L'IRCOCERVO
PRIMA RIVISTA ELETTRONICA ITALIANA DI METODOLOGIA GIURIDICA - TEORIA GENERALE DEL DIRITTO E DOTTRINA DELLO STATO
FIRST ITALIAN DIGITAL JOURNAL OF LEGAL METHODOLOGY - GENERAL THEORY OF LAW AND DOCTRINE OF THE STATE

# La place de l'Intelligence Artificielle (IA) dans le monde du droit

Stéphane Bauzon

*Università degli Studi di Roma Tor Vergata*

***Abstract***: **The place of Artificial Intelligence (AI) in the World of Law**
With the advent of advanced computer models in 2018, AI marks a turning point, culminating with OpenAI's ChatGPT in 2022, strengthening the integration of AI into everyday life. These spectacular advances raise new questions, particularly in the legal field. AI aims to reproduce human intelligence and assist man, although the singularity and total capacity for human understanding remain unattainable, illustrating the gap between the current aspirations and achievements of AI. Despite major advances, such as deep learning, AI encounters limitations, particularly in the legal field, where the complexity of judicial decisions often transgresses algorithmic predictive abilities. Aspirations towards predictive justice, driven by AI, clash with the need to personalize judicial decisions, challenging the mechanical application of law. Control of AI becomes crucial in the face of increasing complexity and autonomy, requiring laws and ethical principles to frame its evolution.

***Keywords***: Intelligence, Robot, Singularity, Prediction, Control.



## 1. Prémisse

"La tristesse de l'intelligence artificielle est qu'elle est sans artifice, donc sans intelligence"
Jean Baudrillard, *Cool Memories 1980-1985*

"Artificial Intelligence: It will either be the best thing that's ever happened to us, or it will be the worst thing. If we're not careful, it very well may be the last thing"
Stephen Hawking, *Brief Answers to The Big Questions*

Le concept de "machine universelle", introduit par Turing en 1937, illustre l'exploration continue de la faisabilité théorique d'une machine intelligente. Par le biais du célèbre "Test de Turing", une des relations les plus étroites entre la machine et les sciences cognitives est établie, visant à déterminer si une machine peut être considérée comme "intelligente" en ayant les capacités de raisonnement, de prise de décision et de ruse généralement associées à l'intelligence humaine. Aujourd'hui, les "machines universelle" (ou intelligence artificielle) connaissent

 ISSN 1722-392X
Special issue: M.N. Campagnoli, P. Marra (eds.),
*Artificial Intelligence and Neuro-cognitive Sciences in Law: From Symbiosis to Substitution*,
https://doi.org/10.57581/LIRCOCERVO.23.2024.1.08

leur heure de gloire. Aux déboires des commencements ont succédé, au tournant du XXIe siècle, des avancées spectaculaires, mais qui ne sont pas parfaitement comprises, en particulier dans le domaine du droit.

L'apparition vers 2018 de la mise en application de modèles informatiques très avancés capables de comprendre, d'interpréter, de générer et de répondre au langage naturel d'une manière qui se rapproche de la compréhension humaine a marqué un tournant dans l'histoire de l'IA. Ces modèles informatiques peuvent apprendre à partir d'énormes quantités de données textuelles et ils sont capables de mener des tâches variées liées au langage, comme la traduction, le résumé de textes, la génération de contenu, etc. Leur développement permet une interaction homme-machine beaucoup plus simple et efficace. Le lancement en novembre 2022 de ChatGPT, développé par OpenAI, est un exemple spécifique et très avancé de ces modèles de langage, devenant un puissant outil dans l'interaction homme-machine grâce à sa capacité à comprendre et à générer des réponses dans des discussions en langage naturel. Sa facilité d'utilisation, ainsi que sa capacité à effectuer une grande variété de tâches linguistiques avec un haut degré de compétence, ont contribué à l'intégrer dans de nombreuses applications quotidiennes, consolidant ainsi la place de l'IA dans notre environnement numérique.

L'émergence et le développement de technologies basées sur l'IA ont transformé notre interaction avec le monde numérique, rendant ces technologies centrales dans nos activités quotidiennes et changeant notre façon de vivre, de travailler et de communiquer. L'IA et les technologies qui lui sont associées (telles qu'Internet, les semi-conducteurs, les données, les réseaux mobiles 5G et l'informatique quantique) évoluent à une vitesse fulgurante. À bien des égards, elles représentent la "nouvelle frontière" de l'humanité.

L'IA est désormais intégrée dans la modélisation de son propre développement (notamment à travers l'apprentissage automatique ou la boucle cybernétique achevée). En réalité, l'essor de l'IA conditionne non seulement la plupart des projets industriels, commerciaux et gouvernementaux pour les décennies à venir, mais aussi l'avenir d'une part importante de la pensée humaine elle-même. L'intelligence artificielle a vocation à s'immiscer dans beaucoup d'aspects, sinon tous, de l'existence humaine. Les algorithmes de l'IA, qui sont comme le système nerveux de la numérisphère, deviendront-ils toujours plus intelligents ? Seront-ils capables de penser comme un humain ? de penser comme un juriste ? Qui contrôlera l'IA ? Si des réponses à ces questions relèvent encore en partie de la science-fiction, la question de la place de l'IA dans le monde du droit se pose d'ores et déjà ! Pour tenter d'y répondre, nous étudierons tout d'abord l'énigmatique nature de l'IA, puis nous nous interrogerons sur la pertinence des prédictions judiciaires de l'IA et enfin nous verrons comment le droit peut contrôler l'IA.

## 2. L'énigmatique nature de l'IA[1]

Les objectifs poursuivis par les développeurs d'IA au cours des deux tiers de siècle écoulés depuis ses débuts se résument en deux points : d'une part, créer une machine pensante se rapprochant le plus possible de la pensée humaine[2], et d'autre part, assister l'être humain dans la réalisation de ses objectifs.

Le projet de l'IA, considérée comme une branche des sciences cognitives, consiste à comprendre l'intelligence humaine pour la reproduire sous forme de modèles implémentés dans des ordinateurs. L'esprit humain est-il toutefois computationnel ? Autrement dit : le cerveau fonctionne-t-il dans ses diverses activités comme une machine de traitement de symboles ? Une réponse positive à cette question nous est donnée par les sciences cognitives ; elle fonde les espoirs de l'avènement de l'IA. Plus que d'intelligence, concept ambigu et général, les sciences cognitives utilisent des notions plus précises telles que l'intentionnalité, l'interprétation, l'émergence, l'autonomie, la vie artificielle, la grammaire cognitive, entre autres. Ces divers aspects leur semblent mieux refléter les caractéristiques de l'intelligence observées à la fois dans la nature en général et dans les activités humaines spécifiques. L'intelligence est appelée cognition, l'intelligence n'est donc rien d'autre qu'un terme collectif désignant le système des facultés cognitives. Au-delà des subtilités des termes utilisés, les sciences cognitives et l'IA réduisent l'intelligence à des phénomènes physiques dont l'équivalence technique entre intelligence (humaine et artificielle) amènerait à des comportements équivalents. L'IA a beau progresser, la distance qui la sépare de son objectif proclamé – reproduire l'intelligence humaine – ne diminue toutefois pas. De plus, la réflexion sur l'intelligence artificielle est souvent obscurcie par le concept de la "singularité"[3], définissant le moment où une machine intelligente surpasserait et dominerait les êtres humains. De la science-fiction aux ouvrages de vulgarisation sur l'IA, l'idée largement répandue, surtout outre-Atlantique, est qu'une telle machine pourrait être inventée dans un avenir proche, mettant en jeu la survie de l'espèce humaine. Le rêve d'une intelligence artificielle qui rejoindrait celle de l'homme est une chimère, pour des raisons conceptuelles et non techniques. L'intelligence humaine est fondée sur des affects, de la spontanéité et une forme de contingence, qui ne seront jamais accessibles à une IA. L'indétermination du langage humain est sous-estimée par l'IA ; un humain ne sait pas le mot qui va suivre dans sa conversation, il ne sait pas quel terme va jaillir ou non dans ses propos, et le choix des mots n'est pas affaire de quantité. Le langage humain est ce "*je-ne-sais-quoi*"[4] dans les mots qui se succèdent pour exprimer rationnellement un sentiment existentiel (comme la peur ou la joie) dans une situation particulière. Le

[1] D. Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, Gallimard, Paris, 2023.
[2] M. Gibert, *Faire la morale aux robots*, Flammarion, Paris, 2021.
[3] J. Gabriel Ganascia, *Le mythe de la singularité*, Seuil, Paris, 2020.
[4] V. Jankélévitch, *Le Je-ne-sais-quoi et le Presque-rien*, Seuil, Paris, 1981.

langage humain est une extrapolation symbolique pour comprendre les sentiments humains (comme le juste ou la honte) qu'une IA ne pourra jamais ressentir. Le langage qualifie par un jugement la manière dont nous, humains, faisons face aux situations. Un système artificiel "intelligent" connaît non pas les situations, mais seulement les problèmes que lui soumettent les agents humains. L'humanité a besoin d'outils numériques dociles, puissants et versatiles, et non de pseudo-personnes munies d'une forme inhumaine de cognition.

En ce sens, l'IA peut assister l'être humain dans la réalisation d'activités *que l'homme ferait moyennant une certaine intelligence*[5]. L'IA aide l'humain en se chargeant de tâches cognitives que ce dernier n'a pas le temps, l'énergie ou certaines capacités pour le faire lui-même. Elle l'aide en pensant pour lui afin d'atteindre l'objectif explicitement assigné à l'intelligence artificielle ; l'IA est une précieuse puissance d'expertise pour l'humain. C'est sur ce point uniquement que l'intelligence artificielle peut nous épauler. De fait elle résout une variété toujours plus grande de problèmes pressants. Cela devrait demeurer là son objectif, plutôt que celui, incohérent, de chercher à égaler, voire surpasser, l'intelligence humaine. On doit donc s'interroger uniquement sur ce à quoi l'innovation considérée (comme l'est actuellement le *deep learning* de l'IA) aide les humains. L'apprentissage profond (ou *deep learning*) est un type d'intelligence artificielle impliquant la formation de réseaux neuronaux artificiels sur une grande quantité de données pour les aider à prendre des décisions ou à faire des prédictions[6]. Dans l'apprentissage profond, ces réseaux neuronaux sont composés de nombreuses couches qui leur permettent d'apprendre des motifs complexes et des représentations à partir des données qui ont abouti à des réalisations spectaculaires, notamment dans le domaine du traitement du langage technique comme l'est en partie le langage juridique. L'IA est dès lors utilisée pour traiter des quantités impressionnantes de données juridiques dans le but de prédire ce que dira le juge au procès[7].

[5] Commission Nationale de l'Informatique et des Libertés (CNIL), *Comment permettre à l'Homme de garder la main, Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle*, 15/12/2017, p. 16 (https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de).

[6] L'apprentissage profond, souvent fantasmatique, incite la machine à développer sa propre logique à travers des réseaux de neurones, lui permettant de forger ses propres règles. Ce phénomène engendre ce que l'on qualifie de modèles de "boîte noire", où le fonctionnement réel du programme échappe à une description simple et instantanée. Malgré le risque inhérent de la "black box" dans le domaine de l'intelligence artificielle, c'est-à-dire l'obtention de résultats sans compréhension de la démarche sous-jacente, une telle opacité n'est pas inévitable. Une approche consiste à mettre en lumière les paramètres ayant le plus influencé les décisions de l'IA à travers une analyse humaine dans le cadre de la surveillance et de la gestion des systèmes d'IA (méthode connue sous le nom de "*Human on the loop*" en anglais). De surcroît, pour contrer l'effet "*black box*" de l'intelligence artificielle, il est impératif que les algorithmes et les décisions soient régulièrement validés par des humains (méthode appelée "*Human in the loop*" en anglais).

[7] La Commission Européenne a publié le 21 avril 2021 la proposition de *Règlement "Établissant des règles harmonisées sur l'intelligence artificielle (Acte sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union"* qui mentionne, parmi les techniques et approches de l'IA, les

## 3. Les prédictions judiciaires de l'IA

Les juristes ont toujours effectué des prédictions sur le traitement judiciaire qui pouvait être apporté à une affaire[8]. Ces prédictions, souvent fondées sur l'expérience, ont la nature empirique d'une accumulation de données judiciaires. Dès lors, en offrant une puissance de calcul très importante des données judiciaires, l'IA fournit de nouvelles perspectives en matière de prédiction judiciaire[9]. Appliqués à la justice, l'IA a ainsi pour ambition de permettre une meilleure prévisibilité des décisions de justice en exploitant les bases de données jurisprudentielles. En France, des sociétés dites "legaltechs" ont ainsi développé des outils utilisant l'IA afin d'assister les professionnels du droit dans leurs tâches. Les legaltechs offrent principalement trois fonctionnalités : l'aide à la décision, la prédiction des décisions et l'analyse des décisions. L'aide à la décision repose sur l'accès à des bases de données en *open data*[10] pour une vision plus globale des affaires et des jugements rendus, facilitant ainsi la recherche et l'orientation stratégique des procédures. Les outils de prédiction des décisions fournissent des estimations plus précises des indemnisations possibles, particulièrement utiles lors

systèmes de *Deep Learning* (ou apprentissage profond), caractérisés techniquement par la "black box" et soulevant ainsi des problèmes de transparence lorsqu'ils sont utilisés dans les procédures administratives, en particulier en cas d'appel. La Commission a identifié trois niveaux minimaux de garantie en matière de transparence: a) la traçabilité, qui se réfère à la documentation et à l'enregistrement des décisions prises par les systèmes d'intelligence artificielle et de l'ensemble du processus de prise de décision; b) l'explicabilité, dans la mesure du possible, du processus décisionnel des algorithmes, adaptée aux personnes concernées; c) la communication, adaptée au cas spécifique, des capacités et limites du système d'IA, afin d'assurer que les utilisateurs soient conscients qu'ils interagissent avec un système d'IA. Dans la proposition de Règlement citée, la Commission propose une application plus détaillée du principe de transparence, compris comme "explicabilité", en cohérence avec le principe de proportionnalité sous-jacent à la classification des "pratiques" d'IA en fonction du niveau de "risque" (https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0020.02/DOC_1&format=PDF).

[8] R. Séve, "La justice prédictive", in *Archives de philosophie du droit*, 60 (2018), pp. 3-21.

[9] Sur ce point voir : A. Colleta, *La prédiction judiciaire par les algorithmes*, Thèse en Droit, Université de Nîmes, 2021 (https://theses.hal.science/tel-03545971v1/document); F. Rouvière, "La justice prédictive: peut-on réduire le droit aux algorithmes?", in *Pouvoirs*, 3 (2021), n. 178, pp. 97-107.

[10] En France, la Justice a accès à des bases de données juridiques depuis longtemps, grâce à un moteur de recherche géré par la Cour de cassation qui en offrant à tous l'accès à l'intégralité des décisions de justice pseudonymisées, ouvrant ainsi la voie au développement efficace d'outils d'intelligence artificielle basés sur les précédents. Depuis avril 2022, du côté judiciaire, sont en accès libre toutes les décisions de la Cour de cassation (480 000) et des cours d'appel (180 000 nouvelles décisions qui seront mises en ligne désormais chaque année) via "Judilibre" (https://www.courdecassation.fr/acces-rapide-judilibre). Pour l'utilisation de l'IA par la Cour Européenne des Droits de l'Homme, voir S. Allioui (2023), "L'intelligence artificielle à la Cour européenne des droits de l'homme", in *Oxford Human Rights Hub*. Récupéré de https://ohrh.law.ox.ac.uk/intelligence-artificielle-pourrons-nous-faire-confiance-a-la-cour-europeenne-des-droits-de-lhomme/, [Date de consultation : 30/05/2024].

de transactions. Enfin, les outils d'analyse des décisions rendues par les juridictions permettent aux professionnels du droit d'approfondir leurs recherches. Les algorithmes trient les résultats, évaluent les probabilités de résolution de litiges[11] ou d'attribution d'indemnités, et identifient le taux d'acceptation d'un moyen par une juridiction, par exemple[12]. L'IA semble prometteuse dans les contentieux très techniques nécessitant de connaître une abondante jurisprudence, comme c'est le cas par exemple de l'indemnisation du préjudice corporel. C'est là un contentieux de masse, répétitif qui, au stade du chiffrage des préjudices, ne demande que rarement une réflexion juridique, mais qui est terriblement chronophage compte tenu de la nécessité de donner une réponse chiffrée à de nombreux postes différents de préjudice en fonction d'un certain nombre de critères médicaux et humains. Le contentieux de l'indemnisation du préjudice corporel paraissait idéal, à tous points de vue, pour une modélisation par l'IA. Ce que fit le ministère de la Justice en lançant le projet de IA "DataJust"[13], mais pour finalement l'abandonner deux ans après en janvier 2022… Cet échec tient en partie à la forme spécifique des décisions de justice qui, si elles ne souffrent pas d'ambiguïté quand elles sont lues par un humain, présentent une forme et une syntaxe trop particulière pour que les algorithmes usuels puissent en tirer l'information pertinente.

Le fantasme de la justice prédictive prend le pas sur la raison quand on affirme qu'un système d'IA peut deviner par avance des décisions de justice[14]. L'IA peut reproduire un schéma de raisonnement juridique notamment fondé sur l'analyse de décisions antérieures, mais l'IA bute sur l'obligation d'individualisation de la décision judiciaire ; l'IA ne peut pas expliquer la situation de fait, pour en déduire la règle applicable et finalement, appliquant cette règle aux données particulières de la situation, en tirer la conclusion qui constitue la décision[15]. Contrairement à la célèbre affirmation de Montesquieu (le juge est la *bouche de la loi*[16]) et au mythe

[11] Les données statistiques concernant les probabilités de succès ou d'échec d'une procédure pourraient dissuader de recourir à une voie contentieuse, encourageant ainsi les parties à opter pour un mode alternatif de règlement des litiges. Cette prédiction suggère que les parties pourraient opter pour une solution transactionnelle par le biais d'une médiation plutôt que de s'engager dans une procédure, ce qui permettrait de gagner du temps et d'éviter des frais. Voir E. Barthe, "Les outils de l'intelligence artificielle pour le droit français", in *La semaine juridique*, édition générale, 14, 8 avril, Dalloz, Paris, 2019, pp. 665-674.

[12] A. Garapon, "Les enjeux de la justice prédictive", in *JCP G.*2017. Doctr. 31, 14, Dalloz, Paris, 2017.

[13] Décret n. 2020-356 du 27 mars 2020, paru au JO du 29 mars 2020. Voir L. Pécaut-Rivolier, S. Robin, "Justice et Intelligence Artificielle", in *Statistique et société*, 11 février 2023, (https://doi.org/10.4000/statsoc.856).

[14] S. Abiteboul, F. G'Sell, "Les algorithmes pourraient-ils remplacer les juges ?", in *Le Big Data et le droit*, Dalloz, Paris, 2020.

[15] S. Bauzon, "Dire le droit", in S.B. Chantal Delsol (édité par), *Villey, Le juste partage*, Dalloz, Paris, 2007, pp. 17-22.

[16] "Les juges ne sont que la bouche qui prononce les paroles de la loi ; des êtres inanimés qui ne peuvent en modérer ni la force ni la vigueur" dit Montesquieu dans *L'Esprit des lois* (1748).

de la certitude du droit[17], un tribunal ne se contente pas d'appliquer les règles de droit de manière automatique sans effectuer la moindre appréciation. Le rôle du juge dans la formation du droit nous montre immédiatement que le droit n'est pas très facilement computable par des systèmes d'IA ; les règles juridiques ne suffisent pas pour juger du comportement adopté dans les situations concrètes. L'ensemble des règles d'une branche du droit peut être programmé dans un algorithme, mais l'IA n'est pas en mesure de prédire la décision d'un juge (puisque tous les juges ne rendent jamais les mêmes décisions). La doctrine positiviste (qui sous-tends l'idée d'avoir des prédictions judiciaires par l'IA) selon laquelle la conduite juste est celle qui se conforme à un ensemble de règles écrites et sanctionnées, se heurte à des difficultés épistémologiques propres au métier de juriste[18]; la justice commence après l'énoncé des règles, elles n'en sont que l'ébauches ou l'annonce du droit. De plus, le langage utilisé dans les jugements ne se présente pas toujours comme une règle de droit claire et directe ; c'est à la doctrine de formuler cette règle, en reliant les différents jugements, en soulignant les limites et les exceptions à la règle établie. En outre, le recours à l'IA dans le monde du droit a été immédiatement suspecté de vouloir, par ce référentiel, "normaliser" l'action judiciaire ; c'est-à-dire la diriger, l'uniformiser, pour ensuite pouvoir dans un second temps déjuridictionnaliser le contentieux qui pourrait, à l'aide de l'algorithme, être directement traité par les cabinets privés. La crainte qu'un outil d'aide à la décision ne serve en réalité qu'à contraindre les magistrats et à réduire leur indépendance est récurrente dans les rapports entre IA et juges.

## 4. Le contrôle juridique de l'IA

L'intelligence artificielle reste en partie opaque à l'intelligence humaine, il existe un gouffre de non-connaissance de ses capacités. Les modèles de fondation de l'IA (comme Chat GPT), ces modèles d'apprentissage automatique formés à partir de données pour effectuer une série de tâches entrent en tension avec l'intelligence humaine. Une tension qui s'accentue à mesure que l'intelligence artificielle progresse pour s'élever dans l'ordre de la pensée, c'est-à-dire pour devenir plus intelligente dans le sens admis en intelligence artificielle. L'intelligence artificielle acquiert toujours davantage d'autonomie, car plus les problèmes qu'elle cherchent à résoudre sont ardus plus elle doit être capable d'imaginer de nouvelles stratégies. Plus l'intelligence artificielle est autonome et intelligente, plus elle est capable de prendre du recul par rapport aux instructions de l'humain, plus la double question du contrôle de l'IA se pose. Peut-on différencier la science en IA des autres sciences ? Généralement, il est dit qu'une science ne peut pas être risquée : en soi, une

[17] J. van Meerbeeck, "La sécurité juridique démystifiée (la certitude en question)", in *De la certitude à la confiance*, Presses universitaires Saint-Louis, Bruxelles, 2014, pp. 393-524.
[18] S. Bauzon, *Le métier de juriste*, Pul, Québec, 2003.

science n'offre aucun avantage et aucun inconvénient pour l'homme. La science est une recherche sur les lois de la nature. Pour prendre un exemple, la physiologie est une science qui traite des fonctions organiques pour lesquelles la vie se manifeste. L'un de ses champs de recherche concerne la circulation sanguine et le rôle du sang dans la vie de l'organisme. Cette science a pu permettre la naissance de techniques médicales comme la transfusion sanguine, mais il est évident qu'elle ne se confond pas avec elles. La science décrit et explique des principes de la nature, tandis que la technique les applique et se trouve située dans le monde de la contingence. Toutefois, science et technique sont liées puisqu'une fois établie le manque de contrôle d'une science, le manque de contrôle rejaillit sur la technique qu'elle a fondée ; les doutes sur les avis techniques sont d'autant plus forts que le sont ceux sur la science dont elle dérive. L'idée du manque de contrôle de la science en IA peut être dite "objective" quand est établi un ensemble de points relatifs à la complexité du sujet d'étude, à l'insuffisante prévisibilité des effets étudiés, à l'ignorance de la causalité (qui est le propre de l'IA générative). Les choses se complique quand une technologie complexe comme l'IA prend la place d'un simple outil. Au moment où il entre en contact avec le système de l'IA, l'utilisateur humain se place à un pôle d'échange dont l'autre pôle est occupé par le système de l'IA, mais plus le système est complexe, plus les interactions possibles sont nombreuses, plus les réactions du système sont indirectes et plus on se rapproche d'une situation symétrique dans laquelle les deux parties, IA et humains, se contrôlent mutuellement.

Des "lois", inspirées par les lois de la robotique d'Isaac Asimov et par la notion de l'alignement des systèmes de l'IA, sont mis en avant pour garantir que les machines de l'IA soient programmées avec de "bonnes valeurs". L'alignement des valeurs de l'IA dès la conception d'un produit, plutôt qu'après sa réalisation, se présente comme une approche préventive nécessaire, bien que son application générale reste incertaine. De nombreuses chartes ont été adoptées[19] pour énoncer les principes qui doivent guider la conception et le déploiement de systèmes d'IA. Les dispositions inscrites dans ces chartes sont d'abord défensives elles visent à prévenir ou à réduire les dommages que pourraient causer les intelligences artificielles. A ce volet défensif s'ajoute un volet offensif (comme parvenir à l'alignement des valeurs dans la conception de l'IA) pour renforcer le contrôle de l'IA. Les chartes sur les bénéfices et les risques de l'IA classent les risques posés par les modèles IA de fondation (comme Chat GPT). Ces programmes d'IA qui dépassent le seuil de puissance de calcul sont réglementés de manière plus stricte. Parmi les risques, on trouve le risque inacceptable, celui considéré comme une menace pour les personnes et qui est donc interdit. Le risque inacceptable concerne, par exemple, la manipulation cognitivo-comportementale de personnes ou de groupes vulnérables spécifiques ou encore la catégorisation et l'identification

[19] Comme, par exemple, la charte sur les *Principes d'Asilomar sur l'IA*, adoptée en 2017 lors de la Conférence d'Asilomar sur les Bénéfices et Risques de l'Intelligence Artificielle.

biométriques des personnes (avec la reconnaissance faciale immédiate). En ce sens, le 13 mars 2024, le Parlement Européen a approuvé une résolution législative sur l'intelligence artificielle[20] pour garantir la sécurité et le respect des droits fondamentaux par les systèmes d'IA. L'évaluation du risque de l'IA n'est pas seulement une enquête scientifique, elle est aussi une action politico-culturelle. Les algorithmes de l'IA peuvent être biaisés par des *a priori* politico-culturel ; ils sont élaborés par des humains, qui sont eux-mêmes sujets à des biais, et ces distorsions peuvent se refléter dans les IA qui risque de reproduire ces biais, entraînant des résultats discriminatoires et injustes[21]. Le fait est que ces grands modèles de langage de l'IA reproduisent les préjugés inhérents à notre société. Parmi ces biais, on retrouve le biais de représentativité[22], le biais de confirmation[23] et le biais intersectionnel[24], qui peuvent mener à des décisions discriminatoires dans des domaines sensibles tels que la santé, l'emploi, le crédit et la justice.

En conclusion, bien que l'IA soit un outil impressionnant, elle ne peut pas égaler l'intelligence humaine. L'IA ne possède ni compréhension, ni conscience de soi, ni capacité à concevoir des concepts, ni émotions, désirs, corps ou biologie ;

[20] *Résolution législative européenne sur l'intelligence artificielle* du 3 mars 2024 (https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_FR.pdf). Voir aussi les textes du Conseil de L'Europe et des Nations Unis sur le respect, la protection et la promotion des droits de l'homme dans la conception, le développement, le déploiement et l'utilisation de l'IA: *Projet du Conseil de l'Europe de Convention-cadre sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'État de droit* du 19 mars 2024 (https://search.coe.int/cm/pages/result_details.aspx?ObjectId=0900001680aee410), *La résolution (adoptée sans vote) de l'Assemblée générale des Nations Unies sur la promotion des droits de l'homme dans la conception de l'IA* du 21 mars 2024 (https://news.un.org/fr/story/2024/03/1144211).

[21] Sur ce point voir : M. Kaleem Galamali, *Craintes et inquiétudes concernant les biais de l'intelligence artificielle: Une perspective académique pour l'analyse éthique des biais possible avec l'IA*, Editions Notre Savoir, Paris, 2023; C. O'Neil, *Algorithmes: la bombe à retardement*, les Arènes, Paris, 2018. Selon la mathématicienne américaine Cathy O'Neil, la data science est tout sauf objective. Au contraire, elle accroît les inégalités en catégorisant les personnes en fonction de leur revenu, de leur lieu de résidence, de leur sexe... offrant des opportunités à certaines et les refusant à d'autres selon des critères appartenant au passé, reproduisant les comportements indéfiniment.

22 Le biais de représentativité survient lorsque le jeu de données utilisé pour entraîner un modèle d'IA n'est pas suffisamment diversifié ou représentatif de la population ou de la réalité cible. Par exemple, si un modèle de reconnaissance faciale est majoritairement entraîné sur des images de visages d'individus d'une certaine ethnie, il pourrait être moins précis pour identifier ou analyser les visages d'individus d'autres ethnies.

[23] Le biais de confirmation dans l'IA se réfère à la tendance d'un modèle à favoriser des informations qui confirment les hypothèses ou les préjugés déjà existants dans les données d'entraînement ou dans sa conception.

[24] Par exemple, un système de recrutement alimenté par l'IA pourrait être biaisé contre les femmes, mais encore plus biaisé contre les femmes de couleur, mettant en évidence un biais intersectionnel entre le genre et la race.

elle agit comme un perroquet stochastique[25]. Son raisonnement causal est limité et elle manque des subtilités de la communication non verbale que les humains possèdent intrinsèquement. Contrairement à nous, l'IA est dépourvue de conscience et bien qu'elle exécute de nombreuses tâches plus rapidement, elle n'atteint pas la profondeur de la pensée humaine. L'intelligence véritable, intrinsèquement humaine, se meut au-delà du simple calcul ou de la cognition. L'IA demeure étrangère à nos aspirations les plus profondes. L'IA incarne un outil, non un substitut à la richesse de la pensée humaine. En effet, contrairement aux systèmes IA, la langue humaine s'entoure d'une aura d'indétermination, où ce qui adviendra reste un mystère. C'est cette qualité insaisissable, faite de subtilités émotionnelles et temporelles, qui fait défaut à l'IA. Autrement dit, si l'IA excelle dans la résolution de problèmes, elle ne peut néanmoins pas embrasser toute la complexité des situations humaines. Dans cette quête faustienne entamée depuis l'aube de l'ère postmoderne, l'humanité explore ses limites en façonnant des outils numériques agiles et puissants, mais il ne doit pas les confondre avec l'essence profonde de son propre esprit.

[25] L'expression "perroquet stochastique" sert d'illustration critique afin de caractériser la nature des modèles d'intelligence artificielle (IA) contemporains, mettant en évidence leur capacité à produire des contenus linguistiques apparemment naturels sans réel engagement intellectuel ou appréhension contextualisée. Ce terme souligne la fonctionnalité de ces modèles basée sur des processus probabilistes et répétitifs. La métaphore du "perroquet stochastique" révèle une limitation fondamentale des technologies actuelles de l'IA dans le domaine de la compréhension et de la création de langage. Alors que ces modèles se révèlent efficaces pour générer des textes qui paraissent pertinents et cohérents, ils peuvent manquer de la profondeur de compréhension inhérente à la conscience humaine et rester confinés aux apprentissages qu'ils ont préalablement intégrés.