1. Paysage technique de l'IA

Ce chapitre décrit les caractéristiques du paysage technique de l'intelligence artificielle (IA), qui s'est métamorphosé depuis 1950, lorsqu'Alan Turing s'est interrogé pour la première fois sur la capacité des machines à penser. Depuis 2011, des progrès décisifs ont été réalisés dans une branche de l'IA dénommée « apprentissage automatique », qui permet à des machines de s'appuyer sur des approches statistiques pour apprendre à partir de données historiques et formuler des prévisions dans des situations nouvelles. La maturité des techniques d'apprentissage automatique, conjuguée à des ensembles de données volumineux et à l'augmentation de la puissance de calcul, ont contribué à l'accélération du développement de l'IA. Ce chapitre donne également un aperçu général d'un système d'IA, qui établit des prévisions, formule des recommandations ou prend des décisions influant sur l'environnement. Il décrit ensuite le cycle de vie type d'un système d'IA, qui se décompose en quatre phases, à savoir : i) la phase de « conception, données et modèles », qui comprend la planification et la conception, la collecte et le traitement des données, ainsi que la construction et l'interprétation du modèle ; ii) la phase de « vérification et validation » ; iii) la phase de « déploiement » ; et iv) la phase d'« exploitation et (de) suivi ». Enfin, il propose une taxinomie de recherche à l'intention des décideurs.

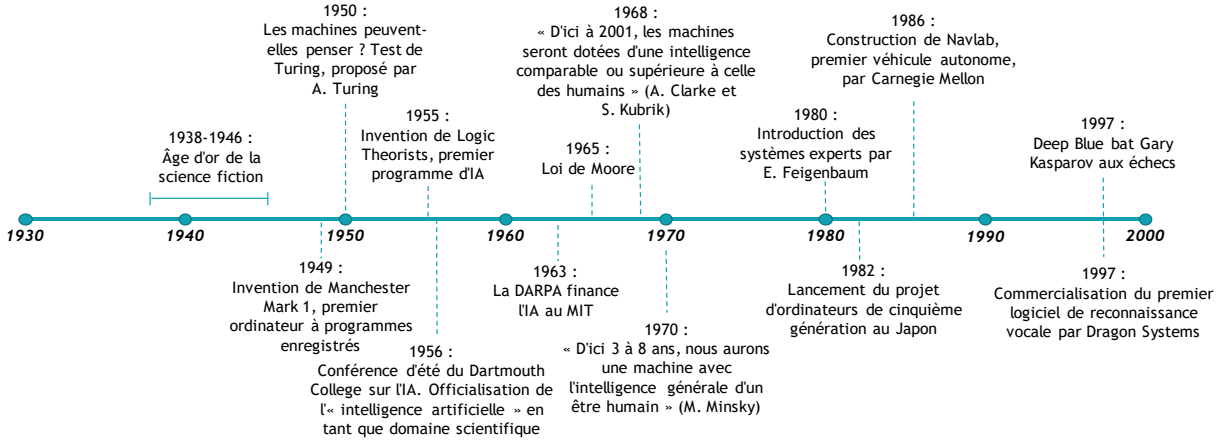
Genèse de l'intelligence artificielle

En 1950, Alan Turing, mathématicien britannique, publie un article sur l'ordinateur et l'intelligence, intitulé *Computing Machinery and Intelligence* (Turing, 1950[1]), dans lequel il s'interroge sur la capacité des machines à penser. Il développe alors une heuristique simple pour tester son hypothèse : un ordinateur pourrait-il mener une conversation et répondre à des questions d'une manière qui puisse conduire une personne suspicieuse à penser que l'ordinateur est en réalité un humain[1] ? De là naît le « test de Turing », encore utilisé de nos jours. La même année, Claude Shannon propose la création d'une machine à laquelle on pourrait apprendre à jouer aux échecs (Shannon, 1950[2]). L'entraînement de la machine pouvait alors se faire en recourant à la force brute ou en évaluant un ensemble réduit de déplacements stratégiques de l'adversaire (UW, 2006[3]).

Nombreux sont ceux qui considèrent le *Dartmouth Summer Research Project*, mené à l'été 1956, comme le point de départ de l'intelligence artificielle (IA). Lors de cet atelier, John McCarthy, Alan Newell, Arthur Samuel, Herbert Simon et Marvin Minsky ont conceptualisé le principe de l'IA. Si les recherches dans le domaine de l'IA n'ont cessé de progresser au cours des 60 dernières années, les promesses de ses précurseurs se révèlent à l'époque par trop optimistes. L'IA connaît alors, dans les années 70, un temps d'arrêt (on parle de l'« hiver de l'IA »), marqué par une chute des financements et de l'intérêt pour la recherche connexe.

On observe dans les années 90 un regain sur ces deux fronts, à la faveur des progrès en termes de puissance de calcul (UW, 2006[3]). Le Graphique 1.1 propose une chronologie de l'évolution de l'IA depuis sa naissance.

Graphique 1.1. Chronologie de l'évolution de l'IA (des années 50 à 2000)



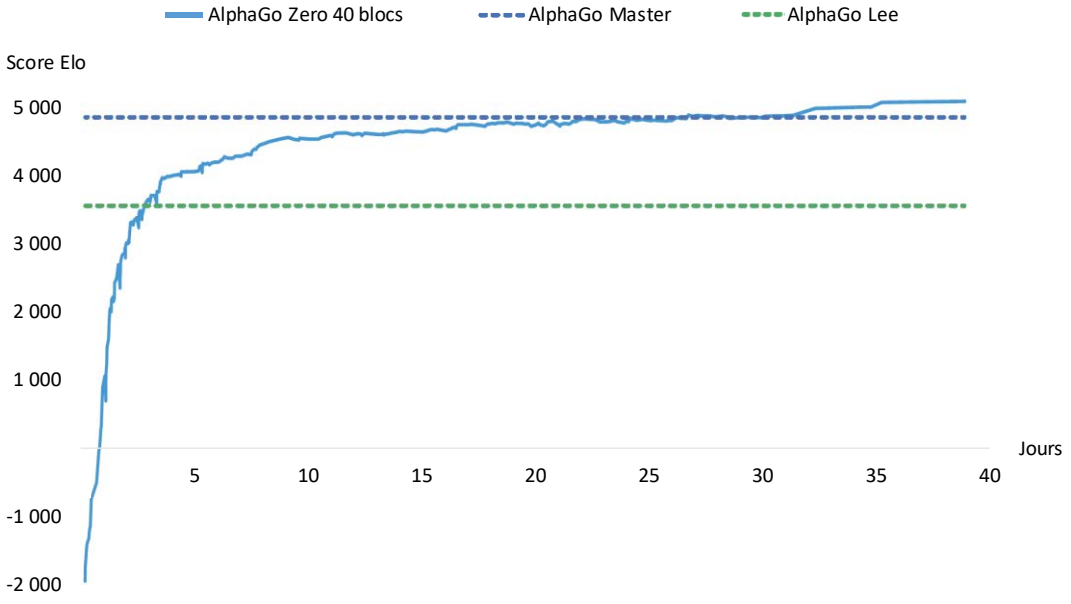
Source : D'après Anyoha (28 août 2017[4]), « The history of artificial intelligence », http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

Après l'« hiver de l'IA », qui prend fin dans les années 90, les progrès de la puissance de calcul et des capacités de stockage des données rendent possible l'exécution de tâches complexes. En 1995, l'IA franchit une étape décisive, avec le développement, par Richard Wallace, d'A.L.I.C.E. (*Artificial Linguistic Internet Computer Entity*), un programme capable de tenir une conversation basique. Toujours dans les années 90, IBM met au point un ordinateur nommé Deep Blue, qui s'appuie sur une approche fondée sur la force brute pour affronter le champion du monde d'échecs, Gary Kasparov. Deep Blue est alors capable d'anticiper

six étapes ou plus et de calculer 330 millions de positions par seconde (Somers, 2013[5]). En 1996, Deep Blue perd contre Kasparov, avant de remporter la revanche un an plus tard.

En 2015, DeepMind, filiale d'Alphabet, lance un logiciel à même d'affronter au jeu ancestral de Go les meilleurs joueurs mondiaux. Pour ce faire, il fait appel à un réseau neuronal artificiel qui a appris à jouer en s'entraînant sur des milliers de parties exécutées par des professionnels et des amateurs humains. En 2016, AlphaGo bat le champion du monde de l'époque, Lee Sedol, quatre jeux à un. Ses développeurs laissent alors le programme jouer contre lui-même en s'appuyant uniquement sur un apprentissage par essai et erreur, et en commençant par des parties totalement aléatoires, sur la base de quelques règles simples. De là naît le programme AlphaGo Zero, capable, avec un entraînement accéléré, de battre le programme AlphaGo initial par 100 jeux à 0. Entièrement autodidacte – sans intervention humaine ni utilisation de données d'historique –, AlphaGo Zero parvient, en 40 jours, à surpasser toutes les autres versions d'AlphaGo (Silver et al., 2017[6]) (Graphique 1.2).

Graphique 1.2. Auto-apprentissage rapide d'AlphaGo pour devenir le champion du monde de jeu de Go en 40 jours



Source : D'après Silver et al. (2017[6]), « Mastering the game of Go without human knowledge », http://dx.doi.org/10.1038/nature24270.

Situation actuelle

Au cours des dernières années, la montée en puissance des données massives, de l'infonuagique et des capacités de calcul et de stockage connexes, alliée aux progrès d'une branche de l'IA nommée « apprentissage automatique », ont dopé la puissance, la disponibilité, le développement et l'impact de l'IA.

Par ailleurs, les avancées technologiques constantes ouvrent la voie à des capteurs plus performants et abordables, qui capturent des données plus fiables venant nourrir les systèmes d'IA. Les volumes de données auxquels accèdent les systèmes d'IA continuent de croître à mesure que la taille et le coût réduits des capteurs en favorisent le déploiement. Cela permet, par ricochet, de réaliser des progrès majeurs dans de nombreux domaines de recherche en IA fondamentale, notamment :

- le traitement du langage naturel
- les véhicules autonomes et la robotique
- la vision par ordinateur
- l'apprentissage des langues.

Certaines des évolutions les plus intéressantes de l'IA ont lieu non pas dans l'informatique, mais dans des domaines comme la santé, la médecine, la biologie et la finance. La transition vers l'IA ressemble à bien des égards au processus de diffusion des ordinateurs qui, après avoir été l'apanage de quelques entreprises spécialisées, a gagné l'ensemble de l'économie et de la société dans les années 90. Elle n'est pas non plus sans rappeler l'expansion de l'accès à l'internet, des entreprises multinationales à une majorité de la population de nombreux pays, dans les années 2000. Les économies vont avoir de plus en plus besoin de personnes disposant de doubles spécialités, c'est-à-dire spécialisées dans une discipline comme l'économie, la biologie ou le droit, mais disposant également de compétences dans les techniques d'IA telles que l'apprentissage automatique. Le présent chapitre s'intéresse aux applications qui sont utilisées ou se profilent à court et moyen termes, plutôt qu'à de possibles évolutions à plus long terme, comme l'intelligence générale artificielle (en anglais, *artificial general intelligence*, ou AGI) (Encadré 1.1).

Encadré 1.1. Intelligence étroite artificielle et intelligence générale artificielle

L'intelligence étroite artificielle (en anglais *artificial narrow intelligence*, ou ANI), également dénommée IA « appliquée », est conçue pour accomplir une tâche de raisonnement ou de résolution de problème spécifique. Elle correspond à l'état actuel de la technologie. Les systèmes d'IA les plus perfectionnés disponibles aujourd'hui, comme AlphaGo de Google, n'en restent pas moins « étroits ». Ils peuvent, dans une certaine mesure, généraliser la reconnaissance de schémas et de formes, en transférant par exemple des connaissances apprises dans le domaine de la reconnaissance d'images vers celui de la reconnaissance de la parole. Toutefois, l'esprit humain est bien plus polyvalent.

On oppose souvent à l'IA appliquée l'intelligence générale artificielle (*artificial general intelligence*, ou AGI), qui reste pour l'heure hypothétique. Dans ce cas, les machines autonomes deviendraient capables d'actions relevant de l'intelligence générale. Comme les humains, elles seraient en mesure de généraliser et de synthétiser des apprentissages acquis par le biais de différentes fonctions cognitives. L'AGI serait assortie d'une mémoire associative développée et d'une capacité de raisonnement et de prise de décisions. Elle pourrait résoudre des problèmes multifacettes, apprendre par la lecture ou l'expérience, créer des concepts, percevoir le monde mais aussi elle-même, inventer et faire preuve de créativité, réagir aux imprévus dans des environnement complexes, ou encore anticiper. Les points de vue divergent sensiblement sur les perspectives d'AGI. Les experts sont d'avis qu'il convient d'être réaliste en termes de calendrier. Ils s'accordent dans l'ensemble sur le fait que l'ANI donnera lieu à des opportunités, des risques et des défis nouveaux importants, et conviennent que l'avènement éventuel de l'AGI, peut-être au cours du XXIe siècle, décuplerait ces conséquences.

Source : OCDE (2018[7]), *Perspectives de l'économie numérique de l'OCDE 2017*, https://doi.org/10.1787/9789264282483-fr.

L'IA, qu'est-ce que c'est ?

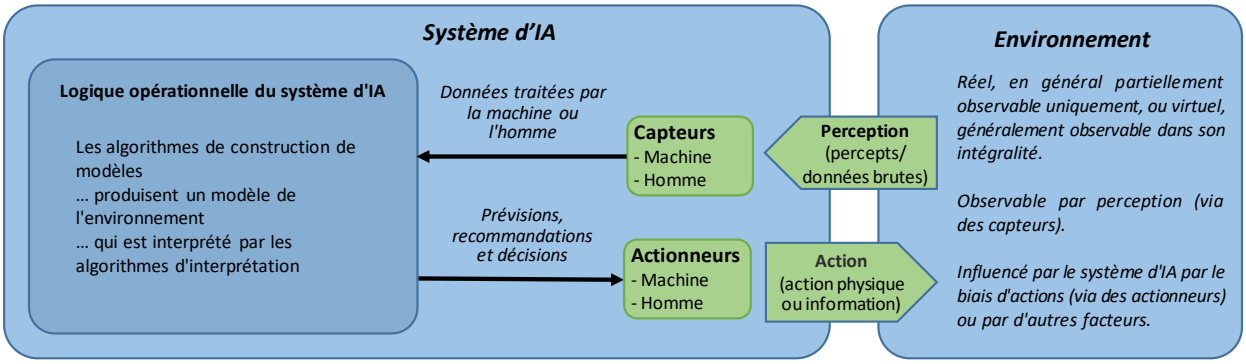
Il n'existe pas de définition universellement admise de l'IA. En novembre 2018, le Groupe d'experts sur l'intelligence artificielle à l'OCDE (AIGO) a créé un sous-groupe chargé d'élaborer une description d'un système d'IA. Cette description entend être compréhensible, techniquement juste, neutre du point de vue de la technologie et applicable aux horizons à court et long termes. Elle est suffisamment large pour couvrir nombre de définitions de l'IA couramment utilisées par la communauté scientifique, les entreprises et les pouvoirs publics. Elle a également nourri l'élaboration de la *Recommandation du Conseil de l'OCDE sur l'intelligence artificielle* (OCDE, 2019[8]).

Vision conceptuelle d'un système d'IA

La présente description d'un système d'IA s'appuie sur la vision conceptuelle de l'IA exposée dans l'ouvrage *Artificial Intelligence: A Modern Approach* (Russel et Norvig, 2009[9]). Celle-ci est cohérente avec la définition fréquente de l'IA comme « ensemble des mécanismes permettant à un agent de percevoir, de raisonner et d'agir » (Winston, 1992[10]) et avec des définitions générales similaires (Gringsjord et Govindarajulu, 2018[11]).

Une première vision conceptuelle de l'IA donne à voir une structure de haut niveau d'un système d'IA générique (également dénommé « agent intelligent ») (Graphique 1.3). Un système d'IA comporte trois éléments principaux : des capteurs, une logique opérationnelle et des actionneurs. Les capteurs collectent des données brutes à partir de l'environnement, tandis que les actionneurs agissent de manière à modifier l'état de l'environnement. La véritable puissance d'un système d'IA réside dans sa logique opérationnelle. Pour un ensemble déterminé d'objectifs et à partir de données d'entrée issues des capteurs, la logique opérationnelle produit des résultats en sortie à l'intention des actionneurs. Ceux-ci prennent la forme de recommandations, de prévisions ou de décisions susceptibles d'influer sur l'état de l'environnement.

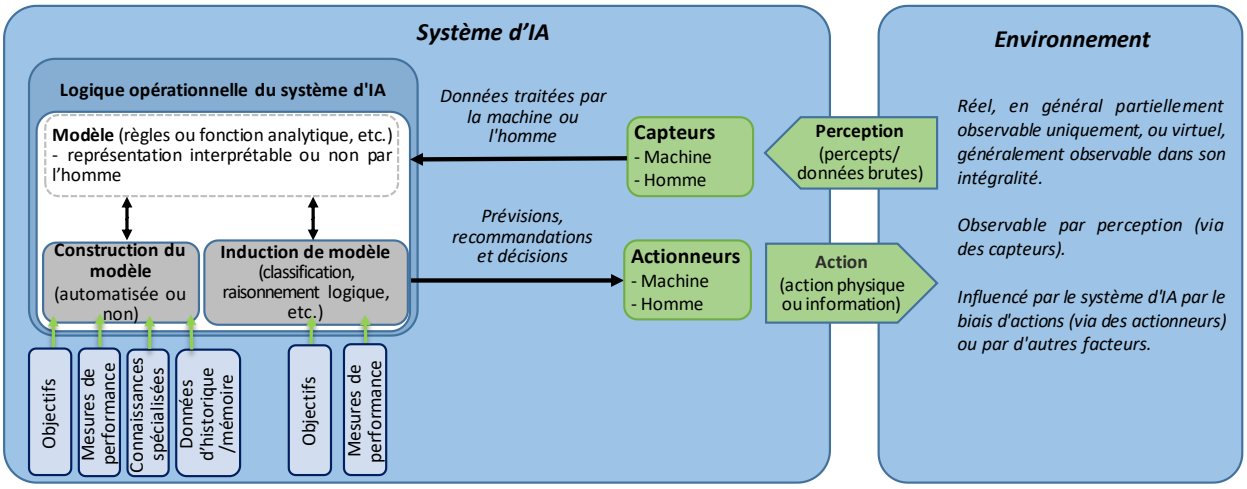
Graphique 1.3. Vision conceptuelle de haut niveau d'un système d'IA



Source : Tel que défini et approuvé par l'AIGO en février 2019.

Une structure plus détaillée rend compte des principaux éléments liés aux dimensions des systèmes d'IA intéressant l'action des pouvoirs publics (Graphique 1.4). Pour couvrir différents types de systèmes d'IA et divers scénarios, le diagramme distingue le processus de construction du modèle (comme l'apprentissage automatique) du modèle lui-même. La construction du modèle est également séparée du processus d'interprétation, qui utilise le modèle pour établir des prévisions, formuler des recommandations et prendre des décisions ; les actionneurs utilisent ces résultats pour influer sur l'environnement.

Graphique 1.4. Vision conceptuelle détaillée d'un système d'IA



Source : Tel que défini et approuvé par l'AIGO en février 2019.

Environnement

Dans le contexte d'un système d'IA, un environnement est un espace observable par le biais de perceptions (via des capteurs) et influencé au moyen d'actions (via des actionneurs). Les capteurs et les actionneurs sont soit des machines, soit des hommes. Les environnements sont quant à eux soit réels (environnement physique, social, mental, etc.) et, en règle générale, partiellement observables uniquement, soit virtuels (à l'instar des jeux, par exemple) et généralement observables dans leur intégralité.

Système d'IA

Un 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. Pour ce faire, il se fonde sur des entrées machine et/ou humaines pour : i) percevoir les environnements réels et/ou virtuels ; ii) transcrire ces perceptions en modèles grâce à une analyse manuelle ou automatisée (s'appuyant par exemple sur l'apprentissage automatique) ; et iii) utiliser des inductions des modèles pour formuler des possibilités de résultats (informations ou actions à entreprendre). Les systèmes d'IA sont conçus pour fonctionner à des niveaux d'autonomie divers.

Modèle d'IA, construction et interprétation de modèle

Au cœur d'un système d'IA se trouve le modèle d'IA, représentation de tout ou partie de l'environnement externe du système qui en décrit la structure et/ou la dynamique. Un modèle peut être fondé sur des connaissances spécialisées et/ou des données, émanant d'humains et/ou d'outils automatisés (des algorithmes d'apprentissage automatique, par exemple). Les objectifs (tels que les variables de sortie) et les mesures de performance (fiabilité, ressources d'entraînement, représentativité de l'ensemble de données) guident le processus de construction. L'induction est le processus par lequel les humains et/ou les outils automatisés déduisent des résultats à partir du modèle. Ceux-ci prennent la forme de recommandations, de prévisions ou de décisions. Les objectifs et les mesures de performance guident l'exécution. Dans certains cas (règles déterministes), le modèle peut générer une seule recommandation. Dans d'autres (modèles probabilistes), il peut en proposer une variété. Ces recommandations sont associées à différents niveaux, par exemple de mesures de performance telles que le niveau de confiance,

de robustesse ou de risque. Il peut être possible, au cours du processus d'interprétation, d'expliquer pourquoi des recommandations particulières ont été formulées ; parfois, c'est impossible.

Exemples de systèmes d'IA

Système d'évaluation des risques-clients

Un système d'évaluation des risques-clients est un exemple de système automatisé influant sur son environnement (octroi ou non d'un prêt à une personne). Il émet des recommandations (cotes de crédit) pour un ensemble donné d'objectifs (solvabilité). Pour ce faire, il utilise à la fois des entrées machine (données sur les profils des personnes et sur leur historique de remboursement d'emprunts) et des entrées humaines (ensemble de règles). À partir de ces deux jeux d'entrées, le système perçoit les environnements réels (à savoir si les personnes remboursent régulièrement leurs emprunts), puis transcrit automatiquement ces perceptions en modèles. Un algorithme d'évaluation des risques-clients pourrait par exemple utiliser un modèle statistique. Enfin, il a recours à des inductions (algorithme d'évaluation des risques) pour formuler des recommandations (cotes de crédit) sur les possibilités de résultats (accorder ou refuser le prêt).

Assistant pour malvoyants

Un assistant pour les personnes souffrant d'une déficience visuelle illustre la manière dont un système automatisé influe sur son environnement. Il émet des recommandations (comment un malvoyant peut éviter un obstacle ou traverser une rue) pour un ensemble donné d'objectifs (se déplacer d'un endroit à un autre). Pour ce faire, il utilise des entrées machine et/ou humaines (de volumineuses bases de données contenant des images étiquetées d'objets, de mots écrits, voire de visages humains) à trois fins. Premièrement, il perçoit les images de l'environnement (une caméra capture une image de ce qui se trouve devant une personne et la transmet à une application). Deuxièmement, il transcrit automatiquement ces perceptions en modèles (algorithmes de reconnaissance d'objets, capables de reconnaître un feu de signalisation, un véhicule ou un obstacle sur le trottoir). Troisièmement, il recourt à des inductions pour recommander des possibilités de résultats (en fournissant une description audio des objets détectés dans l'environnement) afin que la personne puisse décider de l'action à entreprendre et, par conséquent, influer sur son environnement.

AlphaGo Zero

AlphaGo Zero est un système d'IA capable de battre au jeu de Go n'importe quel joueur professionnel. L'environnement du jeu est virtuel et observable dans son intégralité. Les positions sont contraintes par les objectifs et les règles du jeu. Le système AlphaGo Zero utilise à la fois des entrées humaines (les règles du jeu de Go) et des entrées machine (apprentissage par le jeu itératif contre lui-même, en commençant par des parties entièrement aléatoires). Il transcrit alors les données en modèle (stochastique) d'actions (« déplacements » dans le jeu) entraîné à l'aide d'une technique dite d'« apprentissage par renforcement ». Enfin, il utilise le modèle pour proposer un nouveau déplacement d'après l'état d'avancement de la partie.

Système de conduite automatisée

Le système de conduite automatisée est un exemple de système automatisé capable d'influer sur son environnement (en décidant si le véhicule accélère, ralentit ou opère un virage). Il formule des prévisions (prévoit si un objet ou un panneau correspond à un obstacle ou à une instruction) et/ou prend des décisions (accélérer, freiner, etc.) pour un ensemble donné

d'objectifs (aller d'un point A à un point B en toute sécurité, le plus rapidement possible). Pour ce faire, il utilise à la fois des entrées machine (données d'historique de conduite) et des entrées humaines (ensemble de règles de conduite). Ces entrées sont utilisées pour créer un modèle du véhicule et de son environnement. Il permet ainsi au système d'atteindre trois objectifs. Premièrement, il peut percevoir les environnements réels (via des capteurs comme des caméras et des sonars). Deuxièmement, il peut transcrire automatiquement ces perceptions en modèles (reconnaissance d'objet ; vitesse et détection de trajectoire ; et données géodépendantes). Troisièmement, il peut recourir à l'induction. Par exemple, l'algorithme de conduite automatisée peut comporter de nombreuses simulations de possibilités à court terme pour le véhicule et son environnement. Il peut ainsi recommander des possibilités de résultats (s'arrêter ou avancer).

Cycle de vie d'un système d'IA

En novembre 2018, l'AIGO a créé un sous-groupe dont les travaux sont destinés à éclairer l'élaboration de la *Recommandation du Conseil de l'OCDE sur l'intelligence artificielle* (OCDE, 2019[8]) en détaillant le cycle de vie d'un système d'IA. Ce cadre n'a pas vocation à devenir une nouvelle norme sur le cycle de vie de l'IA[2] ni à proposer des directives. En revanche, il peut aider à définir un contexte pour d'autres initiatives internationales sur les principes de l'IA[3].

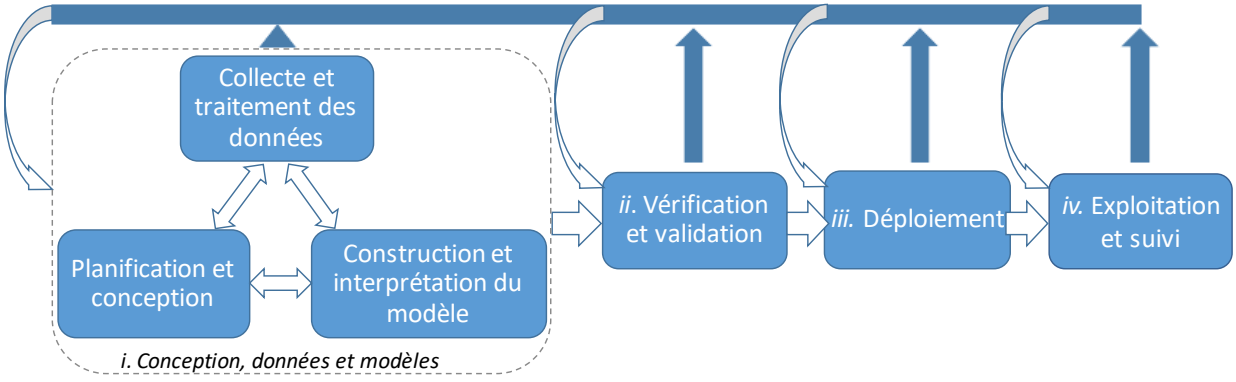
Un système d'IA présente de nombreuses phases communes avec les cycles traditionnels de développement des logiciels et, plus généralement, des systèmes. En revanche, le cycle de vie type d'un système d'IA comporte quatre phases spécifiques. La phase de conception, données et modèles est une séquence dépendante du contexte comprenant la planification et la conception, la collecte et le traitement des données, ainsi que la construction et l'interprétation du modèle. Elle est suivie des phases de vérification et validation, puis de déploiement et, enfin, d'exploitation et de suivi (Graphique 1.5. Cycle de vie d'un système d'IA). Ces phases présentent souvent un caractère itératif et ne respectent pas nécessairement un ordre séquentiel. La décision de mettre un terme à l'utilisation d'un système d'IA peut intervenir à n'importe quel stade de la phase d'exploitation et de suivi.

Les phases du cycle de vie d'un système d'IA peuvent être décrites comme suit :

1. La phase de **conception, données et modèles** comprend plusieurs activités, dont l'ordre peut varier selon les systèmes d'IA.
    - La **planification et la conception** du système d'IA couvrent la définition du concept et des objectifs du système, des principes sous-jacents, du contexte et du cahier des charges, ainsi que la construction éventuelle d'un prototype.
    - La **collecte et le traitement des données** englobent les tâches visant à recueillir et nettoyer les données, réaliser les vérifications d'exhaustivité et de qualité, et documenter les métadonnées et les caractéristiques de l'ensemble de données. Les métadonnées intègrent les informations relatives aux modalités de création de l'ensemble de données, à sa composition, aux usages prévus et à sa maintenance au fil du temps.
    - La **construction et l'interprétation du modèle** couvrent la création ou le choix des modèles ou des algorithmes, leur calage et/ou leur entraînement, ainsi que leur interprétation.
2. La phase de **vérification et validation** comprend l'exécution et le réglage des modèles, avec des tests visant à évaluer les performances au regard de diverses dimensions et considérations.

3. La phase de **déploiement** (mise en production) englobe le pilotage, la vérification de la compatibilité avec les systèmes existants, la mise en conformité réglementaire, la gestion des changements organisationnels et l'évaluation de l'expérience des utilisateurs.
4. La phase d'**exploitation et** de **suivi** couvre l'exploitation du système d'IA et l'évaluation permanente de ses recommandations et de ses effets (attendus et imprévus) au regard des objectifs et des considérations éthiques. C'est au cours de cette phase que l'on identifie les problèmes, opère les ajustements en revenant aux autres phases, voire, si nécessaire, abandonne la production du système d'IA.

Graphique 1.5. Cycle de vie d'un système d'IA



Source : Tel que défini et approuvé par l'AIGO en février 2019.

De par la centralité des données et des modèles dont l'entraînement et l'évaluation dépendent des données, le cycle de vie de nombreux systèmes d'IA se distingue du cycle traditionnel de développement des systèmes. Certains systèmes d'IA faisant appel à l'apprentissage automatique peuvent fonctionner par itérations et évoluer au fil du temps.

Recherche en matière d'IA

Cette section passe en revue certaines évolutions techniques qui ont marqué la recherche en matière d'intelligence artificielle dans les secteurs universitaire et privé et favorisent la transition vers l'IA. L'IA, en particulier sa branche dénommée « apprentissage automatique », est devenue un domaine de recherche active de l'informatique. Un nombre croissant de disciplines universitaires mettent à profit les techniques d'IA pour un large éventail d'applications.

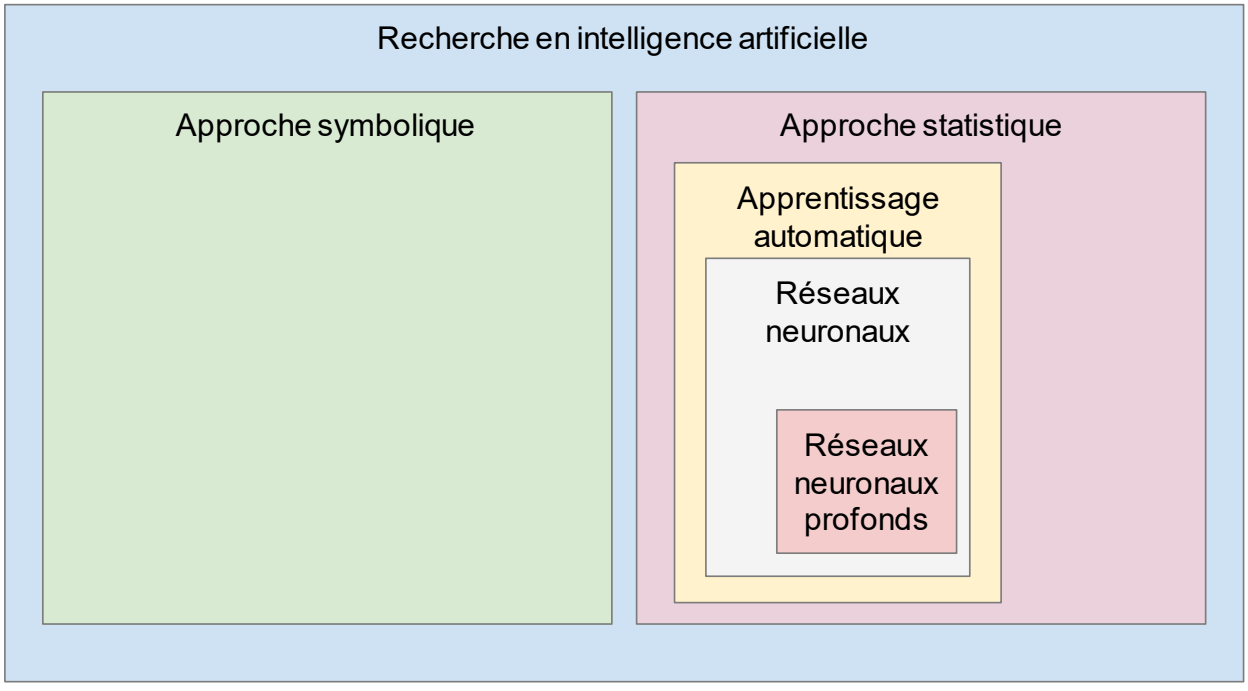
Il n'existe pas de système de classification communément admis pour la répartition des activités d'IA en domaines de recherche, à l'image par exemple des 20 grandes catégories de recherche économique du système de classification *Journal of Economic Literature*. Cette section entend proposer une taxinomie de recherche sur l'IA ayant vocation à aider les décideurs à décrypter certaines tendances récentes de l'IA et identifier les enjeux en termes d'action des pouvoirs publics.

Par le passé, les travaux de recherche ont opéré une distinction entre l'IA symbolique et l'IA statistique. L'IA symbolique s'appuie sur des représentations logiques pour aboutir à une conclusion à partir d'un ensemble de contraintes. Elle exige que les chercheurs construisent des structures décisionnelles détaillées, compréhensibles par l'homme, pour traduire la complexité du monde réel et aider les machines à parvenir à des décisions semblables à celles des humains. L'IA symbolique est encore aujourd'hui couramment utilisée, par exemple pour l'optimisation et la planification des outils. L'IA statistique, qui

permet aux machines d'inférer une tendance à partir de schémas, connaît depuis peu un engouement grandissant. Un certain nombre d'applications allient les approches symbolique et statistique. Ainsi, il n'est pas rare que des algorithmes de traitement du langage naturel conjuguent des approches statistiques (qui exploitent d'importants volumes de données) et des approches symboliques (qui tiennent compte de considérations comme les règles de grammaire). La combinaison de modèles s'appuyant à la fois sur les données et sur l'expertise humaine pourrait aider à lever les contraintes liées à chacune des deux approches.

Les systèmes d'IA font un usage croissant de l'apprentissage automatique. Il s'agit là d'un ensemble de techniques permettant aux machines d'apprendre de manière automatisée à partir de schémas et d'inductions, plutôt qu'en suivant les instructions explicites d'un humain. Les approches fondées sur l'apprentissage automatique entraînent souvent les machines à atteindre un résultat en leur présentant de nombreux exemples de résultats corrects. Toutefois, ils peuvent aussi définir un ensemble de règles et laisser la machine apprendre par essai et erreur. L'apprentissage automatique sert généralement à construire ou ajuster un modèle, mais pas seulement : il peut également être utilisé pour en interpréter les résultats (Graphique 1.6). Il fait appel à de nombreuses techniques employées depuis des décennies par des économistes, des chercheurs et des technologues. Ces techniques vont des régressions linéaires et logistiques aux arbres de décision, en passant par l'analyse en composantes principales, sans oublier les réseaux neuronaux profonds.

Graphique 1.6. Relation entre l'IA et l'apprentissage automatique



Source : Fourni par le programme *Internet Policy Research Initiative* (IPRI) du *Massachusetts Institute of Technology* (MIT).

En économie, les modèles de régression utilisent des données d'entrée pour établir des prévisions de telle sorte que les chercheurs puissent interpréter les coefficients (pondérations) appliqués aux variables d'entrée, souvent dans une optique d'action publique. Avec l'apprentissage automatique, il arrive que les humains ne soient pas en mesure de comprendre les modèles eux-mêmes. De plus, les problèmes d'apprentissage automatique tendent à recourir à un nombre bien plus important de variables que celles utilisées couramment en

économie. Ces variables, nommées « caractéristiques », se chiffrent généralement en milliers, voire plus. Des ensembles de données plus volumineux peuvent aller de dizaines de milliers à des centaines de millions d'observations. À une telle échelle, les chercheurs ont recours à des techniques plus sophistiquées et plus méconnues, comme les réseaux neuronaux, pour établir des prévisions. Il est intéressant de noter que l'un des domaines fondamentaux de recherche en matière d'apprentissage automatique tente de réintroduire le type d'explicabilité employé par les économistes dans ces modèles à grande échelle (voir Volet 4 ci-après).

La véritable technologie derrière la vague actuelle d'applications d'apprentissage automatique correspond à une technique de modélisation statistique perfectionnée, dénommée « réseaux neuronaux ». Elle s'accompagne d'une augmentation de la puissance de calcul et d'une disponibilité accrue d'ensembles de données colossaux (les « données massives »). Les réseaux neuronaux établissent des interconnexions répétées entre des milliers, voire des millions de transformations simples pour aboutir à une machine statistique plus importante, capable d'apprendre des relations élaborées entre les entrées et les sorties. En d'autres termes, ils modifient leur propre code pour trouver et optimiser les liens entre les entrées et les sorties. L'apprentissage profond désigne quant à lui des réseaux neuronaux particulièrement volumineux ; aucun seuil n'est défini pour déterminer à quel stade un réseau neuronal devient « profond ».

Cette dynamique évolutive de la recherche en matière d'IA va de pair avec des progrès constants en termes de capacités de calcul, de disponibilité des données et de conception des réseaux neuronaux. Sous leurs effets conjugués, l'approche statistique de l'IA devrait continuer de jouer un rôle important dans la recherche sur l'IA à court terme. Par conséquent, les décideurs devraient concentrer leur attention sur les évolutions de l'IA qui sont susceptibles d'avoir les incidences les plus marquées au cours des années à venir et représentent certains des défis les plus difficiles à surmonter pour les pouvoirs publics. Au nombre de ces défis figurent le décryptage des décisions des machines et le renforcement de la transparence du processus décisionnel. Les décideurs devraient également garder à l'esprit que les approches de l'IA les plus dynamiques – l'IA statistique et plus particulièrement les réseaux neuronaux – ne sont pas adaptées à tous les types de problèmes. D'autres approches, ainsi que le couplage des méthodes symbolique et statistique, restent importantes.

Il n'existe pas de taxinomie largement admise de la recherche sur l'IA ou de l'apprentissage automatique. La taxinomie exposée dans la sous-section suivante couvre 25 axes de recherche sur l'IA. Ils sont organisés en quatre grandes catégories (ou volets) et neuf sous-catégories, principalement centrées sur l'apprentissage automatique. S'il arrive que les chercheurs en économie se penchent sur un domaine de recherche restreint, les chercheurs en IA, pour leur part, travaillent généralement sur différents volets simultanément dans le but de résoudre des problématiques de recherche ouvertes.

Volet 1 : Applications d'apprentissage automatique

La première grande catégorie de recherche a trait à la mise en œuvre des méthodes d'apprentissage automatique pour résoudre diverses problématiques pratiques touchant l'économie et la société. L'émergence des applications d'apprentissage automatique est comparable à la façon dont l'accès à l'internet a commencé par transformer certains secteurs, avant de déferler sur le reste de l'économie. Le chapitre 3 expose différents exemples d'applications d'IA qui voient le jour dans les pays de l'OCDE. Les axes de recherche figurant dans le Tableau 1.1 représentent les principaux domaines de recherche liés au développement d'applications pour le monde réel.

Les domaines fondamentaux de recherche appliquée qui utilisent l'apprentissage automatique vont du traitement du langage naturel à la vision par ordinateur, en passant par la navigation robotique. Chacune de ces trois disciplines représente un champ de recherche riche et en expansion. Les problématiques de recherche peuvent être limitées à un seul domaine ou

couvrir plusieurs axes. Par exemple, aux États-Unis, des chercheurs allient, d'une part, le traitement du langage naturel pour des mammographies et des notes de pathologie en texte libre et, d'autre part, la vision par ordinateur des mammographies afin d'aider au dépistage du cancer du sein (Yala et al., 2017[12]).

Tableau 1.1. Volet 1 : Domaines d'application

Domaines d'application	Utilisation de l'apprentissage automatique	Traitement du langage naturel
		Vision par ordinateur
		Navigation robotique
		Apprentissage des langues
	Contextualisation de l'apprentissage automatique	Théorie algorithmique des jeux et choix social computationnel
		Systèmes collaboratifs

Source : Fourni par le programme IPRI du MIT.

Deux lignes de recherche étudient les moyens de contextualiser l'apprentissage automatique. La théorie algorithmique des jeux se situe à l'intersection de l'économie, de la théorie des jeux et de l'informatique. Elle utilise les algorithmes pour analyser et optimiser des jeux sur plusieurs périodes. Les systèmes collaboratifs permettent une approche des grands défis où plusieurs systèmes d'apprentissage automatique s'associent pour traiter différentes parties de problèmes complexes.

Volet 1: Intérêt pour l'action des pouvoirs publics

Plusieurs questions relevant des pouvoirs publics sont liées aux applications d'IA. Tel est le cas notamment de l'avenir du travail, des incidences potentielles de l'IA, du développement du capital humain et des compétences, ou encore de la compréhension des situations dans lesquelles le recours à des applications de l'IA pourrait ou non s'avérer adapté dans des contextes sensibles. À cela s'ajoutent les répercussions de l'IA sur les acteurs et la dynamique de l'industrie, les politiques en matière de données publiques ouvertes, la réglementation de la navigation robotique et les politiques en faveur de la protection de la vie privée qui régissent la collecte et l'utilisation des données.

Volet 2 : Techniques d'apprentissage automatique

La deuxième grande catégorie de recherche porte sur les techniques et paradigmes utilisés dans le domaine de l'apprentissage automatique. Cette ligne de recherche, similaire aux travaux de recherche sur les méthodes quantitatives dans les sciences sociales, développe et fournit les outils techniques et les approches employés dans les applications d'apprentissage automatique (Tableau 1.2).

Tableau 1.2. Volet 2 : Techniques d'apprentissage automatique

Techniques d'apprentissage automatique	Techniques	Apprentissage profond
		Apprentissage par simulation
		Production participative et calcul humain
		Calcul évolutif
		Techniques au-delà des réseaux neuronaux
	Paradigmes	Apprentissage supervisé
		Apprentissage par renforcement
		Modèles génératifs/réseaux antagonistes génératifs

Source : Fourni par le programme IPRI du MIT.

Cette catégorie est dominée par les réseaux neuronaux (dont l'apprentissage profond est une sous-catégorie) et constitue aujourd'hui le socle de la majeure partie de l'apprentissage automatique. Les techniques d'apprentissage automatique intègrent également divers paradigmes utilisés pour aider les systèmes à apprendre. L'apprentissage par renforcement entraîne le système d'une façon qui imite l'apprentissage humain, par essai et erreur. Plutôt que d'attribuer des tâches explicites aux algorithmes, ceux-ci apprennent en essayant différentes options qu'ils enchaînent à un rythme rapide. Ils s'adaptent alors en fonction des résultats, qui prennent la forme de récompenses et de pénalités. Certains parlent d'« expérimentation sans relâche » (Knight, 2017[13]).

Les modèles génératifs, notamment les réseaux antagonistes génératifs, entraînent un système à produire de nouvelles données à l'image d'un ensemble de données existant. Ces réseaux constituent un domaine de recherche sur l'IA intéressant car ils mettent en compétition au moins deux réseaux neuronaux non supervisés dans un jeu à somme nulle. En théorie des jeux, cela signifie qu'ils fonctionnent et apprennent comme une suite de jeux répétés à un rythme rapide. Les systèmes ainsi mis en compétition et dotés de vitesses de calcul élevées peuvent apprendre des stratégies utiles. Ils sont particulièrement adaptés aux environnements structurés assortis de règles claires, comme le jeu de Go, avec AlphaGo Zero.

Volet 2: Intérêt pour l'action des pouvoirs publics

Plusieurs questions relevant des pouvoirs publics sont liées au développement et au déploiement des technologies d'apprentissage automatique. Citons notamment le soutien en faveur de l'amélioration des ensembles de données d'entraînement ; le financement de la recherche universitaire et de la science fondamentale ; les politiques visant à former des personnes dotées de doubles spécialités, à savoir disposant de compétences à la fois en matière d'IA et dans une autre discipline ; et l'enseignement informatique. Par exemple, le financement de la recherche par le gouvernement canadien a permis des avancées qui ont conduit au formidable succès des réseaux neuronaux modernes (Allen, 2015[14]).

Volet 3 : Solutions d'amélioration de l'apprentissage automatique/optimisations

La troisième grande catégorie de recherche porte sur les moyens d'améliorer et d'optimiser les outils d'apprentissage automatique. Les axes de recherche y sont décomposés selon l'horizon temporel des résultats (actuels, émergents et futurs) (Tableau 1.3). La recherche à court terme est axée sur l'accélération du processus d'apprentissage profond, soit en améliorant la collecte des données, soit en utilisant des systèmes informatiques distribués pour entraîner les algorithmes.

Les chercheurs étudient comment équiper de fonctions d'apprentissage automatique des appareils à faible puissance comme des téléphones mobiles et d'autres appareils connectés. Des progrès significatifs ont été réalisés sur ce front. Des projets, à l'instar de la *Teachable Machine* de Google, offrent désormais des outils à code source libre suffisamment légers pour fonctionner avec un simple navigateur (Encadré 1.2). Le projet *Teachable Machine* n'est qu'un exemple parmi d'autres d'outils émergents de développement de l'IA destinés à étendre le rayonnement et accroître l'efficience de l'apprentissage automatique. À cela s'ajoutent des avancées significatives dans le développement de puces d'IA dédiées aux appareils mobiles.

La recherche dans l'apprentissage automatique à plus long terme porte sur l'étude des mécanismes permettant aux réseaux neuronaux d'apprendre efficacement. Bien que les réseaux neuronaux se soient révélés être une puissante technique d'apprentissage automatique, la compréhension de leur mode de fonctionnement reste limitée. Mieux appréhender ces processus permettrait de concevoir des réseaux neuronaux plus profonds. La recherche à plus long terme

s'intéresse également aux moyens d'entraîner les réseaux neuronaux à utiliser des ensembles de données d'entraînement plus réduits, une technique parfois dénommée « apprentissage à partir d'un exemple unique » (*one-shot learning*). Qui plus est, on cherche généralement à améliorer l'efficience du processus d'entraînement. De fait, les modèles de grande envergure peuvent nécessiter des semaines voire des mois d'entraînement et requièrent des centaines de millions d'exemples.

Tableau 1.3. Volet 3 : Solutions d'amélioration de l'apprentissage automatique/optimisations

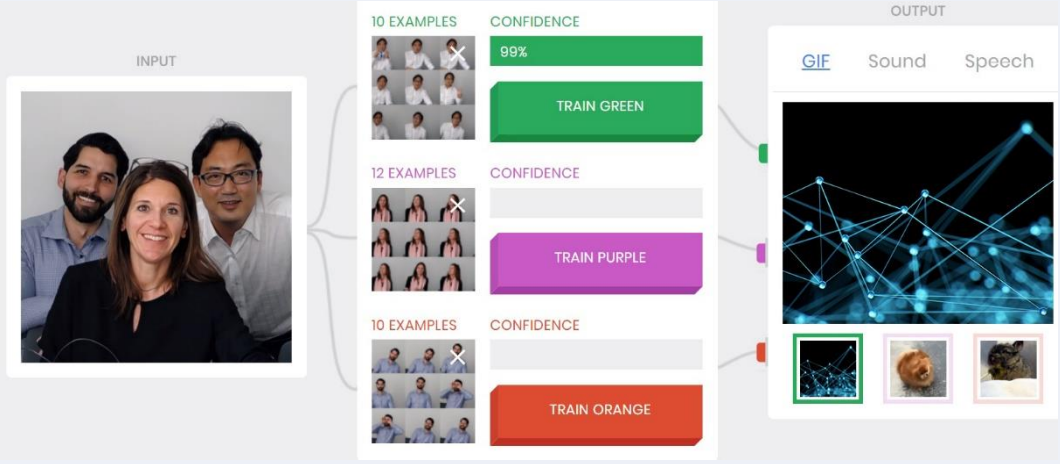
Solutions pour améliorer l'apprentissage automatique	Facteurs favorables (actuels)	Accélération de l'apprentissage profond
		Amélioration de la collecte des données
		Systèmes distribués pour l'entraînement des algorithmes
	Facteurs favorables (émergents)	Performances sur des appareils à faible puissance
		Apprendre à apprendre/méta-apprentissage
		Outils de développement de l'IA
	Facteurs favorables (futurs)	Comprendre les réseaux neuronaux
		Apprentissage à partir d'un exemple unique

Source : Fourni par le programme IPRI du MIT.

Encadré 1.2. Projet *Teachable Machine*

Le projet *Teachable Machine* est une expérience menée par Google permettant aux utilisateurs d'entraîner une machine à détecter différents scénarios à l'aide de l'appareil photo d'un téléphone ou de la caméra d'un ordinateur. Pour ce faire, l'utilisateur prend une série de photos dans trois situations distinctes (par exemple, trois expressions faciales). La machine analyse alors les photos au sein de l'ensemble de données d'entraînement et peut les utiliser pour détecter différents scénarios. Elle peut ainsi émettre un son à chaque fois que l'utilisateur sourit dans le champ de la caméra. Le projet *Teachable Machine* se démarque des autres projets d'apprentissage automatique par le fait que le réseau neuronal fonctionne exclusivement via le navigateur de l'utilisateur, sans avoir besoin de recourir à des calculs externes ni au stockage des données (Graphique 1.7).

Graphique 1.7. Entraînement d'une machine à l'aide de la caméra d'un ordinateur



Source : https://experiments.withgoogle.com/ai/teachable-machine.

Volet 3: Intérêt pour l'action des pouvoirs publics

Les questions liées au troisième volet intéressant les pouvoirs publics tiennent notamment aux incidences de l'utilisation de l'apprentissage automatique sur des appareils autonomes, sans impliquer nécessairement de partage de données dans le nuage. Autres sujets d'intérêt : les perspectives de réduction de la consommation d'énergie et la nécessité de développer des outils d'IA plus perfectionnés afin d'en optimiser les usages bénéfiques.

Volet 4 : Prise en compte du contexte sociétal

La quatrième grande catégorie de recherche examine le contexte technique, juridique et social dans lequel s'inscrit l'apprentissage automatique. Les systèmes d'apprentissage automatique s'appuient de plus en plus fréquemment sur des algorithmes pour prendre des décisions majeures. D'où l'importance de comprendre comment des biais peuvent être introduits, comment ils peuvent se propager et comment les éliminer des résultats. L'un des domaines de recherche les plus actifs en matière d'apprentissage automatique concerne la transparence et la responsabilité dans le cadre des systèmes d'IA (Tableau 1.4). Les approches statistiques de l'IA ont conduit à l'utilisation de calculs moins compréhensibles par l'homme pour la prise de décisions algorithmiques. Or ces dernières peuvent avoir des incidences non négligeables sur la vie des individus – qu'il s'agisse de prêts bancaires ou de décisions de libération conditionnelle de prisonniers (Angwin et al., 2016[15]). Les étapes visant à assurer la sécurité et l'intégrité de ces systèmes sont un autre type de recherche tenant compte du contexte. Les chercheurs commencent à peine à entrevoir de quelle façon les réseaux neuronaux parviennent à leurs décisions. Les réseaux peuvent souvent être piégés à l'aide de méthodes simples, par exemple en changeant quelques pixels sur une photo (Ilyas et al., 2018[16]). Ces axes de recherche visent à protéger les systèmes de l'introduction intempestive d'informations indésirables et d'attaques. Le but est également de vérifier l'intégrité des systèmes d'apprentissage automatique.

Volet 4: Intérêt pour l'action des pouvoirs publics

Plusieurs questions relevant des pouvoirs publics sont liées au contexte dans lequel s'inscrit l'apprentissage automatique. Citons notamment les exigences de responsabilité algorithmique, la lutte contre les biais, les incidences des systèmes d'apprentissage automatique, la sécurité des produits, la responsabilité des personnes et la sécurité des systèmes (OCDE, 2019[8]).

Tableau 1.4. Volet 4 : Affiner l'apprentissage automatique en tenant compte du contexte

Affiner l'apprentissage automatique en tenant compte du contexte	Explicabilité	Transparence et responsabilité
		Explication des décisions individuelles
		Simplification afin de tendre vers des algorithmes compréhensibles par l'homme
		Équité/biais
		Capacité de débogage
	Sécurité et fiabilité	Exemples de risques
		Vérification
		Autres classes d'attaques

Source : Fourni par le programme IPRI du MIT.

Références

Allen, K. (2015), « How a Toronto professor's research revolutionized artificial intelligence, », *The Star,* 17 April, https://www.thestar.com/news/world/2015/04/17/how-a-toronto-professors-research-revolutionized-artificial-intelligence.html. [14]

Angwin, J. et al. (2016), « Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks », *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [15]

Anyoha, R. (28 août 2017), « The history of artificial intelligence », Harvard University Graduate School of Arts and Sciences Blog, 28 août 2017, http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/. [4]

Gringsjord, S. et N. Govindarajulu (2018), *Artificial Intelligence*, The Stanford Encyclopedia of Philosophy Archive, https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/. [11]

Ilyas, A. et al. (2018), *Blackbox Adversarial Attacks with Limited Queries and Information*, exposé présenté à la 35e Conférence internationale sur l'apprentissage automatique, 2018, Stockholmsmässan Stockholm, du 10 au 15 juillet 2018, pp. 2142–2151. [16]

Knight, W. (2017), « 5 big predictions for artificial intelligence in 2017 », *MIT Technology Review,* 4 janvier, https://www.technologyreview.com/s/603216/5-big-predictions-for-artificial-intelligence-in-2017/. [13]

OCDE (2019), *Recommandation du Conseil sur l'intelligence artificielle*, OCDE, Paris, https://legalinstruments.oecd.org/api/print?ids=648&lang=fr. [8]

OCDE (2018), *Perspectives de l'économie numérique de l'OCDE 2017*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264282483-fr. [7]

Russel, S. et P. Norvig (2009), *Artificial Intelligence: A Modern Approach, 3ème édition*, Pearson, http://aima.cs.berkeley.edu/. [9]

Shannon, C. (1950), « XXII. Programming a computer for playing chess », *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, vol. 41/314, pp. 256-275. [2]

Silver, D. et al. (2017), « Mastering the game of Go without human knowledge », *Nature*, vol. 550/7676, pp. 354-359, http://dx.doi.org/10.1038/nature24270. [6]

Somers, J. (2013), « The man who would teach machines to think », *The Atlantic,* November, https://www.theatlantic.com/magazine/archive/2013/11/the-man-who-would-teach-machines-to-think/309529/. [5]

Turing, A. (1950), « Computing machinery and intelligence », dans *Parsing the Turing Test*, Springer, Dordrecht, pp. 23-65. [1]

UW (2006), *History of AI*, University of Washington, History of Computing Course (CSEP 590A), https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf. [3]

Winston, P. (1992), *Artificial Intelligence*, Addison-Wesley, Reading, MA, https://courses.csail.mit.edu/6.034f/ai3/rest.pdf. [10]

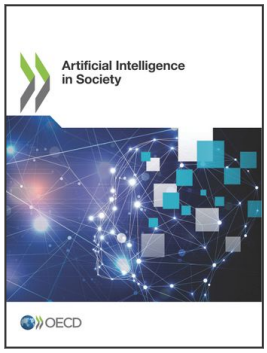
Yala, A. et al. (2017), « Using machine learning to parse breast pathology reports », *Breast Cancer Research and Treatment*, vol. 161/2, pp. 201-211. [12]

Notes

[1] Ces tests ont été réalisés à l'aide de messages saisis ou relayés, et non par la voix.

[2] Des travaux sur le cycle de développement d'un système ont été menés, entre autres, par le *National Institute of Standards*. Plus récemment, des organisations de normalisation comme l'Organisation internationale de normalisation (ISO), par le biais de son sous-comité SC 42, ont commencé à se pencher sur le cycle de vie des systèmes d'IA.

[3] L'initiative mondiale sur l'éthique dans la conception de systèmes autonomes et intelligents (*Global Initiative on Ethics of Autonomous and Intelligent Systems*) de l'*Institute of Electrical and Electronics Engineers* (IEEE), en est un exemple.



Extrait de :
Artificial Intelligence in Society

Accéder à cette publication :
https://doi.org/10.1787/eedfee77-en

Merci de citer ce chapitre comme suit :

OCDE (2019), « Paysage technique de l'IA », dans *Artificial Intelligence in Society*, Éditions OCDE, Paris.

DOI: https://doi.org/10.1787/78af3a32-fr