -- Father of AI and Lisp -- Dies at 84 », *Wired* (24 octobre 2011), en ligne : <https://www.wired.com/2011/10/john-mccarthy-father-of-ai-and-lisp-dies-at-84/>.

Naveen Joshi, « 7 Types Of Artificial Intelligence » *Forbes* (19 juin 2019), en ligne : <https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/>.

Documents de travail

Colleret, M. et Y. Gingras, *L'intelligence artificielle au Québec : un réseau tricoté serré*, note de recherche 2020-07, Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal (UQÀM), 2020, en ligne : <cirst2.openum.ca/files/sites/179/2020/12/Note_2020-07_IA.pdf>

Future of Privacy Forum, *The Privacy Expert's Guide to Artificial Intelligence and Machine Learning*, 2018, en ligne : <https://iapp.org/media/pdf/resource_center/FPF_Artificial_Intelligence_Digital.pdf>.

Garnelo, M., K. Arulkumaran et M. Shanahan, « Towards Deep Symbolic Reinforcement Learning », 2016, en ligne : <http://arxiv.org/abs/1609.05518>.

Joint Technology Committee, *Introduction to AI for Courts*, JTC Resource Bulletin version 1.0, 2020, en ligne : <https://www.ncsc.org/__data/assets/pdf_file/0013/20830/2020-04-02-intro-to-ai-for-courts_final.pdf>.

Lighthill, J., *Artificial Intelligence: A General Survey*, 1972, en ligne : <http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm>.

McCarthy, J. et al., A Proposal for the Darmouth Summer Research Project on Artificial Intelligence, Darmouth, 1995.

Zhu, J., *Glossaire en intelligence artificielle*, 2020, en ligne : <https://cyberjustice.openum.ca/glossaire-ia-laboratoire-de-cyberjustice-2020/>.

Sources électroniques

Allen, D., M. West and R. John, « How artificial intelligence is transforming the world », (24 avril 2018), en ligne : <https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/>.

Bastien, L., « Les quatre V du Big Data expliqués par IBM », (2016), en ligne : *LeBigData.fr* <https://www.lebigdata.fr/infographie-quatre-v-big-data-expliques-ibm>.

Conseil de l'Europe, « Histoire de l'intelligence artificielle », en ligne : <https://www.coe.int/fr/web/artificial-intelligence/history-of-ai>.

Emerj, « Emerj AI Opportunity Landscape Service », en ligne : <https://emerj.com/ai-opportunity-landscape-inquiry/>.

Faggella, D., « What is Artificial Intelligence? An Informed Definition », en ligne : <https://emerj.com/ai-glossary-terms/what-is-artificial-intelligence-an-informed-definition/>.

Gingras Y., « L'intelligence sociologique confrontée à l' « intelligence artificielle » *Ateliers SociologIA* (2019), en ligne : <https://www.cirst.uqam.ca/nouvelles/2019/ateliers-sociologia-conferences-disponibles-en-ligne/

Gingras, Y., « L'intelligence artificielle n'existe pas », (3 juin 2018), en ligne : *Radio-Canada* <http://ici.radio-canada.ca/premiere/emissions/les-annees-lumiere/segments/chronique/74731/science-critique-gingras-intelligence-artificielle-n-existe-pas>.

Grands Dossiers, « Intelligence : de quoi parle-t-on ? » *Sciences Humaines* (2007), en ligne : <https://www.scienceshumaines.com/intelligence-de-quoi-parle-t-on_fr_21032.html>.

Launchbury, J, « A DARPA Perspective on Artificial Intelligence » (2020) en ligne : <https://www.darpa.mil/attachments/AIFull.pdf>.

Lussier-Lejeune, F., « Le développement sociohistorique de l'intelligence artificielle sous l'angle des économies de la promesse » (9 décembre 2020), en ligne : <https://www.cyberjustice.ca/programme-virtuel/epistemologie-de-lia/>.

Narayanan, A., « How to recognize AI snake oil» *Princeton University* (2021), en ligne : <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>.

Oracle, « Dynamisez vos activités grâce à l'intelligence artificielle », en ligne : <https://www.oracle.com/ca-fr/artificial-intelligence/>.

Documents internationaux

Ad Hoc Expert Group for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence, *Document final: avant-projet de recommandation sur l'éthique de l'intelligence artificielle*, 2020.

CNIL, *Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, Synthèse du débat publique animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, 2017.

European Commission Joint Research Centre, AI Watch: Historical evolution of artificial intelligence: analysis of the three main paradigm shifts in AI, Publications Office of the European Union, 2020.

European Commission, AI watch: defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence, Publications Office of the European Union, 2020.

High-Level Expert Group on Artificial Intelligence, *A definition of AI: Main Capabilities and Disciplines*, Publications Office of the European Union, 2019.

High-Level Expert Group on Artificial Intelligence, *Shaping Europe's digital future, Publications Office of the European Union*, 2018, en ligne : <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>.

OCDE, Recommendation of the Council on Artificial Intelligence, 2019.

Dictionnaire et ouvrages généraux

LAROUSSE, « Intelligence artificielle », *Encyclopédie Larousse en ligne*, en ligne : <https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257>.

Office québecois de la langue française « Troll d'Internet », en ligne : *Fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26522696>.

Office québecois de la langue française, « Apprentissage profond », en ligne : *Fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26532876>.

Office québécois de la langue française, « mégadonnées », en ligne : fiche terminologique <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26507313>.

Office québecois de la langue française, « Robotique », en ligne : *Fiche terminologique* <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=2078479>.

UNIVERSALIS, « Norbert Wiener », *Encyclopædia Universalis*, en ligne : <https://www.universalis.fr/encyclopedie/norbert-wiener/>

Chapitre 2

Monographies et ouvrages collectifs

Clark, A., *Surfing Uncertainty : Prediction, Action, and the Embodied Mind*, Oxford Scholarship Online, 2016, DOI : <doi.org/10.1093/acprof:oso/9780190217013.001.0001>

Downer, A., Smart and Spineless : Exploring Invertebrate Intelligence, Twenty-First Century Books, 2015

Fleming, S.M. et C.D. Frith, dir, *The Cognitive Neuroscience of Metacognition*, Springer, 2014, doi : <doi.org/10.1007/978-3-642-45190-4>

Haugeland, J., L'esprit dans la machine. Fondements de l'intelligence artificielle, Odile Jacob, 1989

Hawkins, J. et S. Blakeslee, On Intelligence : How a New Understanding of the Brain will Leaad to the Creation of Truly Intelligent Machines, Times Books, 2004

Kurzweil, R., The Singularity is Near : When Humans Transcend Biology, Viking, 2005

Levi, P. et S. Kernbach, dir, *Symbiotic Multi-Robot Organism. Reliability, Adaptability, Evolution*, Springer, 2010, doi : <10.1007/978-3-642-11692-6>

Pagel, J.F., Dream Science : Exploring the Forms of Consciousness, Academic Press, 2014

Pagel, J.F. et P. Kirshtein, *Machine Dreaming and Consciousness*, Academic Press, Elsevier, 2017

Parker, L.E., F.E. Schneider et A.C. Schultz, *Multi-Robot Systems. From Swarms to Intelligent Automata. Volume III*, Proceedings from the 2005 International Workshop on Multi-Robot Systems, Springer, 2005, doi : <doi.org/10.1007/1-4020-3389-3>

Reeves, H., Patience dans l'azur. L'évolution cosmique, 2e éd, Seuil, 1988

Werbos, P.J., The Roots of Backpropagation : From Ordered Derivatives to Neural Networks and Political Forecasting, New York, John Wiley & Sons, 1994

Articles de revues et études d'ouvrages collectifs

AI Multiple, *995 experts opinion : AGI/singularity by 2060 [2021 update]*, 2 février 2021, en ligne : <research.aimultiple.com/artificial-general-intelligence-singularity-timing/>

Apley, D.W. et J. Zhu, « Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models » (2019), en ligne : <arxiv.org/abs/1612.08468>

Asada, M. et al., « Cognitive Developmental Robotics As a New Paradigm for the Disgn of Humanoid Robots » (2001) 37:2-3 Robotics and Autonomous System 185, doi : <doi.org/10.1016/S0921-8890(01)00157-9>

Asimov, I., « Runaround » (mars 1942) Astouding Science Fiction, en ligne : <web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html>

Ast, F., « A Short Literature Review on Collective Intelligence », *Medium* (13 septembre 2015), en ligne : <medium.com/astec/a-brief-literature-review-on-collective-intelligence-2b7f7e4f4561>

Battaglia, P.W. et al., « Relational inductive biases, deep learning, and graph networks » (2018), en ligne : <arxiv.org/abs/1806.01261>

Bau, D. et al., « GAN Dissection : Visualizing and Understanding Generative Adversarial Networks » (2018), en ligne : <arxiv.org/abs/1811.10597>

Bau, D. et al., « Network Dissection : Quantifying Interpretability of Deep Visual Representations » (2017) Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, en ligne : <arxiv.org/abs/1704.05796>

Bengio, Y. et al., « Towards Biologically Plausible Deep Learning » (2015), en ligne : <arxiv.org/pdf/1502.04156>

Betti, A., M. Gori et G. Marra, « Backpropagation and Biological Plausibility » (2018), en ligne : <arxiv.org/abs/1808.06934>

Bhatia, R., « Back-Propagation : Is It The Achilles Heel Of Today's AI », *Analytics India Magazine* (7 novembre 2017), en ligne : <analyticsindiamag.com/back-propagation-is-it-the-achilles-heel-of-todays-ai>

Blum, C., A.F.T. Winfield et V.V. Hafner, « Simulation-Based Internal Models for Safer Robots » (2018) Front Robot AI, doi : <doi.org/10.3389/frobt.2017.00074>

Breed, M.D. et J. Moore, « Chapter 6 – Cogntion » dans *Animal Behavior*, 2e éd, Elsevier, 2015, 175, doi : <doi.org/10.1016/C2013-0-14008-1>

Brooks, R.A., « A Robust Layered Control System For A Mobile Robot » (1986) 2:1 IEEE Journal of Robotics and Automation 14, doi : <doi.org/10.1109/JRA.1986.1087032>

Brown, C., « Fish intelligence, sentience and ethics » (2014) 18:1 Animal Cognition, doi : <doi.org/10.1007/s10071-014-0761-0>

Brown, T.B. et al, « Language Models are Few-Shot Learners » (2020), en ligne : <arxiv.org/abs/2005.14165>

Cadieu, C.F. et al., « Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition » (2014) 10 PLoS Comput Biol, e1003963, doi : <doi.org/10.1371/journal.pcbi.1003963>

Call, J. et M. Tomasello, « Does the chimpanzee have a theory of mind? 30 years later » (2008) 12:5 Trends in Cognitive Sciences 187, doi : <doi.org/10.1016/j.tics.2008.02.010>

Chatila, R. et al., « Toward Self-Aware Robots » (2018) Front Robot AI, doi : <doi.org/10.3389/frobt.2018.00088>

Chella, A. et al., « Developing Self-Awareness in Robots via Inner Speech » (2020) Front Robot AI, doi : <doi.org/10.3389/frobt.2020.00016>

Chella, A. et A. Pipitone, « A cognitive architecture for inner speech » (2020) 59 Cognitive Systems Research 287, doi : <doi.org/10.1016/j.cogsys.2019.09.010>

Chen, M. et al., « Generative Pretraining from Pixels » (2020), en ligne : <cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf>

Chen, Z.-M. et al., « Multi-Label Image Recognition with Graph Convolutional Networks » (2019) Computer Vision and Pattern Recognition, en ligne : <arxiv.org/abs/1904.03582>

Couzin, I.D., « Collective cognition in animal groups » (2008) 13:1 Trends in Cognitive Sciences 36, doi : <doi.org/10.1016/j.tics.2008.10.002>

Cruse, H. et M. Schilling, « Mental states as emergent properties. From walking to consciousness » dans Metzinger, T. et J. Windt, dir, *Open* Mind, Francfort, PUB, 2015, doi : <doi.org/10.15502/9783958570436>

Cusumano-Towner, M.F. et al., « Gen : a general-purpose probabilistic programming system with programmable inference » (2019) PLDI 221, doi : <doi.org/10.1145/3314221.3314642>

Cziko, G.A., « Unpredictability and Indeterminism in Human Behavior : Arguments and Implications for Educational Research » (1989) 18:3 Educational Researcher 17, doi : <doi.org/10.2307/1174887>

Dargazany, A.R., *Deep learning research landscape & roadmap in a nutshell : past, present and future – Towards deep cortical learning*, 7 août 2019, en ligne : <arxiv.org/pdf/1908.02130.pdf>

Davidson, J.E. et I.A. Kemp, « Contemporary Models of Intelligence » dans R.J. et S.B. Kaufman, dir, *The Cambridge Handbook of Intelligence*, Cambridge, University Press, 2011, 58, doi : <doi.org/10.1017/CBO9780511977244.005>

De Boeck, P. et al., « An Alternative View on the Measurement of Intelligence and Its History » dans Sternberg, R.J., *The Cambridge Handbook of Intelligence*, Cambridge University Press, 2020, 47, doi : <doi.org/10.1017/9781108770422.005>

Dosovitskiy, A. et al., « An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale » (2020), en ligne : <arxiv.org/abs/2010.11929>

Emery, N.J., « Cognitive ornithology : The evolution of avian intelligence » (2006) 361:1465 Philosophical Transactions of the Royal Society B Biological Sciences 23, doi : <doi.org/10.1098/rstb.2005.1736>

Feinberg, T.E. et J. Mallatt, « Phenomenal Consciousness and Emergence : Eliminating the Explanatory Gap » (2020) 11 Front Psychol, doi : <doi.org/10.3389/fpsyg.2020.01041>

Feinerman, O. et A. Korman, « Individual versus collective cognition in social insects » (2017) 220 J Exp Biol 73, doi : <doi.org/10.1242/jeb.143891>

Fesce, R., « Subjectivity as an Emergent Property of Information Processing by Neuronal Networks » (2020) Front Neurosci, doi : <doi.org/10.3389/fnins.2020.548071>

Friedman, J.H., « Greedy Function Approximation : A Gradient Boosting Machine » (2001) 29 The Annals of Statistics 1189, doi : <doi.org/10.1214/aos/1013203451>

Fukushima, K., « Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position » (1980) 36 Biological Cybernetics 193, doi : <doi.org/10.1007/BF00344251>

Gallup, G.G., Jr, « Chimpanzees : Self-Recognition » (1970) 167:3914 Science 86, doi : <doi.org/10.1126/science.167.3914.86>

Gallup, G.G., « Self recognition in primates : A comparative approach to the bidirectional properties of consciousness » (1977) 32:5 American Psychologist 329, doi : <doi.org/10.1037/0003-066X.32.5.329>

Gautam, A. et S. Mohan, « A review of research in multi-robot systems » dans *2012 IEEE 7th International Conference on Industrial and Information Systems (ICIIS)*, Chennai (Inde), 2012, 1, doi : <doi.org/10.1109/ICIInfS.2012.6304778>

Gibson, J.J., « The Theory of Affordances » dans Shaw, R.E. et J. Bransford, *Perceiving, acting, and knowing : toward an ecological psychology*, Hillsdale (NJ), Lawrence Erlbaum Associates, 1977, 67

Gibson, J.J., « The Theory of Affordances » dans *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979, 127

Good, I.J., « Speculations Concerning the First Ultraintelligent Machine » (1966) 6 Advances in Computers 31, doi : <doi.org/10.1016/S0065-2458(08)60418-0>

Guevara, R., D.M. Mateos et J.L.P. Velázquez, « Consciousness as an Emergent Phenomenon : A Tale of Different Levels of Description » (2020) 22:9 Entropy (Basel) 921, doi : <doi.org/10.3390/e22090921>

Hall, P., S. Ambati et W. Phan, « Ideas on interpreting machine learning. Mix-and-match approaches for visualizing data and interpreting machine learning models and results », *O'Reilly*, 15 mars 2017, en ligne : <www.oreilly.com/radar/ideas-on-interpreting-machine-learning/>

Hawkins, J., « What Intelligent Machines Need to Learn From the Neocortex », *IEEE Spectrum* (2 juin 2017), en ligne : <spectrum.ieee.org/computing/software/what-intelligent-machines-need-to-learn-from-the-neocortex>

Hildt, E., « Artificial Intelligence : Does Consciousness Matter? » (2019) Front Psychol, doi : <doi.org/10.3389/fpsyg.2019.01535>

Hoffmann, M. et al., « Robot in the Mirror : Toward an Embodied Computational Model of Mirror Self-Recognition » (2021) 35 Künstiliche Intelligenz 37, doi : <doi.org/10.1007/s13218-020-00701-7>

Illing, B., W. Gerstner et J. Brea, « Biologically plausible deep learning – But how far can we go with shallow networks? » (2019) 118 Neural Networks 90, doi : <doi.org/10.1016/j.neunet.2019.06.001>

Jiang, J. et al., « Anomaly Detection with Graph Convolutional Networks for Insider Threat and Fraud Detection » dans *MILCOM 2019 – 2019 IEEE Military Communications Conference (MILCOM)*, Norfolk (VA), 2019, 109, doi : <doi.org/10.1109/MILCOM47813.2019.9020760>

John, R.A. et al., « Self healable neuromorphic memtransistor elements for decentralized sensory signal processing in robotics » (2020) 11 Nature Communications, doi : <doi.org/10.1038/s41467-020-17870-6>

Kammer, M. et al., « A Perceptual Memory System for Affordance Learning in Humanoid Robots » dans Honkela, T. et al, dir, *Artificial Neural Networks and Machine Learning – ICANN 2011*, part II, Springer, 2011, 349, doi : <doi.org/10.1007/978-3-642-21738-8_45>

Khaligh-Razavi, S.-M. et N. Kriegeskorte, « Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation » (2014) 10 PLoS Comput Biol e1003915, doi : <doi.org/10.1371/journal.pcbi.1003915>

Khasahmadi, A.H. et al., « Memory-Based Graph Networks » (2020) ICLR, en ligne : <arxiv.org/abs/2002.09518>

Kipf, T.N. et M. Welling, « Semi-Supervised Classification with Graph Convolutional Networks » (2017) ICLR, en ligne : <arxiv.org/abs/1609.02907>

Kulkarni, T. et al., « Unsupervised Learning of Object Keypoints for Perception and Control » (2019) NeurIPS, en ligne : <arxiv.org/abs/1906.11883>

Lara, B. et al., « Embodied Cognitive Robotics and the learning of sensorimotor schemes » (2018) 26:5 Adaptive Behavior, doi : <doi.org/10.1177/1059712318780679>

Lecue, F., « On the Role of Knowledge Graphs in Explainable AI » (2019) 11:1 Semantic Web 1, DOI : <doi.org/10.3233/SW-190374>

Lee, D. et al., « Joint Interaction and Trajectory Prediction for Autonomous Driving using Graph Neural Networks » (2019) Machine Learning for Autonomous Driving NeurIPS, en ligne : <arxiv.org/abs/1912.07882>

Lettvin, J.Y. et al., « What the Frog's Eye Tells the Frog's Brain » (1959) 47:11 Proceedings of the IRE 1940, doi : <doi.org/10.1109/JRPROC.1959.287207>

Levinson, F.H., « Man and Superman : Life Near An Approaching Technology Singularity », *One Million by One Million Blog* (5 juillet 2017), en ligne : <www.sramanamitra.com/2017/07/05/man-and-superman-life-near-an-approaching-technology-singularity/>

Li, J., D. Cai et X. He, « Learning Graph-Level Representation for Drug Discovery » (2017), en ligne : <arxiv.org/abs/1709.03741>

Li, Y. et al., « Gated Graph Sequence Neural Networks » (2016) ICLR, en ligne : <arxiv.org/abs/1511.05493>

Lillicrap, T.P. et al., « Backpropagation and the brain » (2020) 21 Nature Reviews Neuroscience 335, doi : <doi.org/10.1038/s41583-020-0277-3>

Lin, J. et al., « Generalized and Scalable Optimal Sparse Decision Trees » (2020) ICML, en ligne : <arxiv.org/abs/2006.08690>

Liu, Y. et al., « Cognitive Modeling for Robotic Assembly/Maintenance Task in Space Exploration » dans *International Conference on Applied Human Factors and Ergonomics*, Los Angeles (CA), Springer, 143, doi : <doi.org/10.1007/978-3-319-60642-2_13>

Loor (de), P., A. Mille et M. Réguigne-Khamassi, « Intelligence artificielle : l'apport des paradigmes incarnés » (2015) 64:2 Revue de l'Association pour la recherche cognitive 27, doi : <doi.org/10.3406/intel.2015.1011>

Lou, H.C., J.P. Changeux et A. Rosenstand, « Towards a cognitive neuroscience of self-awareness » (2017) 83 Neuroscience & Biobehavioral Reviews 765, doi : <doi.org/10.1016/j.neubiorev.2016.04.004>

Lundberg, S. et S.-I. Lee, « A Unified Approach to Interpreting Model Predictions » (2017) NIPS, en ligne : <arxiv.org/abs/1705.07874>

MacLean, E.L., « Unraveling the evolution of uniquely human cognition » (2016) 113:23 PNAS 6348, doi : <doi.org/10.1073/pnas.1521270113>

Marcus, G., « The Next Decade in AI : Four Steps Towards Robust Artificial Intelligence » (2020), en ligne : <arxiv.org/abs/2002.06177>

McCarthy, J., « Programs with Common Sense » (1959)

McFadden, C., « Researchers in Japan to Make Artificial Skin That Can Feel Pain », *Interesting Engineering* (17 février 2020), en ligne : <interestingengineering.com/researchers-in-japan-to-make-artificial-skin-that-can-feel-pain>

Metz, C., « A.I. Is Learning From Humans. Many Humans », *New York Times* (16 août 2019), en ligne : <www.nytimes.com/2019/08/16/technology/ai-humans.html>

Ribeiro, M.T., S. Singh et C. Guestrin, « "Why Should I Trust You?" Explaining the Predictions of Any Classifier », *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, 2016, en ligne : <arxiv.org/pdf/1602.04938v1.pdf>

Rosenblatt, F., « The Perceptron : A Probabilistic Model for Inforation Storage and Organization in the Brain » (1958) 65:6 Psychological Review 386, doi : <doi.org/10.1037/h0042519>

Roth, D., « Learning to resolve natural language ambiguities : a unified approach » (1998) AAAI 806, en ligne : <dl.acm.org/doi/10.5555/295240.295894>

Rudrauf, D. et al., « The role of consciousness in biological cybernetics : emergent adaptive and maladaptive behaviours in artificial agents governed by the projective cosnciousness model » (2020), en ligne : <arxiv.org/abs/2012.12963>

Rumelhart, D.E., G.E. Hinton et R.J. Williams, « Learning representations by back-propagating errors » (1986) 323 Nature 533, doi : <doi.org/10.1038/323533a0>

Russin, J., R. C. O'Reilly et Y. Bengio, « Deep Learning Needs a Prefrontal Cortex », *Bridging AI and Cognitive Science*, ICLR 2020, en ligne : <baicsworkshop.github.io/pdf/BAICS_10.pdf>

Sabourin, C., *Systèmes cognitifs artificiels : du concept au développement de comportements intelligents en robotique autonome*, Université Paris Est Crétail, 2016, en ligne : <hal.archives-ouvertes.fr/tel-01352195>

Sanborn, A.N. et N. Chater, « Bayesian Brains without Probabilities » (2016) 20:12 Trends in Cognitive Sciences 883, doi : <doi.org/10.1016/j.tics.2016.10.003>

Sarkar, D(DJ), « A Comprhensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning », *Towards Data Science* (14 novembre 2018), en ligne : <towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a>

Scarselli, F. et al., « The graph neural network model » (2009) 20:1 IEEE Transactions on Neural Networks 61, DOI : <doi.org/10.1109/TNN.2008.2005605>

Scharre, P., *Autonomous Weapons and Operational Risk*, Ethical Autonomy Project, Center for a New American Security, février 2016, en ligne : <s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf>

Scheffel, J., « On the Solvability of the Mind-Body Problem » (2020) 30 Axiomathes 289, doi : <doi.org/10.1007/s10516-019-09454-x>

Schneegans, S. et G. Schöner, « 13 – Dynamic Field Theory as a Framework for Understanding Embodied Cognition » dans Calvo, P. et A. Gomila, dir, *Handbook of Cognitive Science*, coll « Perspectives on cognitive science », Elsevier Science, 2008, 241, doi : <doi.org/10.1016/B978-0-08-046616-3.00013-X>

Scriven, M., « An Essential Unpredictabilty in Human Behavior » dans Wolman, B.B. et E. Nagel, dir, *Scientific Psychology : Principes and Approaches*, 1965, 411

Searle, J.R., « Minds, brains, and programs » (1980) 3:3 Behavioral & Brain Sciences 417

Sejnowski, T.J., « The unreasonable effectiveness of deep learning in artificial intelligence » (2020) 117:48 PNAS 30033, DOI : <doi.org/10.1073/pnas.1907373117>

Shapiro, L., « The Embodied Cognition Research Programme » (2007) 2:2 Philosophy Compass 338, doi : <doi.org/10.1111/j.1747-9991.2007.00064.x>

Simonite, T., « Facebook's AI Chief : Machines Could Learn Common Sense from Video » (2017) MIT Technology Review, en ligne : <www.technologyreview.com/2017/03/09/153343/facebooks-ai-chief-machines-could-learn-common-sense-from-video>

Smith, J.D., « Inaugurating the Study of Animal Metacognition » (2010) 23:3 Int J Comp Psychol 401

Souza, T., « Connecting the Dots : Using AI & Knowledge Graphs to Identify Investment Opportunities », *Towards Data Sicence* (28 mars 2019), en ligne : <towardsdatascience.com/knowledge-graphs-in-investing-733ab34abe>

Srivastava, N. et al,, « Dropout : A Simple Way to Prevent Neural Networks from Overfitting » (2014) 15:56 Journal of Machine Learning Research 1929

Storrs, K.R. et R.W. Fleming, « Unsupervised Learning Predicts Human Perception and Misperception of Gloss » (2020), doi : <doi.org/10.1101/2020.04.07.026120>

Sun, M. et al, « Graph convolutional networks for computational drug development and discovery » (2020) 21:3 Briefings in Bioinformatics 919, DOI : <doi.org/10.1093/bib/bbz042>

Tayo, B.O., Ph.D., « Simplicity vs Complexity in Machine learning – Finding the Right Balance », *Towards Data Science* (11 novembre 2019), en ligne : <towardsdatascience.com/simplicity-vs-complexity-in-machine-learning-finding-the-right-balance-c9000d1726fb>

Toews, R., « The Next Generation of Artificial Intelligence », *Forbes* (12 octobre 2020), en ligne : <www.forbes.com/sites/robtoews/2020/10/12/the-next-generation-of-artificial-intelligence/?sh=39e7f5859eb1>

Tomasello, M., A.C. Kruger et H.H. Ratner, « Cultural learning » (1993) 16:3 Behavioral & Brain Sciences 495

Trewavas, A., « The foundations of plant intelligence » (2017) Interface Focus, doi : <doi.org/10.1098/rsfs.2016.0098>

Turing, A.M., « Computing Machinery and Intelligence » (1950) LIX:236 Mind 433, doi : <doi.org/10.1093/mind/LIX.236.433>

Van Hoeck, N., P.D. Watson et A.K. Barbey, « Cognitive neuroscience of human counterfactual reasoning » (2015) 9 Front Hum Neurosci 420, doi : <doi.org/10.3389/fnhum.2015.00420>

Vaswani, A. et al., « Attention Is All You Need » (2017), en ligne : <arxiv.org/abs/1706.03762>

Veličković, P. et al., « Graph Attention Networks » (2018) ICLR, en ligne : <arxiv.org/abs/1710.10903>

Vinge, V., « The Coming Technological Singularity : How to Survive in the Post-Human Era » (1993) VISION-21 Symposium, en ligne : <edoras.sdsu.edu/~vinge/misc/singularity.html>

Voss, P., « The Third Wave of AI », *Becoming Human AI* (25 septembre 2017), en ligne : <becominghuman.ai/the-third-wave-of-ai-1579ea97210b>

Wachter, S., B. Mittelstadt et C. Russell, « Counterfactual Explanations without Opening the Black Box : Automated Decisions and the GDPR » (2018) Harvard Journal of Law & Technology, en ligne : <arxiv.org/abs/1711.00399>

Weng, J., « Developmental Robotics : Theory and Experiments » (2004) 1:2 International Journal of Humanoid Robotics 199, doi : <doi.org/10.1142/S0219843604000149>

Whittington, J.C.R. et R. Bogacz, « Theories of Error Back-Propagation in the Brain » (2019) 23:3 Trends in Cognitive Sciences Review 235, doi : <doi.org/10.1016/j.tics.2018.12.005>

Williams, D., « Predictive coding and thought » (2020) 197 Synthese 1749, doi : <doi.org/10.1007/s11229-018-1768-x>

Wilson, H.J., P.R. Daugherty et C. Davenport, « The Future of AI Will Be About Less Data, Not More », *Harvard Business Review* (14 janvier 2019), en ligne : <hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more>

Winfield, A.F.T., « Experiments in Artificial Theory of Mind : From Saftey to Story-Telling » (2018) Front Robot AI, doi : <doi.org/10.3389/frobt.2018.00075>

Worrall, S., « There Is Such a Thing as Plant Intelligence », *National Geographic* (21 février 2016), en ligne : <www.nationalgeographic.com/science/article/160221-plant-science-botany-evolution-mabey-ngbooktalk>

Yu, C., Y. Chai et Y. Liu, « Literature review on collective intelligence : a crowd science perspective » (2018) 2:3 International Journal of Crowd Science, doi : <doi.org/10.1108/IJCS-08-2017-0013>

Zentall, T.H., « Animal intelligence » dans Sternberg, R.J. et S.B. Kaufman, dir, *Cambridge handbooks in psychology. The Cambridge handbook of intelligence*, 3e éd, Cambridge University Press, 2011, 309, doi : <doi.org/10.1017/CBO9780511977244.017>

Zhong, J., C. Weber et S. Wermter, « Robot Trajectory Prediction and Recognition Based on a Computational Mirror Neurons Model » dans Honkela, T. et al., dir, *Artificial Neural Networks and Machine Learning – ICANN 2011*, part II, Springer, 2011, 333, doi : <doi.org/10.1007/978-3-642-21738-8_43>

Chapitre 3

Table de la Législation

Loi de l'impôt sur le revenu (Canada), LRC 1985, c 1 (5e suppl)

Loi sur les impôts (Québec), LRQ c I-3

Règlement édictant trois programmes pilotes d'immigration permanente (projet), (2020) 152 G.O. II, 4592

Monographies et ouvrages collectifs

Bloch, L., *L'Internet, vecteur de puissance des États-Unis?* Diploweb, 2017, en ligne : <www.diploweb.com/-L-Internet-vecteur-de-puissance-des-Etats-Unis-.html>

Julia, L., L'Intelligence artificielle n'existe pas, FIRSTFORUM, 2019

Supiot, A., *La Gouvernance par les nombres. Cours au Collège de France 2012-2015*, Fayard, 2015, en ligne : <www.college-de-france.fr/site/alain-supiot/La-gouvernance-par-les-nombres-film.htm>

Articles de revues et études d'ouvrages collectifs

« L'intelligence artificielle ne peut être vaincue : un maître de go abandonne », *Le Point* (27 novembre 2019), en ligne : <www.lepoint.fr/high-tech-internet/l-intelligence-artificielle-ne-peut-etre-vaincue-un-maitre-de-go-abandonne-27-11-2019-2350129_47.php>

Arrow, K., « Economic Welfare and the Allocation of Resources for Invention » dans Universities-Nationa Bureau Committee for Economic Research, Committee on Economic Growth of the Social Science Research Council, *The Rate and Direction of Inventive Activity : Economic and Social Factors*, Princeton University Press, 1962, 609, en ligne : <www.nber.org/system/files/chapters/c2144/c2144.pdf>

Aubin, C., « Intelligence artificielle et brevets » (2018) 30:3 Les Cahiers de la propriété intellectuelle 947, en ligne : <www.lescpi.ca/articles/v30/n3/intelligence-artificielle-et-brevets/>

Colleret, M. et Y. Gingras, *L'intelligence artificielle au Québec : un réseau tricoté serré*, note de recherche 2020-07, Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal (UQÀM), 2020, en ligne : <cirst2.openum.ca/files/sites/179/2020/12/Note_2020-07_IA.pdf>

Computing Research Association (CRA), *The CRA Taulbee Survey*, 2001–2019, en ligne : <cra.org/resources/taulbee-survey/>

D'Allegro, J., « How Google's Self-Driving Car Will Change Everything », *Investopedia* (20 décembre 2020), en ligne : <www.investopedia.com/articles/investing/052014/how-googles-selfdriving-car-will-change-everything.asp>

Dagenais, M., P. Mohnen et P. Therrien, « Les firmes canadiennes répondent-elles aux incitations fiscales à la recherche-développement? » (2004) 80: 2-3 L'Actualité économique 175, DOI : <doi.org/10.7202/011385ar>

Dass, R., « 5 Key Challenges faced by Self-driving cars », *Medium* (14 septembre 2018), en ligne : <medium.com/@ritidass29/5-key-challenges-faced-by-self-driving-cars-ed04e969301e>

De Fraja, G., « Optimal public funding for research : a theoretical analysis » (2016) 47:3 The RAND Journal of Econoics 498, en ligne : <www.jstor.org/stable/43895655>

Dufour, P. et Y. Gingras, « La politique scientifique et technologique du gouvernement du Canada » dans R. Dalpé et R. Landry, dir, *La politique technologique au Québec*, Montréal, Presses de l'Université de Montréal, 1993, 129, en ligne : <archipel.uqam.ca/557/1/Politique_scientifique_technologique_Canada.pdf>

Ferraris, F.S.G., « Les voitures autonomes, une révolution qui pourrait sauver des vies », *Le Devoir* (9 janvier 2017), en ligne : <www.ledevoir.com/societe/transports-urbanisme/488715/voitures-autonomes-une-revolution-qui-pourrait-sauver-des-vies>

Fortin, L.-E., « La politique technologique québécoise » (1985) 8 Politique 23, DOI : <doi.org/10.7202/040496ar>

Gillies, D.J., « Technological Determinism in Canadian Telecommunications : Telidon Technology, Industry and Government » (1990) 15:2 Canadian Journal of Communication 1, en ligne : <digital.library.ryerson.ca/islandora/object/RULA%3A4805>

Gingras, Y., B. Godin et M. Trépanier, « La place des universités dans les politiques scientifiques et technologiques canadiennes et québécoises » dans D. Bertrand et P. Beaulieu, dir, *L'État québécois et les universités : Acteurs et enjeux*, Sainte-Foy, Presses de l'Université du Québec, 1999, 69, en ligne : <archipel.uqam.ca/537/1/Place_universite_dans_politiques_scientifique.pdf>

Guellec, D., « Les politiques de soutien à l'innovation technologique à l'aune de la théorie économique » (2001) 4-5: 150-151 Économie & Prévision 95, DOI : <doi.org/10.3917/ecop.150.0095>

Hu, S. et T. Jiang, « Artificial Intelligence Technology Challenges Patent Laws » dans *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Changsha (Chine), IEEE, 2019, 241, DOI : <doi.org/10.1109/ICITBS.2019.00064>

Joly, P.-B., « Le régime des promesses technoscientifiques » dans Marc Audétat, *Sciences et technologies émergentes : Pourquoi tant de promesses*, Paris, Hermann, 2015, 31, en ligne : <www.researchgate.net/publication/297622208_Le_regime_des_promesses_technoscientifique>

Kegels, C., « La politique d'innovation dans une économie de la connaissance » (2009) 1-2 Reflets et perspectives de la vie économique 151, DOI : <doi.org/10.3917/rpve.481.0151>

Knight, W., « AlphaGo Zero Shows Machines Can Become Superhuman Without Any Help », *MIT Technology Review* (18 octobre 2017), en ligne : <www.technologyreview.com/2017/10/18/148511/alphago-zero-shows-machines-can-become-superhuman-without-any-help/>

LaMonica, M., « Should the Government Support Applied Research? », *MIT Technology Review* (10 septembre 2012), en ligne : <www.technologyreview.com/2012/09/10/183924/should-the-government-support-applied-research/>

Le Roux, M. et G. Ramunni, « L'OCDE et les politiques scientifiques » (2000) 3 Revue pour l'histoire du CNRS, DOI : <doi.org/10.4000/histoire-cnrs.2952>

Leith, P., « The rise and fall of the legal expert system » (2016) 30:3 Int Rev Law, Comp & Tech 94, DOI : <doi.org/10.1080/13600869.2016.1232465>

Lomazzi, L., M. Lavoie-Moore et J. Gélinas, *Financer l'intelligence artificielle, quelles retombées économiques et sociales pour le Québec?* Institut de recherche et d'informations socio-économiques (IRIS), mars 2019, en ligne : <cdn.iris-recherche.qc.ca/uploads/publication/file/Intelligence_artificielle_IRIS_WEB4.pdf>

Mansfield, E. et L. Switzer, « The effects of R&D tax credits and allowances in Canada » (1985) 14:2 Research Policy 97, DOI : <doi.org/10.1016/0048-7333(85)90017-4>

Marcus, G., *Deep Learning : A Critical Appraisal*, 2018, en ligne : <arxiv.org/pdf/1801.00631.pdf>

Maxmen, A., « Self-driving car dilemmas reveal that moral choices are not universal » (2018) 562 Nature 469, DOI : <doi.org/10.1038/d41586-018-07135-0>

Miailhe, N., « Géopolitique de l'intelligence artificielle : le retour des empires? » (2018) 3 Politique étrangère 105, en ligne : <www.ifri.org/sites/default/files/atoms/files/geopolitique_de_lintelligence_artificielle.pdf>

Moreau, N., « Internet in Canada », *The Canadian Encyclopedia* (3 décembre 2012), en ligne : <www.thecanadianencyclopedia.ca/en/article/internet>

Oosterlinck, A., K. Debackere et G. Cielen, « Balancing basic and applied research » (2002) 3:1 EMBO Rep 2, doi : <doi.org/10.1093/embo-reports/kvf016>

Perrault, R. et al, *The AI Index 2019 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford (CA), décembre 2019, en ligne : <hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf>

Richards, W., S. Yusufali et R. Marsh, « Télidon », *L'Encyclopédie canadienne* (28 janvier 2007), en ligne : <www.thecanadianencyclopedia.ca/fr/article/telidon>

Roberts, H. et al., « The Chinese approach to artificial intelligence : an analysis of policy, ethics, and regulation » (2021) 36 AI & Society 59, doi : <doi.org/10.1007/s00146-020-00992-2>

Shoham, Y. et al., *The AI Index 2018 Annual Report*, AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford (CA), décembre 2018, en ligne : <hai.stanford.edu/sites/default/files/2020-10/AI_Index_2018_Annual_Report.pdf>

Van Lente, H., « Navigating foresight in a sea of expectations : Lessons from the sociology of expectations » (2012) 24:8 Technology Analysis & Strategic Management 769, DOI : <doi.org/10.1080/09537325.2012.715478>

Van Pottelsberghe De La Potterie, B., « Les politiques de science et technologie et l'objectif de Lisbonne » (2004) XLIII:1 Reflets et perspectives de la vie économique 69, DOI : <doi.org/10.3917/rpve.431.0069>

Waldrop, M., *DARPA and the Internet Revolution*, DARPA, 2015, en ligne : <www.darpa.mil/attachments/(2O15)%20Global%20Nav%20-%20About%20Us%20-%20History%20-%20Resources%20-%2050th%20-%20Internet%20(Approved).pdf>

Yanisky-Ravid, S. et R. Jin, « Summoning a New Artificial Intelligence Patent Model : In The Age Of Pandemic » (2020), version préimprimée, DOI : <doi.org/10.2139/ssrn.3619069>

Documents gouvernementaux

Alberta Treasury Board and Finance, *Fiscal Plan : A Plan for Jobs and the Economy 2020-23*, Budget 2020, Edmonton, février 2020 à la p 171, en ligne : <open.alberta.ca/dataset/05bd4008-c8e3-4c84-949e-cc18170bc7f7/resource/79caa22e-e417-44bd-8cac-64d7bb045509/download/budget-2020-fiscal-plan-2020-23.pdf>

Bush, V., *Science : The Endless Frontier*, Washington (DC), United States Government Printing Office, 1945, en ligne : <www.nsf.gov/od/lpa/nsf50/vbush1945.htm#ch1.3>

Commission d'examen sur la fiscalité québécoise, *Compétitivité, efficacité, équité. Se tourner vers l'avenir du Québec*, vol 1 « Une réforme de la fiscalité québécoise », rapport final, gouvernement du Québec, mars 2015, en ligne : <www.groupes.finances.gouv.qc.ca/examenfiscalite/uploads/media/Volume1_RapportCEFQ_01.pdf>

Commission européenne, *AI Watch. National strategies on Artificial Intelligence. Au European perspective in 2019*, JRC Technical Report, Union européenne, 2020, en ligne : <publications.jrc.ec.europa.eu/repository/bitstream/JRC119974/national_strategies_on_artificial_intelligence_final_1.pdf>

Commission européenne, *L'intelligence artificielle pour l'Europe*, communication de la Commission au Parlement européen, au Conseil européen, au Conseil, au Comité économique et social européen et au Comité des régions, COM(2018) 237 final, Bruxelles, 25 avril 2018, en ligne : <ec.europa.eu/transparency/regdoc/rep/1/2018/FR/COM-2018-237-F1-FR-MAIN-PART-1.PDF>

Commission européenne, *LIVRE BLANC : Intelligence artificielle. Une approche européenne axée sur l'excellence et la confiance*, COM(2020) 65 final, Bruxelles, 19 février 2020, en ligne : <www.eesc.europa.eu/fr/our-work/opinions-information-reports/opinions/livre-blanc-sur-lintelligence-artificielle>

Conseil d'État de Chine, 新一代人工智能发展规划 *[Plan de développement de l'intelligence artificielle de nouvelle génération]*, 8 juillet 2017, en ligne : <www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm>; traduit en anglais par Graham Webster, Rogier Creemers, Paul Triolo et Elsa Kania, *Full Translation : China's 'New Generation Artificial Intelligence Development Plan'*, 1er août 2017, en ligne : <www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>

Conseil de l'Europe, *Initiatives sur l'IA*, en ligne : <www.coe.int/fr/web/artificial-intelligence/national-initiatives>

Defense Advanced Research Projects Agency (DARPA), *AI Next Campaign*, en ligne : <www.darpa.mil/work-with-us/ai-next-campaign>

Defense Advanced Research Projects Agency (DARPA), *DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies*, 7 septembre 2018, en ligne : <www.darpa.mil/news-events/2018-09-07>

Gouvernement du Canada, *Évolution du Programme de la RS&DE – une perspective historique*, 2015, en ligne : <www.canada.ca/fr/agence-revenu/services/recherche-scientifique-developpement-experimental-programme-encouragements-fiscaux/evolution-programme-perspective-historique.html>

Gouvernement du Québec, *Vos priorités, votre budget : Plan budgétaire*, budget 2019-2020, mars 2019, en ligne : <www.budget.finances.gouv.qc.ca/budget/2019-2020/fr/documents/PlanBudgetaire_1920.pdf>

Ministère de l'Économie, de la Science et de l'Innovation, *Stratégie québécoise de la recherche et de l'innovation 2017-2022*, gouvernement du Québec, 2017, en ligne : <www.economie.gouv.qc.ca/fileadmin/contenu/documents_soutien/strategies/recherche_innovation/SQRI/sqri_complet_fr.pdf>

Ministre d'État au développement culturel du Québec, *La politique québécoise du développement culturel*, vol 2 « Les trois dimensions d'une politique : genres de vie, création, éducation », 1978, en ligne : <classiques.uqac.ca/contemporains/Quebec_gouvernement_du/Politique_qc_devel_culturel_t2/Politique_qc_devel_culturel_t2.pdf>

Ministre des Finances du Canada, *Pour assurer le renouveau économique. Documents budgétaires*, déposés à la Chambre des Communes, 23 mai 1985, en ligne : <www.budget.gc.ca/pdfarch/1985-pap-fra.pdf>

Ministre des Finances du Québec, *Budget 1987-1988. Discours sur le budget et Renseignements supplémentaires*, 30 avril 1987, en ligne : <www.budget.finances.gouv.qc.ca/budget/archives/fr/documents/1987-88_fine.pdf>

Ministre des Finances du Québec, *Budget 1988-1989. Discours sur le budget et Renseignements supplémentaires,* 12 mai 1988, en ligne : <www.budget.finances.gouv.qc.ca/budget/archives/fr/documents/1988-89_fine.pdf>

Ministre des Finances du Québec, *Budget 1998-1999. Renseignements supplémentaires sur les mesures du budget*, 31 mars 1998, en ligne : <www.budget.finances.gouv.qc.ca/budget/1998-1999/fr/PDF/rensupfr.pdf>

Ministre des Finances du Québec, *Budget 1999-2000. Discours sur le budget*, 9 mars 1999 à la p 18, en ligne : <www.budget.finances.gouv.qc.ca/budget/1999-2000/fr/PDF/disc-fr.pdf>

Ministre des Finances du Québec, *Budget 1999-2000. Renseignements supplémentaires sur les mesures du budget*, 9 mars 1999 à la p 24, en ligne : <www.budget.finances.gouv.qc.ca/budget/1999-2000/fr/PDF/disc-fr.pdf >

Ministre des Finances du Québec, *Budget 2006-2007. Renseignements additionnels sur les mesures du budget*, mars 2006, en ligne : <www.budget.finances.gouv.qc.ca/budget/2006-2007/fr/pdf/RenseignementsAdd.pdf>

National Research Council, *Funding a Revolution. Government Support for Computing Research*, Washington, DC, National Academies Press, 1999, doi : <doi.org/10.17226/6323>

National Research Council, *Furthering America's Research Enterprise*, Washington, DC, The National Academies Press, 2014, doi : <doi.org/10.17226/18804>

National Science and Technology Council (Networking and Information Technology Research and Development Subcommittee), *The National Artificial Intelligence Research and Development Strategic Plan*, Executive Office of the President of the United States, octobre 2016, en ligne : <www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf>

Secrétariat du Conseil du Trésor, *Stratégie de transformation numérique gouvernementale 2019-2023*, gouvernement du Québec, 2019, en ligne : <cdn-contenu.quebec.ca/cdn-contenu/adm/min/secretariat-du-conseil-du-tresor/publications-adm/strategie/StrategieTNG.pdf?1559512998>

Strategic Council for AI Technology, *Artificial Intelligence Technology Strategy*, gouvernement du Japon, 31 mars 2017, en ligne : <ai-japan.s3-ap-northeast-1.amazonaws.com/7116/0377/5269/Artificial_Intelligence_Technology_StrategyMarch2017.pdf>

Documents internationaux

Convention relative à l'Organisation de Coopération et de Développement Économiques, Paris, 14 décembre 1960, en ligne : <www.oecd.org/fr/general/conventionrelativealorganisationde cooperationetdedeveloppementeconomiques.htm>

OECD.AI Policy Observatory, National AI policies & strategies, en ligne : <oecd.ai/dashboards>

Organisation de coopération et de développement économique (OCDE), La Science et la politique des gouvernements. L'influence de la science et de la technique sur la politique nationale et internationale, 1963

Organisation de coopération et de développement économique (OCDE), L'OCDE hébergera le Secrétariat du nouveau Partenariat mondial sur l'intelligence artificielle, 15 juin 2020, en ligne : <www.oecd.org/fr/presse/l-ocde-hebergera-le-secretariat-du-nouveau-partenariat-mondial-sur-l-intelligence-artificielle.htm>

Organisation mondiale de la Propriété intellectuelle (OMPI), Artificial Intelligence. WIPO Technology Trends 2019, Genève, OMPI, 2019, en ligne : <www.wipo.int/edocs/pubdocs/en/wipo_pub_ 1055.pdf>

Partenariat mondial sur l'intelligence artificielle (PMIA), À propos, en ligne : <gpai.ai/fr/a-propos/>

UNESCO, Commission mondiale d'éthique des connaissances scientifiques et des technologies (COMEST), Étude préliminaire sur l'éthique de l'intelligence artificielle, SHS/COMEST/EXTWG-ETHICS-AI/2019/1, Paris, 26 février 2019, en ligne : <unesdoc.unesco.org/ark:/48223/pf0000367823_fre>

UNESCO, Commission mondiale d'éthique des connaissances scientifiques et des technologies (COMEST), Étude préliminaire sur l'éthique de l'intelligence artificielle, SHS/COMEST/EXTWG-ETHICS-AI/2019/1, Paris, 26 février 2019 à la p 27, en ligne : <unesdoc.unesco.org/ark:/48223/ pf0000367823_fre>

UNESCO, Composition du groupe d'experts ad hoc (GEAH) pour la recommandation sur l'éthique de l'intelligence artificielle, SHS/BIO/AHEG-AI/2020/INF.1 REV, Paris, 11 mars 2020, en ligne : <unesdoc.unesco.org/ark:/48223/pf0000372991>

UNESCO, Élaboration d'une Recommandation sur l'éthique de l'intelligence artificielle, en ligne : <fr.unesco.org/artificial-intelligence/ethics>